Towards the characterization of the eukaryotic selenoproteome

a computational approach

Sergi Castellano Hereza

Ph.D. dissertation

Barcelona, May 2004

Departament de Ciències Experimentals i de la Salut
Universitat Pompeu Fabra

Towards the characterization of the eukaryotic selenoproteome

a computational approach

Sergi Castellano Hereza

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 Guigo Serra **Sergi Castellano Hereza**

Barcelona, May 2004

To my parents
To my brother and sister

Dipòsit legal: B.49649-2004
ISBN: 84-689-0206-3

Preface

The election of selenoprotein genes as the subject of a *PhD* within the area of gene prediction is, and no effort is made here to deny that, a rare case. Not only because it focus on particular type of genes scarcely present in the genome, but because these genes subvert some biological rules we all trust and carefully implement in our gene finding schemas. However, this dissertation and, in general this line of research on selenoproteins, is not the result of a random turn or a crazy night promise[1]. On the contrary, it is originally rooted on a very specific question posed by previous experimental results on selenoproteins in the fly model system and, not negligible, on the belief that the study of exceptions, if so, can provide great insight into biology.

These works in *Drosophila* were initiated by Montserrat Corominas and Florenci Serras at the Universitat de Barcelona. In their lab, Berta Alzina[2] (1999), presented her *PhD* work in which the fly *sps1* gene (*seld*) was described and functionally characterized by means of the *sps1*^{ptuf} mutant. The *sps1* gene is a key enzyme of the selenoprotein biosynthesis pathway and, so far, its mutant is the only available in such a top control position in eukaryotes. It was then shown to be a recessive mutation which blocked the biosynthesis of selenoproteins and caused larval lethality, suggesting an important role of selenoproteins in cell function. A first clue to this was the observation that homozygous larvae had extremely reduced and abnormal imaginal discs, the cells of which accumulated free radicals and entered apoptosis. Thus, selenoproteins may be instrumental to maintain a certain redox state in the cell. In addition, *sps1* has a paralogue, termed *sps2*, which, interestingly, differs in the presence of a selenocysteine residue in place of a cysteine in the predicted active site of the enzyme.

With such grounds, additional functional studies were in need to clarify the implications of selenoproteins in the regulation of the redox state in the cell and, to this, the description of the fly selenoproteome became of outstanding relevance. Specially, because, no actual selenoprotein besides the apparently non-redox *sps2* gene, was known at that time in the fly. Therefore, lacking the the downstream-acting selenoproteins in the system, those to be blame for the phenotypes observed, not much could be said. Subsequently, the first of these premises was further tackled in a second thesis in the same lab. Marta Morey (2003) finished her *PhD* concluding that the interference in selenoprotein biosynthesis results in accumulation of ROS and consequently in a toxic intracellular environment, which can trigger apoptosis to eliminate the deleterious cells. This line of investigation is currently followed up, in the same group, by Cristina Pallarès.

At the same time, the first version of the *D. melanogaster* genome was released, and the second of our premises, the finding of the complete selenoprotein set in *Drosophila*, became achievable. At that point, I was already hooked[3] by selenoproteins, which later on led to writing to my funding agency for a change in the *PhD* topic, originally the development of methods for syntenic gene prediction. Anyway, we and others, analyzed the fly genome and two novel selenoprotein families were uncover. Recent work has just started to discuss the function of these genes, with so far controversial results (Morozova et al., 2003; Hirosawa-Takamori et al., 2004). Other genomes and other new selenoprotein families came later.

Bioinformatics is steadily gaining importance worldwide and, in particular, in our city. In this direction, selenium-dependent proteins attract a growing number of researchers with who I have the pleasure of collaborating. On one hand, Charles Chapple is already going through my `bash` and `gawk` scripts to, surprisingly, translate them into the more fancy `perl` language. Let me just say that, as part of his thesis, he is already successful in the finding of novel selenoprotein genes and in understanding their

[1] Though, it could well be with no influence in its scientific validity.

[2] Who, serendipitously, I first met in year 1994 in a NGO to discuss third world issues (a topic, sadly, as current as before!).

[3] An evergrowing number of people would prefer here the term *obsessed* for the sake of precision. They are not to be heard and several euphemisms will alternatively be used throughout this text.

distribution across the eukaryotic lineage. In addition, a new focus on selenoproteins is the one followed in the lab next door, where Jordi Villà and Alfons Nonell, who is starting his *PhD*, work on understanding the differential role of selenium (Se) and sulfur (S) in enzyme reactivity. Their results may help us to comprehend the preferential and organism-specific usage of Se and selenoproteins.

We, sort of Catalan selenoproteins gang, are just starting in terms of melting selenium and bioinformatics. However, and while it has not given to me the ability to foresee the future, I anticipate it as a pot in which effort and curiosity will be put together.

Sergi Castellano
Barcelona, May 2004

Contents

List of Figures

List of Tables

Acknowledgements

Possibly, you are among those, the smart majority, who will only go through the acknowledgements section of this dissertation. If so, keep reading and, if I unwrapped my memories correctly, you may spot your name along these lines.[4] On the other hand, if you are a formal member of the tribunal eager to judge my *PhD* work, feel free to skip this section. However, If you do insist in reading, proceed but be indulgent. Please.

I believe that, to acknowledge, is to mirror oneself's absences. Here, I share mines. I not only recover with ease, but I delight to recall from tiny details to important moments and people that, in these intense years, have made the difference to me. Though, and I say this with past awareness, I have often lack the vision to appreciate them immediately or I have done it naively. A complacent claim is not my intention, and this scientific work was indeed the result of many brilliant but as many other sweet-and-sour instants. Not surprisingly, as happens in any other aspect of this unnecessarily demanding life we are all in. To those who were involved in such moments because of me, thank me or sue me. In any case, sorry in advance. In addition, comments here may be highly biased depending on factors I do not need to worry you about. Nevertheless, this is what I choose to remember and this can be considered as a subjective and invalid disclaimer.

In a less pedantic writing, this is no more than a private account for places, events and people, which, in the years to come, I will reread with pleasure. Years away from Barcelona. For such, it is not necessarily short, neither written with the reader in mind nor less important than any other chapter in this book. With the passing of time, in fact, it is bounded to become of greater relevance to me.

Thanks.

To susanna Thanks for coming at the right time. For your patience. For losing sleep just to listen to me and make sure everything was in place. For your sweetness. For your car and to those traffic lights which, unexpectedly, turned green faster than usual in the late way home. For your mother. For your strength. For your *PhD*. For introducing me to txell. For your new life in London. For sheltering me when I most needed it. Thanks for your caring friendship.

To robert Thanks for moving back to Barcelona. For your perspective. For your companion. For your help. For our late-night whiskey sessions. For your *truita de patates* (best in town!). For your surprising jerseys. For the hard construction works you did at my place. For the Cuban restaurant the other day. For our coffee time. For your experience. For your advise. For your calm. For your exceptionality. Thanks for a lifetime tuition.

To txell Thanks for your love to me. For your generosity. For your doubts. For your patience. For our consented ambiguity. I have possibly made more mistakes with you than ever before. For taking me out, to the beach, to the theater, to dance and to everywhere. For just walking together holding hands uncountable afternoons. For that weekend. For cooking with me. For Baileys and Martini. For driving me up and down. For the SONAR festival. For waiting for me again and again at the airport. For taking care of my business when I am abroad. For sleeping with me when I could not sleep. For all your gifts, calls and letters. In fact, you just called me while I was writing this paragraph to ask me how my broken arm was doing and to tell me that, finally, you had differentiated pancreatic stem cells. Though, I did not understand your explanation. My fault. I will miss you, sergi.

[4] Altogether with some personal notes that may or may note coincide with your own opinion.

To genis Thanks for sharing with me your doubts. For telling me what others feel when I am to stupid to be aware of. For preparing the doctorate seminars on Sunday's nights. For training `geneid`. For watering my plants. For your genis-shirts. For introducing me to pizza terradellas. For those dinners in Lisbon (remember the onions soup?). For helping to your best. For your honesty. For your personal style in everything. Thanks for your authenticity.

To bet Thanks for your proximity. For your spontaneousness. For your kindness. For your happiness. For your patience. For your beauty. For helping me out with the neverending cheese. For making me laugh. For teaching me Italian. For hugging me. For your affection. For touching my emptiness. Thanks for giving me an undeserved glimpse of you.

To xavi Thanks for your smart quotations. For saying no more than the necessary. For that dinner at the Sibarit restaurant. For the friendly management. By the way, oscar told me that we could not yet find your home directory. For MOSIX. For keeping the system up. For listening to me. A life-wise guy. For calling me anytime you are back in Barcelona. Thanks for keeping in touch.

To nat Thanks for daring. For your patience. For cheering me up. For your music. For the tematik sessions. For all the beers. For the Ambar bar. You are the best dj!. For not asking me for more than I could give in troublesome times. Thanks for all the good moments together.

To mire Thanks for appearing. For inviting me to your wedding when we seldom knew each other. For one of your workmates!. You got mad at me for preventing you to reschedule the doctorate lecturers. For la Paloma. kofi? fashion-girl!. For coming to "La Passió d'Esparreguera every single year. For the rest of you: DO NOT MISS IT!. For letting me know carme. For carme and our nights smoking I do not know what. Thanks both of you.

To isabel For your impressive friendship. For your guts. For taking me to see cartoon movies. For the worst hangover I have ever had. For painting my balcony. For fabien. I really admire you. Thanks for sharing your hapiness with me.

To nu For your coherence. For your example. For your father. For your pain. For your courage. I am proud that you call me friend. Thanks for your inspiration.

To marta For all your visits. For that new year eve. For worrying about me. Thanks for that and more.

To georgy . For coming back in time. For laughing with me and, most of the times, at me. For taking me out with your friends. For that calotada. For nuria and eli. For that dinner at Rita Blue. For "La Fabulosa Historia de los Inolvidables Marrapodi". For my catastrophic mail. For your mother's chicken. For your father's whiskey. Thanks for being back.

To cocktail-bars All of them.

To lo pep Thanks for your expertise in everything. For the coffes. For your help. For the jokes.

To enrique For coding `geneid`. For debugging it. For helping me.

To er moi and homo Thanks for being the first ones. Freakies forever!

To jordi Thanks for always looking at the bright side of life. For asking no question that night. For that lunch in Strasbourg. "No hay medusas mas".

To charles For your sense of humour. For waking up. For your enthusiasm.

To lab's crew hugo, francisco, oscar (master of technicians!), eduardo, noura, mar, nicolas, dezi, cherraiz, jacobos, fabien, lulla, jan, josep, jordi, ramon, alfons, To rut, best of lucks!. To montse.

To roderic Thanks for being honest. For always listening. For your passion for science. For teching me a lot.

To doctorate people To samitu and our wednesday's cinemas sharing popcorn. To anna and our surrealistic lunch at Laie. To miki and his incredible job stories.

To armando Thanks for your humanity. For that "pollo con chile" you showed me how to cook. For all the fantastic walks we did together when you lived close to the Raval. For no saying. You had it all and now we have nothing. I really miss you.

To helena Thanks for first arguing but later giving me advice many times.

To isabel's lab To **maite** for that snapshor together. To **paco** for such a witty conversation on the mountains.

To susanna's lab Specially to **clara**.

To txell's lab Specially to **encarna** for inviting me to that paella. To **ferran**. To **berta**. To yolanda. To anna.

To aida's lab Specially to **oscar**. To **anna g**. To **monica**. To **david**.

To inma Thanks for teaching me so many things. Thanks for taking me out of the shell. Thanks for not letting me down when I was ill. One kiss.

To apiserum Energy in a bottle with a bit of alcohol.

To other friends To **fiona**. To **miquel**. To **alina**. To **eva**. To **maruxa**. To **mario**.

To universities' group To **joan**. To **txell**. To **nela**. To **ignasi**. To **mireia**. To **evelyn**. To **guille**.

To sparrakas' gang To **vicen**. To **masde**. To **oriol**. To **monne**. To **juanpa**. To **chus**. To **quim**. To **laia**. To **monica**. To **carme**. To **marta**. To **tirsos**.

To physiotherapy's group To **jose**, master of pressure and pain. To **marta**, always happy. To **monica**, always cheerful. To **paulino** for letting me in ahead of schedule those days I was in a hurry.

To liliana Thanks for your friendship.

To montse

To angeles Who taught me the laws of physics while going from Madrid to Valencia to Barcelona. Thanks for inspiring me to do a *PhD*.

To alex Who taught me english many afternoons and nights while dancing at the Dot Light Club. I dare to say that she learned some spanish from me too.

To lourdes Who taught me biology at high-school. Who taught me the passion for details. I want you to have a copy of this work, which is also yours.

To all I forgot and here we are now.

To aida Thanks for always being there, even when you should have not. Thanks for all the japanese restaurants. Thanks for not letting me go. Thanks for your gifts. Thanks for not losing me. Thanks for cooking with me. Thanks for your contradictions. Thanks for taking care of me while my arm was badly broken. Thanks for never say never. Thanks to your mother for such pleasant sunday lunches and weekdays dinners. Thanks for our talks, all of them!. Thanks for your honestity. Thanks for ever say ever. Thanks to Havana 7. Thanks for my ambiguity. Thanks for dancing with me. Thanks for loving me. Thanks for all those weekends together. Thanks for making me that happy. Love, sergi.

Thanks.

On the professional side, I would like to acknowledge the following individuals and institutions. Thanks.

To Roderic Guigó Thanks for your supervision. For letting me travel that much.

To Montserrat Corominas and Florenci Serras Thanks for all the useful conversations.

To Marta Morey and Berta alzina Thanks for all your hard work.

To Vadim Gladyshev Thanks for inviting me to a short stay at the University of Nebraska in January-February 2001. For your wide knowledge.

To Gregory Kryukov Thanks for your collaboration and friendship-

To Alain Krol Thanks for your smart perspective.

To Marla Berry Thanks for all your work.

To CIRIT Thanks for a PhD scholarship in years 2000-2003.

To IMIM Thanks for a PhD scholarship in year 2004.

To Cold Spring Harvour Laboratory Thanks for the partial scholarships in years 2001, 2002 and 2003.

Thanks.

Commentaries

Tips on reading this thesis

Some hints while browsing through.

On the text

Footnotes those comments no needed for following the main flow of the text, and that may interrupt it, are placed in an *ad hoc*[5] footnote section; and

Software names are written in a `typewriter` font and in lowercase as they would be invocated in the command line (eg.[6] `geneid`); and

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

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

On the main chapters

Results section includes the original research papers; and

Methods section includes their supplementary material; thus, a deeper methodological description is provided.

On the appendix section

Abbreviations a list of the shortened words in the text; and

Glossary those terms that may have confusing or alternative meaning in the literature are defined as understood throughout this book; and

A genetic code chart a table of the canonical/standard genetic code useful while reading this pages; and

The GNU[7] General Public License freedom to share and change free software. Unvaluable for the work presented here, hence, the majority of the programs used are under this license or similar ones. By the way, say no to the European Directive on the patentability of computer-implemented inventions (see `http://register.consilium.eu.int/pdf/en/04/cm00/cm00543.en04.pdf`); and

Human genome and human rights the UNESCO 1997 declaration on the human genome interlink between freedom of research and respect to human rights; and

List of publications a complete list of the material, published and unpublished, derived from this work is given here; and

[5] For the particular purpose.

[6] For example

Contact information addresses for some of the authors and coauthors of the articles presented in the Results section; and

Curriculum Vitae a brief scientific account of the author of this dissertation; and

Miscellanea technical annotation of the $\LaTeX 2_\varepsilon$ framework used to produce this thesis book. Download this thesis template.

On the PDF file

Bookmarks in your PDF reader allow easy navigation through the document; and

Thumbnails in your PDF reader put pages just one click away; and

Table of contents page numbering in blue is linked to the corresponding sections; and

References in blue are linked to the Bibliography section at the end of the document; and

URLs in blue are linked to the corresponding electronic resource (if undefined in your PDF viewer, it will prompt you for the application to use).

A PhD in the Spanish scientific system

No matter how you look at it, Spain is at the back of science in Europe. However, the situation is changing (though at a pace that will always seem slow to us). In relation to the so-called *PhD* students, it would be fair if their contribution to science started to be generally recognized and their professional conditions improved in a way that the *PhD* could really be considered the first step of any scientific career. For such a big change, scholarships should turn into real contracts with unemployment and social security coverage (that basic rights have nothing to do with the level of experience is too obvious to worry you about. Only money does it.). This scenario is still far, but fair, given the originality of the work demanded to any *PhD* student (who, in the best of the worlds, would be named simply investigator).

Abstract

Although the genome sequence and gene content are available for an increasing number of organisms, eukaryotic selenoproteins remain poorly characterized. In these proteins, selenium (Se) is incorporated in the form of selenocysteine(Sec), the 21st amino acid. Selenocysteine is cotranslationally inserted in response to UGA codons (a stop signal in the canonical genetic code). The alternative decoding is mediated by a stem-loop structure in the 3'UTR of selenoprotein mRNAs (the SECIS element). Selenium is implicated in male infertility, cancer and heart diseases, viral expression and ageing. In addition, most selenoproteins have homologues in which Sec is replaced by cysteine (Cys).

Genome biologists rely on the high-quality annotation of genomes to bridge the gap from the sequence to the biology of the organism. However, for selenoproteins, which mediate the biological functions of selenium, the dual role of the UGA codon confounds both the automatic annotation pipelines and the human curators. In consequence, selenoproteins are misannotated in the majority of genome projects. Furthermore, the finding of novel selenoprotein families remains a difficult task in the newly released genome sequences.

In the last few years, we have contributed to the exhaustive description of the eukaryotic selenoproteome (set of eukaryotic selenoproteins) through the development of a number of *ad hoc* computational tools. Our approach is based on the capacity of predicting SECIS elements, standard genes and genes with a UGA codon in-frame in one or multiple genomes. Indeed, the comparative analysis plays an essential role because 1) SECIS sequences are conserved between close species (eg. human-mouse); and 2) sequence conservation across a UGA codon between genomes at further phylogenetic distance strongly suggests a coding function (eg. human-fugu). Our analysis of the fly, human and *Takifugu* and *Tetraodon* genomes have resulted in 9 novel selenoprotein families. Therefore, 20 distinct selenoprotein families have been described in eukaryotes to date. Most of these families are widely (but not uniformly) distributed across eukaryotes, either as true selenoproteins or Cys-homologues.

The correct annotation of selenoproteins is thus providing insight into the evolution of the usage of Sec. Our data indicate a discrete evolutionary distribution of selenoprotein in eukaryotes and suggest that, contrary to the prevalent thinking of an increase in the number of selenoproteins from less to more complex genomes, Sec-containing proteins scatter all along the complexity scale. We believe that the particular distribution of each family is mediated by an ongoing process of Sec/Cys interconversion, in which contingent events could play a role as important as functional constraints. The characterization of eukaryotic selenoproteins illustrates some of the most important challenges involved in the completion of the gene annotation of genomes. Notably among them, the increasing number of exceptions to our standard theory of the eukaryotic gene and the necessity of sequencing genomes at different evolutionary distances towards such a complete annotation.

Introduction

1.1 Overview

In the last decade, selenoproteins, a peculiar family of selenium-containing proteins present in all three domains of life, have been thoroughly studied. It has now become clear that, in selenoproteins, a stop codon (UGA) is recoded to cotranslationally insert selenocysteine[1]; this is due to an RNA secondary structure (SECIS) and a few other proteins that direct the recoding process. However, much is yet to know about the actual number of selenoprotein genes and their biological properties. Recently, the advent of large-scale DNA sequencing projects and the parallel development of the bioinformatics field, have provided a powerful approach to their study. In this direction, the final goal of this work was to contribute towards the complete definition and characterization of the eukaryotic selenoproteome (set of all selenoproteins encoded in these genomes) through computational methods.

This introductory chapter shortly describe[2], at the appropriate depth for the work presented here, the key biological mechanisms of interest to selenoproteins and the general bioinformatics background needed for the computational prediction of selenoprotein genes in eukaryotic genomes. The introductory sections of the articles included in this dissertation (see Results) provide extra, but somewhat more specific, elements on the research described here.

1.2 The genetic code

Nearly 40 years ago, Crick and colleagues deduced the general nature of the genetic code from the results of crosses between mutants in the T4 rIIB cistron. The genetic code is written as a set of all 64 possible arrangements of the four RNA nucleotides –U, C, A and G– in the form of a triplet (codon). These seminal works resulted in a particular assignment of each of the 64 codons to either one of the 20 amino acids used in protein synthesis (61 codons) or a termination function (3 codons: UGA, UAG and UAA). In addition, they recognized that the code must be nonoverlapping (codons do not overlap), degenerate (most amino acids are encoded by more than one sense codon), and read from a fixed point (usually the AUG (Met) start codon)[3]; moreover, they predicted that some triplets must be nonsense.[4] An excellent review on this topic is the one by S. Osawa 1995, with a special focus on the evolution of variant genetic codes.

[1] Selenocysteine is recalled as the 21st amino acid and it is designated as Sec and U in the three and one-letter code, respectively.

[2] Comprehensiveness is not intended here and appropriately selected references, including reviews and books, are generously given throughout.

[3] Though, also GUG (Val, 8%) and UUG (Leu, 1%) in bacteria).

[4] The term nonsense codon should be revisited. There is a clear difference, in meaning and molecular function, between the so-called nonsense codons and real unassigned codons. This latter class, while having an assigned amino acid in the canonical code, do not code for one in other genetic codes. Furthermore, neither are recognized by a release factor. This fact, stands for the difference between real stop codons and unassigned ones. Stop codons do stop polypeptide chain growing, but through release factor interaction also direct chain releasing from the ribosome. This is due the release factor mimicry of the anticodon arm of the tRNA, which pairs to the codon triplet by protein-RNA binding; thus, in some sense, stop codons code for release factors. On the contrary, unassigned codons only promote chain elongation but not its release from the ribosome.

1.2.1 The standard code

The established genetic code was then thought to be common to all organisms and viruses, because it was the same in some very different organisms (e.g. yeast, vertebrates and the tobacco mosaic virus). Therefore, it was claimed to be universal. Such apparent universality of the code led Crick to propose the frozen-accident theory. Crick (1968) states that:

> This account for the fact that the code does not change. To account for it being the same in all organisms one must assume that all life evolved from a single organism (more strictly, from a single closely interbreeding population).

The theory further states that the proteins had become so sophisticated in a single pool of progenote cells that any changes in codon meaning would disrupt proteins by making unacceptable amino acid substitutions. Therefore, the genetic code was frozen. These code, the one though to be universal, can be read in Appendix ?.

1.2.2 The nonstandard code

The universality of the genetic code was first challenged in 1981, when mammalian mitochondria were found to use a code which deviated somewhat from the universal.[5] Since then, many other organisms with a deviant code have been uncovered (see below). Taking the canonical code as a reference, two main types of coding deviations can be defined:

Major or complete those that affect a codon type in all genes of a given organism; and

Minor or partial[6] those that affect only a codon type in some genes of a given organism.

Major differences are due a complete change of meaning of a codon in a genome; thus, this affects all proteins encoded. On the contrary, minor changes are due to a specific translation control in certain mRNAs which recodes a particular codon, as happens to UGA in selenoproteins. Therefore, giving raise to a codon ambiguity phenomenon in which the same codon has two (or more) contrasting meanings in the same genome.

Major deviations have been found in both, the mitochondrial and nuclear codes. Some well-known major deviations in the mitochondrial code are, among others,: 1) UGA (stop)[7] codes for Trp in all but green plants; 2) CUU, CUC, CUA and CUG (Leu) codons code for Thr in yeasts; 3) AUA (Ile) codes for Met in yeast and animals; 4) UAA (stop) code for Tyr in Planaria; 5) UAG (stop) is unassigned in green algae; and 6) AGA and AGG (Arg) code for Ser in invertebrates and Gly in tunicates. In addition, major deviations in the nuclear code include (Osawa, 1995) (but are not limited to): 1) in *Mycoplasma* and *Spiroplasma* UGA (stop) codes for Trp; 2) in *Mycoplasma* CGG (Arg) is unassigned; 3) in certain ciliated protozoans, UAA (stop) and UAG (stop) code for Gln; 4) in *Candida* CUG (Leu) codes for Ser; 5) in *Micrococcus* AUA (Ile) and AGA (Arg) are unassigned; and 6) in *Euplotes* UGA (stop) codes for Cys. Clearly, stop codons seem to change easily. This may be due their rareness (they occur once per gene) and the fact that release factors are easy to modify.

On the other hand, besides selenocysteine, other minor and dual coding divergences exist which involve stop codon redefinition by incorporation of one of the usual 20 amino acids. In this process, the specific context of the termination codon is such that occasionally it is decoded by a tRNA rather than a release factor, allowing ribosomes to synthesize an extended polypeptide. Readthrough of these stop codons[8] by recoding involves a variety of stimulatory signals. The following cases have been reported (**?**): 1) in the phage QB coat protein, a UGA (stop) codon is readthrough; 2) in tobacco mosaic virus, a UAG (stop) codon is readthrough; 3) in barley yellow dwarf virus, a UAG codon is readthrough; 4) in

[5] As many deviations from the universal code have now been observed, this term seems excessive and it will not be used here. One code is predominant, but it is not universal; thus, the terms canonical or standard are preferred instead.

[7] The codon standard meaning is given in parenthesis

[8] To my knowledge, no codons other than those signaling termination (UAA, UAG and UGA) have been found to be recoded in specific genes. However, the misincorporation of amino acids in “sense” codons would open a brand new way of protein modification control.

Drosophila headcase (*hdc*) gene a UAA (stop) codon is readthrough; 5) in *Drosophila* kelch gene a UGA is recoded; and 6) in *Drosophila* out of first (*oaf*) gene a UGA is recoded. Again, stop codons seem to be highly susceptible to specific partial recoding and the reason for this could be much the same as for complete alternative decoding, altogether with the slow-to-decode property of stop codons which gives time for recoding.

A special case is the one of pyrrolysine, the 22^nd amino acids. In the archaeal *Methanosarcina barkeri* a UAG codon is recoded to incorporate this residue in monomethylamine methyltransferases. Whether specification of pyrrolysine is due a permanent reassignment of UGA or to recoding of a subset of such codons is not yet clear.

1.2.3 Evolution of the code

The relationship between major and minor changes in the canonical code is unclear. In essence, three theories account for the changes in the genetic code:

Codon capture hypothesis proposes that fluctuations in mutation bias that influence G+C content can eliminate codons from the entire genome (by the conversion to a synonymous codon and the loss of the corresponding tRNA or RF that translate the codon), after which they can be reassigned by neutral processes. Therefore, the production of unassigned codons is probably an intermediate step in codon reassignment during evolution of the genetic code. Depending on the appearance of a new tRNA (or RF) the lost codon could later reappear (by the conversion of another codon, mainly a codon that is synonymous with the altered codon) with the same or other amino acid or termination meaning;

Ambiguous intermediate hypothesis notes that tRNA mutations at locations other than at the anticodon can cause translation ambiguity, and ultimately fixation of the new meaning if it is adaptive. Therefore, there is a state in which some codons have more than one meaning; and

Genome streamlining hypothesis suggests that code change, at least in mitochondria and obligate intracellular parasites, is driven by selection to minimize the translation apparatus.

These theories, and other suggestions about variations in the code, are not mutually exclusive. Different codon reassignments may result from different causes. For instance, codons that have been made rare during extreme G+C pressure might be easier to reassign via an ambiguous intermediate, as the impact of mistranslation at those codons would be ameliorated by their scarcity. However, none of these hypothesis account explicitly for the dual decoding of codons and, specially, for the inclusion of selenoproteins and the recoding machinery involved within a general framework of the genetic code evolution.

1.3 Translation

Translation is a sophisticated, complex and, at the same time, fragile process. It is now widely acknowledged as one of the major gene expression control points in the cell. This is due to the regulation of the efficiency of mRNA in specifying protein synthesis. However, despite the term translational control usually refers to an effect that causes a change in the rate of mRNA translation, in this work this term will be expanded to the mechanisms that regulate the final sequence of the encoded protein.

Translation maps mRNAs into proteins. This decoding[9] step, is usually believed to follow the canonical genetic code (in other words, the canonical correspondence between codons and amino acids). In this way, the ribosome, after finding a specific start site on the mRNA that sets the reading frame, translates the nucleotide sequence into an amino acid sequence one codon at a time. This linear process ends when a stop codon is reached and the growing protein is released. This would be the standard mechanism but it has now become well established that alternative decoding (also named non-standard decoding or recoding) is a widespread mechanism. In what follows, a brief overview of the eukaryotic translational process for a better understanding of the selenoprotein recoding mechanism is given.

[9] Decoding means to map one code into a another, codons into amino acids in this case

1.3.1 Translation initiation

In eukaryotes, translation starts when the preinitiation complex (the small ribosome subunit carrying the initiator tRNA^{Met} and other protein factors) binds the m^{7}G-cap-proximal region of the mRNA. Then, a downstream scanning process in the 5'UTR takes place in the search of a suitable start codon (AUG). In most cases, it happens to be the first one encountered by the ribosome subunit, but exceptions exist. Thus, the ribosome stops when it binds stably at the initiation codon, primarily through the RNA-RNA interaction of the AUG and the CAU anticodon of the bound tRNA^{Met}. Once a AUG is recognized, the initiation factors that were bound to the 40S initiation complex dissociate and the large 60S subunit can bind. Protein synthesis is ready to start.

1.3.2 Translation elongation

The stepwise movement of the ribosome along the mRNA produces the assembly of amino acids into a polypeptide by decoding nucleotides into amino acids as triplets (codons). It requires a group of proteins termed elongation factors that fall into two groups: 1) those necessary for the recruitment of aminoacyl-tRNAs to the ribosome; and 2) those involved in the subsequent translocation event in which the ribosome moves ahead. selenoproteins have specific proteins of the first class as well as its own tRNA^{Sec}.

1.3.3 Translation termination

This step is of outstanding importance in the selenoprotein translational process. Its general framework is now well known. Translation termination is initiated when one of the three stop codons is present in the ribosomal A site, resulting in binding of the Release Factor (RF). Then, the hydrolysis of the peptide bond results in a deacylated tRNA in the ribosomal P site. RF1 is removed from the ribosome in a GTP-dependent reaction involving RF3. Dissociation of the 70S/mRNA complex.

1.4 Selenoproteins

Selenoproteins are proteins that contain selenium in the form of the amino acid selenocysteine. Though could be argued that any selenium-containing polypeptide should be called selenoprotein, this designation is usually reserved only to selenocysteine-containing proteins. This amino acid is inserted co-translationally in response to UGA codons (usually stop signals) and the alternative decoding is due to several coordinated elements: 1) an mRNA secondary structure, the SECIS element, located in eukaryotes within the 3'UTR of the selenoprotein mRNA; 2) a SECIS binding protein; 3) a specific tRNA^{Sec}; and 4) an elongation factor. Below, all these eukaryotic selenoprotein elements are described at length. A comprehensive introduction to selenoproteins can be found in (hat).

1.4.1 Selenium

Selenium is an essential trace element. This nutrient is mostly found in selenoproteins, which usually are involved in redox reactions (Low and Berry, 1996; Stadman, 1996) in which selenium plays a central role. This includes, in many cases, working as antioxidants against the effects of free radicals that are produced during normal oxygen metabolism. Free radicals can damage cells and contribute to the development of some chronic diseases.

Plant foods are the major dietary sources of selenium in most countries throughout the world. The amount of selenium in soil, which varies by region, determines the amount of selenium in the plant foods that are grown in that soil. Selenium also can be found in some meats and seafood. Animals that eat grains or plants that were grown in selenium-rich soil have higher levels of selenium in their muscle

Selenium and selenoproteins have been related to several diseases: in cancer as a chemopreventive agent (Combs and Lu, 2001), in preventing heart disease and other cardiovascular and muscle disorders (Coppinger and Diamond, 2001), in helping to relieve symptoms of arthritis (Kose et al., 1996; Aaseth et al., 1998), in reducing viral expression (Beck, 2001) and in delaying the progression of AIDS (Baum et al., 2001). Selenium may also be essential for normal functioning of the immune system (McKenzie et al., 2001) and the thyroid gland (Kohrle, 2000), in male reproduction (Flohé et al., 2001) and in slowing the aging process (McKenzie et al., 2001). Thus, selenium deficiency (commonly seen in parts of China and Russia with poor selenium soils) is known to affect thyroid function and it is linked to Keshan Disease where enlarged heart and poor heart function is observed (Levander and Beck, 1997). On the other hand, high blood levels of selenium can result in a condition called selenosis (Koller and Exon, 1986). Symptoms include gastrointestinal upsets, hair loss, white blotchy nails, and mild nerve damage. However, selenium toxicity is rare and reported cases have been associated with industrial accidents and a manufacturing error that led to an excessively high dose of selenium in a supplement (Raisbeck et al., 1993; Hathcock, 1997)

1.4.2 Selenocysteine

Selenium is found in selenoproteins in the form of a covalently bound selenocysteine residue, a cysteine analog in which a selenium atom is found in place of sulfur. Selenocysteine was recognized as a constituent of special proteins in 1976 (Cone et al., 1976). Until recently, it was thought that these *extra* amino acids were made by modifying one of the standard amino acids after it was incorporated into protein, a process called posttranslational modification. However, it is now clear that selenocysteine is inserted cotranslationally, directed by an in-frame UGA codon in the mRNA (Chambers et al., 1986; Zinoni et al., 1986). Thus, though usually it's thought that only 20 amino acids are specified by the genetic code, selenocysteine must be considered the twenty-first. As any other standard amino acid, selenocysteine has it's own codon (UGA) and a tRNA with the corresponding anticodon (UCA),

Selenoprotein that have identified functions are enzymes with selenocysteine in their active site. Natural variants containing a cysteine in this position have been identified for many of these enzymes, showing that selenocysteine per se, in most of the selenoproteins, does not possess an essential role. Mutational change of the selenocysteine residue to a cysteine also gives variants that are active (Axley et al., 1991; Berry et al., 1992) but have a lower overall catalytic efficiency (Axley et al., 1991). Thus, although selenocysteine confers a considerable catalytic gain and might give a selective advantage, it can be replaced by a cysteine in most enzymes.

It also should be noted that selenium can also be incorporated nonspecifically into protein (Hatfield et al., 1999). This happens when selenium replaces sulfur in the biosynthesis of cysteine or methionine and the resulting selenoamino acid is inserted in place of the natural amino acid (Müller et al., 1997). This free-selenocysteine incorporation basically depends on the relative abundance of S/Se (the S/Se ratio is between 10^3-10 ^{5} to 1 in the atmosphere) and the organism-specific S/Se biochemical discrimination capacity. Clearly, this misincorporation of selenium into protein has nothing to do with the recoding machinery, though may have some evolutionary implications, and may even be toxic (Hatfield et al., 1999). This nonspecific incorporation will not be discussed further in this work.

The biosynthesis of selenocysteine is as follows:

1. A specific tRNA^{Sec} is charged with serine by seryl-tRNA synthetase; and
2. The seryl moiety is converted into the selenocysteil residue by selenocysteine synthase with selenomonophosphate as selenium donor.

1.4.3 The UGA codon

When genes encoding selenoproteins were sequenced it was found that the codon that corresponded to the selenocysteine was a UGA (Chambers et al., 1986; Zinoni et al., 1986; Berry et al., 1991). Thus, in a single mRNA, UGA codons can have two contrasting meanings, stop or selenocysteine. Furthermore, the UGA codon can have several meanings depending the genome (mitochondrial or nuclear) or the organism.

1.4.4 Transfer RNA: tRNA^{Sec}

tRNA^{Sec} have the UCA anticodon and display sequence and structural differences from canonical tRNAs (Sturchler et al.; Baron et al., 1993). Two solution structures of tRNA^{Sec} have been determined, and show differential features from standard tRNAs.

In pyrrolysine, similar to selenocysteine, instead of a tRNA being charged with this new amino acid, it receives a standard one that is then enzymatically modified while still attached to the tRNA. Pyrrolysine is likely to be produced by the modification of a lysine (Lys) residue (serine (Ser) in selenoproteins) attached to a special lysil-tRNA (seryl-tRNA in selenoproteins). The tRNAs involved in the production of selenocysteine and pyrrolysine are distinct from those decoding the standard amino acids, serine and lysine, but they differ from each other in certain features, for example, the pyrrolysine tRNA has a special anticodon arm.

1.4.5 SECIS structure

The Selenocysteine Insertion Sequence (SECIS) is an RNA stem-loop secondary structure. In eukaryotes is located within the 3′UTR of selenoproteins mRNAs and its so-called canonical structure is the following (Berry et al., 1993; Shen et al., 1995; Walczak et al., 1996; Hubert et al., 1996; Kollmus et al., 1996; Low and Berry, 1996; Walczak et al., 1998; Grundner-Culemann et al., 1999; **?**; Fagegaltier et al., 2000b):

The consensus SECIS structure is composed of an initial helix and internal loop, followed by a second helix containing non-Watson-Crick base pairs UGAN....NGAN (the SECIS core or quartet), an unpaired A preceding the quartet, and an unpaired AA motif in the apical loop that ranges from the core 11 to 12 nucleotides (Walczak et al., 1996; Berry et al., 1997; Walczak et al., 1998) (Figure 1B and 2B). In addition, SECIS elements are divided into two classes, named form 1 and form 2, the latter having an additional small stem-loop at the end of the apical loop (Grundner-Culemann et al., 1999). However, no functional differences have been observed between the two classes and they exhibit comparable activity levels in promoting selenocysteine insertion (III et al., 1998), furthermore, the conversion of form 1 to form 2 or form 2 to form 1 SECIS elements in a given selenoprotein mRNA does not change their wild-type activity level (Grundner-Culemann et al., 1999). This initial consensus definition of the SECIS stem-loop, though correct for most selenoprotein genes, presents some deviations in later discovered SECIS structures. These are 1) the unpaired adenine is replaced by guanine in four SECIS elements (Buettner et al., 1999; **?**); and 2) conserved adenosines in the apical loop are replaced by cytidines in the human SelM SECIS structure (Korotkov et al., 2002). Thus, the SECIS structure has low primary sequence conservation, a high secondary structure conservation and an unknown tertiary structure conservation (though one would expect it to be high). SECIS diversity is still under research.

The presence of a SECIS element is compulsory for the correct recoding of the UGA codon. However, this element, on its own, is not enough to direct the UGA recoding in the translation process. In eukaryotes, at least two additional protein factors are required (see below).

1.4.6 SECIS binding protein: SBP2

The SECIS Binding Protein (SBP2) is essential for the cotranslational insertion of selenocysteineinto selenoproteins (Copeland et al., 2000). This protein binds the SECIS element and its binding specificity comes from a key feature of the SECIS, the highly conserved quartet of non-Watson-Crick base pairs.

1.4.7 Elongation factor: eEFsec

The Selenocysteil-tRNA-specific elongation factor (eEFsec) interacts directly with both tRNAs bearing selenocysteineand SBP2 (Tujebajeva et al., 2000; Fagegaltier et al., 2000a). So, the SECIS element, through a two-protein complex containing SBP2 and eEFsec can recruit selenocysteine-carrying tRNAs. In this way, the SECIS RNA element and two protein factors function together to recode the UGA codon.

1.4.8 Selenoprotein mRNA translation

Selenoprotein mRNA translation by the ribosome is, in overall, highly standard. As usual, 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, comes into play the decoding apparatus for eukaryotic selenocysteine insertion. Briefly, the SECIS element in the downstream untranslated region of the mRNA binds to SBP2. This protein in turn binds to the elongation factor eEFsec, which itself has already recruited the tRNA^{Sec}. The selenocysteine-bound tRNA is then delivered to the waiting UGA codon, for incorporation into the growing polypeptide.

1.4.9 Selenoprotein families

When this research was started, the following selenoprotein families were already known (the biological function is unclear for most of them):

15kDa this protein is directed to the endoplasmatic reticulum, 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); and

Glutathione peroxidases these proteins have an antioxidant function: the degradation of various hydroperoxides by a glutathione dependent catalysis; and

Thioredoxin reductases these proteins reduce oxidized thioredoxin (Trx-S2) at the expenses of NADPH and reduced thioredoxin [Trx-(SH)2] is reoxidez by disulfides in proteins generating thiols; and

SelW this small protein is highly expressed in the muscle and may have an antioxidant role; and

SelP this protein is a glycoprotein and is also the major plasma selenoprotein, accounting for 60% of plasma selenium. It is the only selenoprotein with more than one Sec residue; and

Selenophosphate synthetases these proteins catalyze the synthesis of monoselenophosphate; and

SelT this protein is of unknown function; and

SelR this protein (also called SelX) is a methionine sulfoxide reductase (MsrB); and

SelN this protein is of unknown function; and

Others were found meanwhile:

MsrA this protein is a methionine sulfoxide reductase;

1.4.10 Selenoprotein Distribution

At the start of this *PhD*, all known selenoprotein contained Sec in human, mammalian and other vertebrate genomes, while had Sec, Cys or did not exist in available invertebrate species. This led to the idea of an increase in the usage of Se, Sec and selenoproteins from less to more complex organisms in the eukaryotic lineage. In this scenario, humans and related would have always a Sec-version of any particular selenoprotein (with additional Cys-containing paralogs) and others would have only subsets of this selenoproteome as true selenoproteins.

1.4.11 Selenoprotein evolution

There are no yet comprehensive studies on the evolution of selenoproteins nor their associated translational machinery. However, the unique dual role of the UGA codon has been much under discussion in the genetic code evolution field. In short, the question to be answered here is the ordering of the coding possibilities of the UGA codon in the timescale. Some hypothesis are:

1. Leinfelder and co-workers (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, nearly all Sec residues in protein were switched to Cys. Codon UGA could have "acquired other functions such as its more familiar role in termination" while being retained for rare use in coding for Sec.

2. Jukes (1990) suggests that UGA could not have changed abruptly from coding for Sec to a stop codon. In the anaerobic world, the UGN family box was assigned to both Cys and Sec with anticodon UCA. When oxygen entered the biosphere, nearly all Sec was switched to Cys, which retained UGN codons and anticodon UCA. This anticodon duplicated, and one duplication mutated to GCA; the present Cys anticodon paired with UGY. The other UCA anticodon was captured by a "new amino acid" - Trp - with codon UGR. GC-pressure changed anticodon UCA to CCa, pairing only with UGG, the present Trp codon, and UGA disappeared, except for rare use in coding Sec.
3. Jukes and Osawa (1992) also proposed the opposite point of view. Sec arrived after the establishment of the code for 20 amino acids. As the use of Sec would have become more advantageous for anaerobic organisms than the use of Cys, the system could have undergone evolutionary refinement by positive selection.

In other words, it is unclear whether there was a transition from Sec to stop (though maintaining Sec in a few proteins) or from stop to Sec (though maintaining the stop function in the majority of genes). Alternatively, both stop and Sec meanings could have coexisted from the beginning.

Clearly, evolution of selenoproteins is a complex subject. selenoproteins are unusual genes and proteins and their study involves a wide range of biological processes. Thus, insight into their evolution and to the question raised above is expected to come from gathering different aspects of their biology. In this direction, current relevant questions include:

Selenium evolution Trace element but essential, how does the selenium machinery works?

Genetic code evolution As seen above, the genetic code is not static. Presumably, when decoding originated (at an initial RNA-protein stage), discrimination between alternative decoding forms (same codon coding for two or more amino acids) was weak. But once one mode became predominant, there would have been a selection to lock it in with increasing efficiency. Thus, there is an obvious question to answer: What does the TGA codon code first for? Sec, STOP, none or both?

Amino acids evolution Sec is the 21st amino acid, but what about the analog Cys (TGC or TGT) or Ser from which Sec is derived?

RNA regulatory signals The SECIS structures, are they monophyletic across the three domains of life?

Recoding The TGA codon is recoded to Sec, how recoding proteins interact with both the SECIS element and the ribosome?

Distribution In all domains, but, to sum up, what is the relationship between genes, proteins, selenium machinery, recoding machinery and RNA regulatory elements across the phylogenetic spectrum?

In conclusion, much is yet to know about selenoproteins and related machinery, the SECIS structure, the selenium Vs. sulfur usage and, above all, the parallel evolution of all these elements.

1.5 Gene structure prediction

With millions of bases of genomic DNA from different organisms released every day, computational gene identification has become an outstanding tool as a fast way of pinpointing protein coding elements.[10] In what follows, a brief introduction to the eukaryotic gene structure and to the computational methods used for its prediction is given.

1.5.1 Eukaryotic gene structure

The genes of most eukaryotic organisms are neither continuous nor contiguous. They are separated by long stretches of intergenic DNA and their coding sequences are interrupted by noncoding introns. In general, any eukaryotic gene is made of:

[10] However, any gene prediction should be considered only as a draft model of the real gene as long as not contrasted with additional experimental evidence. On the other hand, prediction of RNA genes will not be discussed here.

Promoter upstream region that regulates transcription (Figure 1.1);

Exons regions not removed by the splicing machinery (Figure 1.2). They are divided into:

> **Coding** regions that will be translated into protein
>
> **Noncoding** untranslated regions (UTRs) that have a regulatory role; and

Introns regions removed by the splicing machinery.[11]

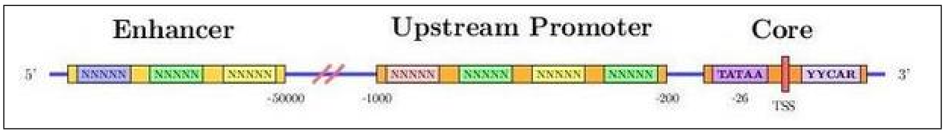


Figure 1.1: An eukaryotic promoter region. Note the short promoter elements (binding sites) and the downstream TATA box.

While the promoter region can be considered part of the gene as is responsible of its expression, the computational methods to predict this regulatory region are still unreliable. For this reason, most gene prediction programs do not attempt to use information on promoters for the location and prediction of the downstream eukaryotic genes, and so is done in the research presented here (see Results).

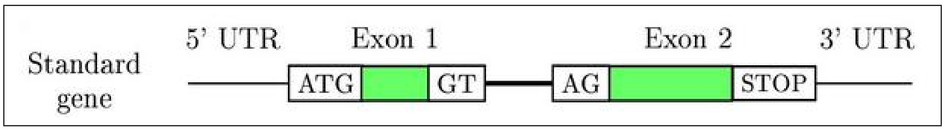


Figure 1.2: Gene structure for an standard eukaryotic two exons gene, partially coding and noncoding.

This complex gene structure and its biological processing is not yet well understood. In other words, we do not really know

1.5.2 Computational methods

Therefore, it is far from being properly defined by computational approaches. For this reason, in the gene finding field, gene structure prediction usually refers only to the prediction of coding exons.[12] This current limitation is due to both the lack of well-defined signals delimiting non-coding exons and the non-protein-biased codon usage in them. In other words, promoters and non-coding exons are not usually predicted by most simple methods. In short, gene prediction programs can be classified into two major classes:

ab initio based on signal and codon bias identification; and

Homology based, only or in addition, on sequence similarity search.

However, gene structure prediction in eukaryotes by computational methods is far from perfect (Burset and Guigó, 1996). The main common steps for most gene prediction programs are:

1. To find all potential exons in a given DNA sequence
2. To assess (score) the likelihood of each exon of being real
3. To build genes that maximize this likelihood

[11] Should be noted that intronless genes also exist.

[12] Most gene prediction programs simply call exons to the coding exons they predict.

Each program may find, assess and build genes in its own way and using different computational techniques (neural networks, hidden markov models, rule-based models and others). At the same time, the biological knowledge embedded in each program is also different. Both things account for the diversity in accuracy measures. In conclusion, the precise identification of the exon/intron structures of genes in genomic DNA sequences is still an open problem.

Signals

Several exon-defining signals have been found to be highly conserved, at least in a few nucleotides:

Start site usually the AUG codon.

Splice site GT is the donor site and AG the acceptor.

Termination one of the three codons: UAA, UAG and UGA.

In addition, these short signals are within a context that shows a biased usage of nucleotides. Therefore, computational approaches exist to make use of this characteristic (see Methods).

Coding potential

At the core of all gene identification programs there exist one or more coding measures. A coding statistics can be defined as a function that computes given a DNA sequence a real number related to the likelihood that the sequence is coding for a protein. Many of such measures have been published in the literature, for example, codon usage bias, base compositional bias between codon positions, or periodicity in base occurences. See Methods for a better introduction.

Accuracy

The following table summarizes the prediction accuracy for several popular gene prediction programs:

	Nucleotide			Exon					Gene						
Program	Sn	Sp	CC	Sn	Sp	$\frac{Sn+Sp}{2}$	ME	WE	Sn	Sp	$\frac{Sn+Sp}{2}$	MG	WG	JG	SG
refGene	0.86	0.96	0.91	0.88	0.95	0.91	0.10	0.02	0.66	0.76	0.71	0.16	0.04	1.00	1.00
GENSCAN	0.88	0.49	0.65	0.72	0.40	0.56	0.12	0.51	0.07	0.04	0.05	0.12	0.53	1.17	1.10
FGENESH++	0.94	0.68	0.79	0.90	0.67	0.78	0.05	0.29	0.58	0.35	0.47	0.09	0.42	1.00	1.05
Ensembl	0.82	0.82	0.81	0.73	0.76	0.75	0.19	0.15	0.43	0.34	0.38	0.09	0.22	1.00	1.07
geneid	0.82	0.63	0.71	0.70	0.54	0.62	0.17	0.36	0.13	0.08	0.10	0.13	0.46	1.12	1.13
geneid$_H$	0.89	0.61	0.73	0.68	0.47	0.57	0.10	0.37	0.13	0.07	0.10	0.12	0.45	1.06	1. 26
geneid$_{H'}$	0.89	0.67	0.77	0.67	0.51	0.59	0.10	0.31	0.13	0.07	0.10	0.13	0.38	1.07	1 .29
geneid$_E$	0.94	0.69	0.80	0.92	0.69	0.81	0.04	0.29	0.68	0.38	0.53	0.09	0.47	1.00	1.05
geneid$_{H'+E}$	0.94	0.74	0.83	0.92	0.71	0.82	0.04	0.25	0.68	0.41	0.55	0.09	0.42	1.00	1.0 5
SGP2	0.84	0.68	0.75	0.72	0.58	0.65	0.14	0.31	0.15	0.09	0.12	0.13	0.41	1.08	1.17
TWINSCAN	0.84	0.70	0.76	0.77	0.63	0.70	0.11	0.26	0.16	0.14	0.15	0.17	0.24	1.14	1.14

Table 1.1: Gene prediction accuracy in the human chromosome 22.

Standard prediction

In this work, standard prediction will be understood as prediction of genes delineated by standard exon-defining signals. An introduction to it based on the geneid program can be found in Methods section.

Non-Standard prediction

In this work, non-standard prediction will be understood as the prediction of selenoprotein genes. In other words, the prediction of genes interrupted by in-frame TGAs. For an in-deep conceptual and methodological introduction to this problem based on `geneid` see sections **??** in page **??** and **??** in page **??**.

1.6 RNA structure prediction

RNA is no longer seen as merely the passive intermediary messenger between DNA genes and proteins. Information storage, transport and control are required for many tasks of the cell and RNA is an optimal molecule for this. For such regulatory steps and other specific functions, specific structures in the mRNA, so-called mRNA elements, motifs or structures, have evolved. Furthermore, many mRNA-independent non-coding RNA molecules exist[13] which adopt sophisticated three-dimensional structures, and some even catalyse biochemical reactions (ribozymes).

There is a growing list of examples that underscores the roles that mRNA motifs, with and without well-defined sequence conservation and/or secondary structure, play in controlling the genetic repertoire of cells and developing organisms. These short regulatory segments of an mRNA can lie anywhere, that is, in the UTRs, coding exons or introns and be functional at different steps of the mRNA processing pathway. Transcription, splicing, polyadenylation, transport to the cytoplasm and protein translation are among the biological processes under the control of these signals. Examples of RNA signals related to the translation of mRNAs, our interest here, include (but are not limited): 1) the Shine-Delgarno sequence in prokaryotes which marks the translation start site (5-9 nts of the AUG. It is complementary and bridges to the 3'end of 16S rRNA); 2) the Iron-responsive elements (IREs) in the 5'UTR or 3'UTR of several mRNAs bind the iron regulatory protein (IRP) and precludes translation or provides stability of the mRNA, respectively; 3) an element in mouse protamine 2 mRNA 3'UTR temporally represses its translation by interaction with a 18kDa protein (Kwon and Hecht, 1993); and, of course 4) the Selenocysteine Insertion Sequence (SECIS) required for the recoding of the UGA stop codon as selenocysteine. A wide introduction to RNA motifs and computational techniques to their identification is Dandekar (2002).

1.6.1 RNA structures

From a computational point of view, it is time to raise the problem of how these RNA structures can be studied. It is seems now clear that, in order to capture the complexity of these RNA molecules, information should be extracted from a spectrum of possible levels of resolution. This is because for given signal, though one level can be ill-defined the others may still be used to define the molecule. In example, one signal may not be conserved at the nucleotide sequence while having a strong secondary structure. Usually, we study signals at:

Primary structure nucleotide sequence of the molecule. Since the advent of rapid methods for sequencing DNA there has been an exponential growth in nucleic acid sequence information and we usually have many instances of the plain sequence of a particular motif in a variety of organisms. This allows to recognize important bits of conserved sequence, which may have a functional role.

Secondary structure RNA molecules have the potential to form into helical structures wherever there are two parts of the sequence that are complementary. Hydrogen bonds are possible between C-G and A-U pairs, and also the less stable G-U pairs are relatively common. Furthermore, RNA double helices are not composed solely of standard Watson-Crick pairs and this perturbations can be functionally important in adopting unusual structures, leading to anchoring sites for metals or proteins. This seems to be the case for selenoproteins, because the SPS2 protein is thought to bind the quartet of non-Watson-Crick pairs in the SECIS element. However, a secondary structure diagram tells us only about the base pairing pattern and gives us no information about the tertiary structure of the molecule.

[13] They exist indeed, but the growing family of small non-coding RNAs that hang around in the cell with a regulatory role will not be a matter of discussion here. The focus will be only in RNA signals on a mRNA sequence that regulate the translational step.

Tertiary structure means the atomic coordinates, as well the relative spatial orientation of the structural elements. Experimentally much less is known about RNA tertiary structure than protein tertiary structure. Currently, it is still unclear how to use this information in genome-wide analysis of RNA structures.

Once the conserved sequence and/or its basepairings and/or its spatial disposition is known from collected examples, it should be embedded into a model in the computer. Naively, a computable model of a RNA signal is a descriptor (a list of characteristics in the simplest design) of the outstanding features of the signal in a way which is parseable by a computer program and can provide novel instances of the element along any nucleotide sequence queried. Clearly, any such model designed to describe and reveal unidentified regulatory RNA signals should take into account as much as biological information as possible from these three levels of description. But, secondary structure in particular, turns out to be and adequate shape definition for RNA: it covers the dominant part of the three-dimensional folding energies, and it is frequently conserved in evolution, sometimes together with a few tertiary interactions. RNA helices are thermodynamically strong and it is thought that stable stable secondary structures form first and that tertiary structures form afterwards as the molecule is able to bend around the flexible single stranded regions. This suggests that many of the relevant intermolecular interactions are indeed strongly influenced by the secondary structure. For this reason, secondary structure information is so reliable in order to infer evolutionary homology between positions in an alignment, where using only similarity of the sequence of symbols could lead to error. The best putative secondary structure can be inferred from experimental information, comparative sequence analysis (CSA) and thermodynamic predictions. In addition, for many RNA elements, sequence conservation is low and the tridimensional structure is either unknown or difficult to use as a signal model[14]. Therefore, the secondary structure information becomes a trade off between Sn/Sp and our capacity of describing an RNA structure and using such a descriptor as a template of the signal.

1.6.2 Computational methods

According to this, how models of these RNA structures can be created? and, moreover, how can databases be searched, using these models as a guide[15], to find unknown instances of these RNA motifs? as mentioned, it is first necessary to compile a list of the known family members including important sequence and structure features from literature and direct experimental data. One broad and simple distinction between template-dependent models is:

Deterministic those methods that input a fix search pattern (model) and output a match/no match result. The key is to understand RNA structures and sequences as a type of elements in a descriptive language. As in human language, both single characters (nucleotides) and higher-order structures (stem-loop structures and more complex tertiary interactions as pseudoknots) are important and can be represented. Thus, this combined motif search depends on how carefully our description of the RNA element is done. Examples of programs that follow this approach are `patscan`, `patsearch`, `rnamot` and `rnabob`; and

Probabilistic those methods that input a representation of the signal based on assigning probabilities to the different elements that compose the element and to the relationship between them. One such methods are the covariance models, where a tree connected by transition probabilities is derived (and generates) from the known sequences of a family. Searches for unrecognized instances are usually very sensitive, though, its limitation is currently slow speed, as only 10-20 base pairs per second can be searched. Examples of programs that implement these types of algorithms are `cove` and ?.

However, so far, these model-oriented programs have only produced as results stretches of sequences which, according to base-pairing rules and similarity to our model, may basepair in the particular way of interest. But they may not. Clearly, there is a weak link between these predicted structures

[14] In the years to come, our ability to high-throughput produce and handle these data will make the difference. However, nowadays our understanding of the link between sequence and RNA spatial structure is weak and progress in the area of RNA *ab initio* modeling is otherwise slow.

[15] Template-free types of searches will not be discussed here.

and biologically real ones. Two possible ways to face this problem are: 1) to assess the thermodynamic stability of the potential RNA elements. Programs exist that measure the Gibbs free energy of the predicted structures, which can be understood as a measure of the tendency of a process to happen, in this case, of adopting a specific secondary folding. Well-known programs are `RNAfold` from the `Vienna RNA package` and `mfold`; and 2) to analyzed the conservation at the sequence and structural level of the predicted element across genomes (see section 1.7).

Ultimately, the candidate RNAs which have passed all these analysis steps should now be tested by direct or indirect biochemical and biological assays. It is at this point where a predicted RNA structure becomes reliable.

1.7 Comparative genomics

The similarity of molecular mechanisms among studied organisms strongly suggests that all organisms on earth had a common ancestor. Thus, any set of species is related and this relationship is called phylogeny. In this way, genes, genomes and organisms are related but at different phylogenetic distances between them. This distance, expressed roughly at the degree of change at the sequence level, is different according to the the molecule type, its particular function, specific contingent events and others. This particularity of the unity and relation behind sequences brings into play comparative genomics, which is the analysis and comparison of genomes from different species.

1.7.1 Conserved regions

The purpose of comparative genomics is to gain a better understanding of how species have evolved and to determine the function of genes and noncoding regions of the genome. Researchers have learned a great deal about the function of human genes by examining their counterparts in simpler model organisms such as the mouse. Genome researchers look at many different features when comparing genomes: sequence similarity, gene location, the length and number of coding regions within genes, the amount of noncoding DNA in each genome, and highly conserved regions maintained in organisms as simple as bacteria and as complex as humans.

1.7.2 Computational methods

Recently, the importance of sequence comparisons between genomes of different species to locate functional domains conserved through evolution (protein coding among them) has been underscored, and new bioinformatics methodologies have been developed to infer protein coding genes from sequence comparisons of the genomes of two different species developed (Batzoglou et al., 2000; Bafna and Hudson, 2000; Wiehe et al., 2001; Korf et al., 2001, Novichkov et al., 2001), which appear to lead to highly accurate predictions. The rationale is that functional regions (protein-coding among them) are more conserved than non-functional ones across the DNA sequence of genomes from different species. In this respect and for the main purpose of this dissertation, the computational prediction of selenoprotein genes, it should be noted that:

genes gene structures, at least between close species, are alike in selenoprotein families; and

proteins selenoproteins are highly similar in their amino acidic sequence across the whole eukaryotic lineage; and

SECIS though only in closer phylogenetic relationships (eg. human-rodent), SECIS elements are similar in sequence.

Comparisons at these three levels of homology are used in the research presented here (See results).

Objectives

The research in this PhD was initially targeted, in late 1999, to the goals enumerated below. In what follows, they are described and an account of their achievement status given.

1. To find all selenoprotein genes that define the eukaryotic selenoproteome by analyzing the genomic sequence data through bioinformatics means. This includes the development of novel computational methods for the prediction of selenoprotein genes. That is to say, the prediction of:
   (a) Genes containing a UGA in-frame encoding Sec
   (b) The SECIS RNA secondary structures in the 3′UTR of these genes
2. To describe these genes in terms of:
   (a) Gene structure
   (b) SECIS structure
   (c) Protein sequence
   (d) Gene families
   (e) Non-selenium-containing homologs
3. To establish the distribution and phylogeny of these selenoprotein families across the eukaryotic domain
4. To propose the evolutionary events responsible for the current selenoproteome map
5. To provide both, the novel selenoprotein data and bioinformatics tools, to the research community

These objectives were established based on data and knowledge of that time, and were intended to explore very basic questions about selenoproteins. These goals have been accomplished to different degrees as related below. Thus, several of these points should be considered as ongoing work and yet many questions, old and new[1], remain unanswered.

For the first goal, it was then generally believed, maybe in a naive interpretation of the data available, that mammals had the most complete selenoproteome, while other less complex organisms only shared subsets of it. In other words, mammals had accumulated or conserved the majority of selenoprotein genes. In consequence, a reasonable hypothesis was that the analysis of several mammalian genomes (specifically of the human and rodent ones because their release was soon to come), would provide the characterization of all, or almost all, eukaryotic selenoproteins. The next step in this rational was to map this set of proteins in other eukaryotic but nonmammalian genomes to get the full picture of the selenoproteinin eukaryotes and its non-selenium homologs. We will show here that, as years have passed, this hypothesis has not been confirmed, and, on the contrary, the opposite may be true. That is to say, that species-specific selenoprotein indeed exist. Therefore, in the present work, this initial goal could only be fulfilled to the rate and the pace at which genome sequences are released and, in consequence,

[1] Among others, this work has raised the intriguing issue of the differential usage of Se among proteins and organisms, close and distant

the achievement of the subsequent goals has been delayed till the analysis of a more representative set of species is done.[2]

None of this would have been possible without the development of a variety of computational techniques able to deal with the oddities of the selenoprotein genes. We are talking about *ad hoc* gene prediction software, RNA secondary structure prediction and comparative genomics approaches. These tools are described at length in the Methods section.

For the second point, the description of selenoproteingenes in terms of their structure and others, we have compiled such information for the majority of selenoprotein in the genomes analyzed. However, the study of the relationships between the different selenoprotein families is still preliminary.

As for the third goal, we provide, to our knowledge, the most comprehensive and updated account for the distribution of selenoprotein genes and their Cys-homologs across the eukaryotic domain (the so-called selenoproteome map). Though, many other wider-spread eukaryotic genomes await analysis.

In relation to the fourth aim, the ordered enumeration of the evolutionary events responsible of the current selenoproteome map, no claim is done at this stage. To face this question, the focus should be shifted to the evolution of the genetic code, at the codon, amino acid and tRNA levels, as well as to the distribution of the usage of Se among proteins and organisms. This work has contributed to the better knowledge of the latest. Therefore, with such partial view, we will treat here this aspect only in a descriptive manner, for the sake of completeness and with no intention of elucidation.

For the last objective, the dissemination of data, we have made available all programs and sequence data with no restriction through our own web service. However, this is not enough. We plan to incorporate the selenoprotein annotation on the major databases.

In conclusion, we and others have developed various computational approaches for the prediction of selenoprotein genes. When applied on several genomes, novel selenoprotein families were uncovered which distributed in a mosaic fashion across the eukaryotic lineage. In addition, we believe that other taxa-specific selenoproteins probably exist.

[2] We currently have started, among others, the analysis of worms and birds.

Results

In this section, the literature consequence of this PhD is presented chronologically and in the form and format of research papers. Thus, the scientific results contributed by this work can directly be learned from the original publications.

3.1 Castellano et al., EMBO reports, 2, 697-702 (2001)

In this paper, we present a novel computational approach to predict genes with a TGA in-frame in coordination with the existence of a suitable SECIS element downstream of such a gene. A modified version of the `geneid` was used for gene prediction and a combination of the `patscan` patter-matcher and the RNAfold program for the assessment of SECIS structures. When applied to the *D. melanogaster* genome, two novel selenoprotein families were found, termed SelH[1] and SelK[2].

- Articles: abstract, full text and pdf
- Supplementary material: additional data
- Datasets: data and software

[1] Previously known as SelM.

[2] Previously known as SelG.

3.2 Kryukov et al., Science, 300, 1439-1443 (2003)

In this work, we provide a comprehensive analysis of the human and rodent selenoproteomes. Through several independent bioinformatics approaches, we believe that all, or almost all, selenoprotein genes in these genomes have been found. Novel selenoprotein families were named SelI, SelO, SelS, SelV and GPx6 (which has Cys in rodents). Therefore, the human selenoproteome consist of 17 distinct families.

- Article: abstract, full text and pdf
- Supplementary material: additional data
- Datasets: data and software

nonmotor tasks. Eyeblink conditioning requires that an animal not only learn the association between a neutral and an aversive stimulus; it is only adaptive if the animal also learns when to expect the aversive stimulus. More generally, the consolidation of motor skills centers on learning to specify the precise timing between successive movements. Cerebellar ataxia is characterized as a disruption in the timing of these events (*3*, *26*, *28*), rather than as a loss of the conceptual knowledge for actions observed in apraxia. Similarly, comparing the duration of two successive events requires a judgment of whether the second event occurred earlier or later than expected. Our results extend previous theories concerning the role of the cerebellum in temporal processing, indicating that this function is limited to tasks that require an explicit specification of the timing of behaviorally meaningful events.

References and Notes

1. R. B. Ivry, S. W. Keele, H. C. Diener, *Exp. Brain Res.* **73**, 167 (1988).
2. R. B. Ivry, H. S. Gopal, in *Attention and Performance XIV*, D. E. Meyer, S. Kornblum, Eds. (MIT Press, Cambridge, MA, 1993), pp. 771–802.
3. D. Timmann, S. Watts, J. Hore, *Exp. Brain Res.* **130**, 441 (2000).
4. R. B. Ivry, S. W. Keele, *J. Cogn. Neurosci.* **1**, 136 (1989).
5. J. A. Mangels, R. B. Ivry, N. Shimizu, *Brain Res. Cogn. Brain Res.* **7**, 15 (1998).
6. H. Ackermann, S. Graeber, I. Hertrich, I. Daum, *Brain Lang.* **60**, 323 (1997).
7. I. Daum, M. M. Shugens, H. Ackerman, W. Lutzenberger, *Behav. Neurosci.* **107**, 748 (1993).
8. H. Topka, J. Valls-Sole, S. G. Massaquoi, M. Hallett, *Brain* **116**, 961 (1993).
9. D. S. Woodruff-Pak, M. Papka, R. B. Ivry, *Neuropsychology* **10**, 443 (1996).
10. M. Jueptner *et al.*, *Neurology* **45**, 1540 (1995).
11. R. Kawashima *et al.*, *J. Neurophysiol.* **83**, 1079 (2000).
12. H. Ackermann, A. Riecker, K. Mathiak, M. Erb, W. Grodd, D. Wilddruber, *Neuroreport* **12**, 4087 (2001).
13. R. B. Ivry, *Int. Rev. Neurobiol.* **41**, 556 (1997).
14. J. Jackson, J. A. Michon, in *Tutorials in Motor Neuroscience*, J. Requin, G. Stelmach, Eds. (Kluwer, Norwell, MA, 1991), NATO Advanced Study Institute series, vol. 62, pp. 169–198.
15. R. B. Ivry, *Ann. NY Acad. Sci.* **682**, 214 (1993).
16. F. Mussa-Ivaldi, in *Biomechanics and Neural Control of Posture and Movement*, J. M. Winters, P. E. Crago, Eds. (Springer-Verlag, New York, 2000), pp. 325–333.
17. We focus on measures of variable error, operationalized as the standard deviation of the cycle durations. This measure has been commonly used in neuropsychological studies [e.g. (*1*)] to assess the contribution of different neural structures to timing based on the assumption that lesions add noise to the system.
18. S. D. Robertson *et al.*, *J. Exp. Psychol. Hum. Percept. Perform.* **25**, 1316 (1999).
19. H. N. Zelaznik, R. M. Spencer, J. Doffin, *J. Mot. Behav.* **32**, 193 (2000).
20. H. N. Zelaznik, R. M. Spencer, R. B. Ivry, *J. Exp. Psychol. Hum. Percept. Perform.* **28**, 575 (2002).
21. By "explicit," we mean that the action representation includes the specification of a temporal goal, as opposed to being a conscious representation as this term implies in the memory literature.
22. Materials and methods are available as supporting material on *Science* Online.
23. P. Killeen, N. Weiss, *Psychol. Rev.* **94**, 455 (1987).
24. Variability due to linear drift in cycle duration was removed by detrending (*34*). A regression line was fit to the time series for each trial, and the residuals were pooled across trials for each participant in each condition. The standard deviation of these residuals was divided by the mean cycle duration to obtain the coefficient of variation scores.
25. One concern with our within-subject comparison is that the difference between hands may be related to other factors, such as handedness. However, the lesions affected the dominant hand for five of the six patients; thus, any effect of handedness would work against our hypothesis.
26. G. Holmes, *Brain* **62**, 1 (1939).
27. A. J. Bastian, T. A. Martin, J. G. Keating, W. T. Thach, *J. Neurophysiol.* **76**, 492 (1996).
28. J. Hore, B. Wild, H. C. Diencer, *J. Neurophysiol.* **65**, 563 (1991).
29. S. W. Kennerley, J. Diedrichsen, E. Hazeltine, A. Semjen, R. B. Ivry, *Nature Neurosci.* **5**, 376 (2002).
30. M. T. Turvey, in *Perceiving, Acting, and Knowing: Toward an Ecological Psychology*, R. Shaw, J. Bransford, Eds. (Erlbaum, Hillsdale, NJ, 1977), pp. 211–265.
31. N. Hogan, T. Flash, *Trends Neurosci.* **10**, 170 (1987).
32. R. J. van Beers, P. Baraduc, D. M. Wolpert, *Phil. Trans. R. Soc. Lond.* **357**, 1137 (2002).
33. E. Burdet, R. Osu, D. W. Franklin, T. E. Milner, M. Kawato, *Nature* **414**, 446 (2001).
34. D. Vorberg, A. Wing, in *Handbook of Perception and Action, Vol. 2: Motor Skills*, H. Heuer, S. W. Keele, Eds. (Academic Press, San Diego, CA, 1996), pp. 181–262.

35. Supported by NSF dissertation improvement grant no. 0121930 (R.M.C.S.) and by NIH grant nos. NS30256, NS40813, and NS17778. We thank S. Keele and S. Grafton for comments on this manuscript.

Supporting Online Material
www.sciencemag.org/cgi/content/full/300/5624/1437/DC1
Materials and Methods
Fig. S1
Tables S1 to S4

19 February 2003; accepted 30 April 2003

Characterization of Mammalian Selenoproteomes

Gregory V. Kryukov,[1] Sergi Castellano,[2] Sergey V. Novoselov,[1] Alexey V. Lobanov,[1] Omid Zehtab,[1] Roderic Guigó,[2] Vadim N. Gladyshev[1]*

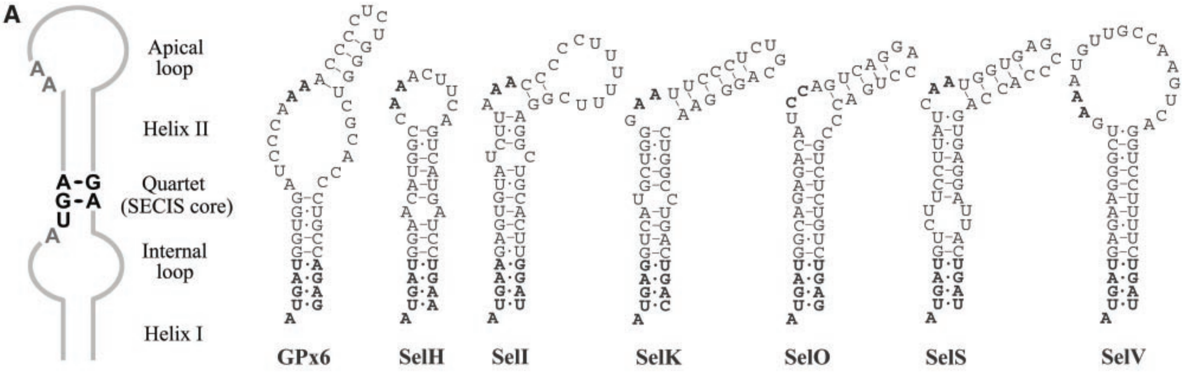
In the genetic code, UGA serves as a stop signal and a selenocysteine codon, but no computational methods for identifying its coding function are available. Consequently, most selenoprotein genes are misannotated. We identified selenoprotein genes in sequenced mammalian genomes by methods that rely on identification of selenocysteine insertion RNA structures, the coding potential of UGA codons, and the presence of cysteine-containing homologs. The human selenoproteome consists of 25 selenoproteins.

In the universal genetic code, 61 codons encode 20 amino acids, and 3 codons are terminators. However, the UGA codon has a dual function in that it signals both the termination of protein synthesis and incorporation of the amino acid selenocysteine (Sec) (*1–3*). Available computational tools lack the ability to correctly assign UGA function. Consequently, there are numerous examples of misinterpretations of UGA codons as both Sec codons (*4*) and terminators (*5*, *6*), including annotations of the human genome (*7*, *8*), where no selenoproteins have been correctly predicted. With 18 human selenoprotein genes previously discovered (*3*), the estimates of the actual number of such genes vary greatly (*9*). All previously characterized selenoproteins except selenoprotein P (*10*) contain single Sec residues that are located in enzyme-active sites and are essential for their activity. Thus, misidentification of UGA codons leads to a loss of crucial biological and functional information. Sec is cotranslationally incorporated into nascent polypeptides in response to UGA codons when a specific stem-loop structure, designated the Sec insertion sequence (SECIS) element, is present in the 3′ untranslated regions (UTRs) in eukaryotes and in archaea, or immediately downstream of UGA in bacteria (*1*, *11–13*). Trans-acting factors, including Sec tRNA, Sec-specific elongation factor, selenophosphate synthetase (SPS), Sec synthase, and a SECIS-binding protein, are also required for Sec biosynthesis and insertion (*1*, *3*, *13–15*). Most known selenoprotein genes have homologs, in which Sec is replaced with cysteine (Cys). However, these proteins are poor catalysts as compared with selenoproteins (*3*).

We hypothesized that the UGA dual-function problem could be solved by identifying selenoprotein genes in sequenced genomes and assigning terminator functions to the remaining in-frame UGAs. The requirement of SECIS elements for Sec insertion and the presence of Cys-containing homologs of selenoproteins suggested two independent bioinformatics methods for selenoprotein identification. In addition, we used an ob-

[1] Department of Biochemistry, University of Nebraska, Lincoln, NE 68588–0664, USA. [2] Grup de Recerca en Informàtica Biomèdica, Institut Municipal d'Investigació Mèdica, Universitat Pompeu Fabra, Centre de Regulació Genòmica, Doctor Aiguader 80, 08003 Barcelona, Catalonia, Spain.

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



B

Selenoprotein	Chromosomal location (number of exons)	Sec location in protein (length of protein)
15kDa	1p22.3 (5)	93 (162)
DI1	1p32.3 (4)	126 (249)
DI2	14q31.1 (2)	133 (265)
DI3	14q32	144 (278)
GPx1	3p21.31 (2)	47 (201)
GPx2	14q23.3 (2)	40 (190)
GPx3	5q33.1 (5)	73 (226)
GPx4	19p13.3 (7)	73 (197)
GPx6	**6p22.1 (5)**	**73 (221)**
H	**11q12.1 (4)**	**44 (122)**
I	**2p23.3 (10)**	**387 (397)**
K	**3p21.31 (5)**	**92 (94)**
M	22q12.2 (5)	48 (145)
N	1p36.11 (12)	428 (556)
O	**22q13.33 (9)**	**667 (669)**
P	5p12 (4)	59, 300,318,330, 345, 352, 367, 369, 376, 378 (381)
R	16p13.3 (4)	95 (116)
S	**15q26.3 (6)**	**188 (189)**
SPS2	-	60 (448)
T	3q24 (6)	36 (182)
TR1	12q23.3 (15)	498 (499)
TR2	3q21.2 (16)	655 (656)
TR3	22q11.21 (18)	522 (523)
V	**19q13.13 (6)**	**273 (346)**
W	19q13.32 (6)	13 (87)

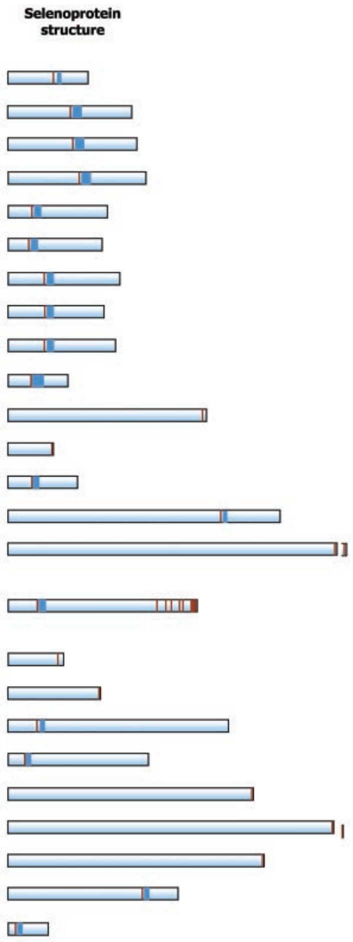


Fig. 1. Mammalian selenoprotein genes. (**A**) Mammalian SECIS element consensus and SECIS elements in newly unidentified human selenoprotein genes. Only the upper portions of SECIS elements are shown. (**B**) Human selenoprotein genes. Proteins are shown in alphabetical order and the newly identified genes are highlighted. On the right, relative lengths of selenoproteins are shown and Sec locations within the proteins are indicated by red vertical lines. The regions in selenoproteins that correspond to downstream α helices are highlighted.

servation that the strong codon bias characteristic of protein-coding regions extends beyond the UGA codon in selenoprotein genes. We previously developed two computer programs, SECISearch 1.0 and geneid, which were used to identify several new selenoprotein sequences (*16–18*), and related approaches have also been developed (*19*). However, these methods were insufficient in identifying selenoprotein genes in mammalian genomes because of their size and complexity.

Our SECIS-based method, as applied to mammalian genomes (fig. S1), consisted of the following principal steps (*20*): (i) We identified candidate SECIS elements in the human genome with SECISearch 2.0. This program analyzed structural and thermodynamic features of SECIS elements and was about 10 times more selective (with the same specificity) than the original version of SECISearch (*16*). (ii) We identified human/mouse and human/rat SECIS pairs with SECISblastn, a program that analyzed evolutionary conservation of mammalian SECIS elements. This program was based on our observation that human, mouse, and rat SECIS elements in orthologous selenoprotein genes exhibited detectable sequence similarity. SECISblastn provided an increase of about 100-fold in the specificity of genomic searches. (iii) We analyzed genomic sequences upstream of candidate SECIS elements with geneid (*18*), a gene prediction program that identified open reading frames (ORFs) that had high coding potential and that contained in-frame TGA codons. (iv) We analyzed predicted human selenoprotein genes with mammalian selenoprotein gene signature (MSGS) criteria (*21*), which screened selenoprotein homologs for the presence and conservation of ORFs, in-frame TGA codons, and SECIS elements.

Primary sequences of more than 95% previously characterized mammalian SECIS elements contain an adenosine that precedes the quartet of non–Watson-Crick base pairs, a TGA_GA motif in the quartet, and two adenosines in the apical loop or bulge (*12*) (the ATGA_AA_GA pattern) (Fig. 1A). In addition, in mammalian SelM SECIS elements, AA is replaced with CC (*22*) (the ATGA_CC_GA pattern). The SECISearch 2.0 screen of mammalian genomes using the ATGA_AA_GA

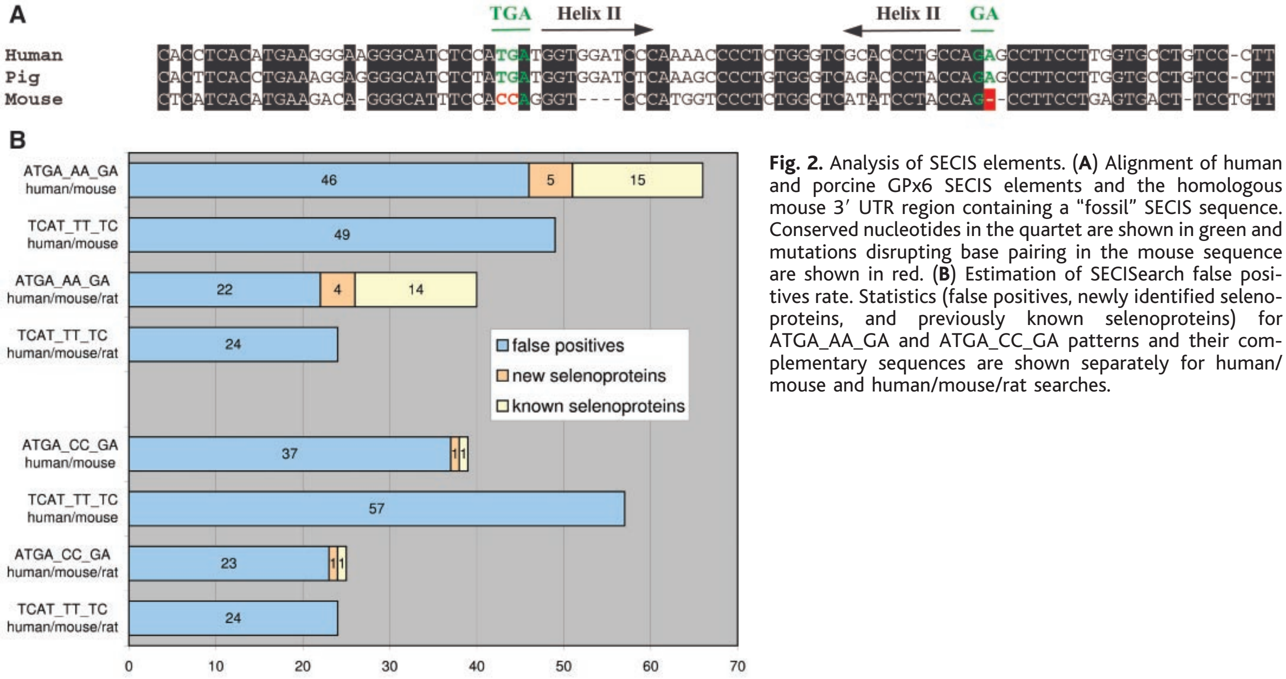


Fig. 2. Analysis of SECIS elements. (**A**) Alignment of human and porcine GPx6 SECIS elements and the homologous mouse 3′ UTR region containing a "fossil" SECIS sequence. Conserved nucleotides in the quartet are shown in green and mutations disrupting base pairing in the mouse sequence are shown in red. (**B**) Estimation of SECISearch false positives rate. Statistics (false positives, newly identified selenoproteins, and previously known selenoproteins) for ATGA_AA_GA and ATGA_CC_GA patterns and their complementary sequences are shown separately for human/mouse and human/mouse/rat searches.

pattern resulted in 7146 human structures. The SECISblastn analysis reduced the number of structures to 1031 human/mouse and 276 human/rat pairs, and subsequent use of contamination, shotgun redundancy, and repetitive element filters resulted in 56 unique human/mouse and 58 unique human/rat pairs, including 40 structures that were common to all three organisms. The geneid analyses of sequences upstream of candidate SECIS elements and a subsequent analysis with MSGS criteria reduced the set to 20 hits. Among these, 15 were already known human selenoproteins and 5 were novel selenoproteins, designated as SelH, SelI, SelK, SelS, and SelV (Fig. 1B, figs. S2 to S6, and figs. S10 and S11).

A similar computational screen using the ATGA_CC_GA pattern (*23*) detected a single true positive selenoprotein (SelM) and one novel selenoprotein (SelO) (Fig. 1, A and B; fig. S7; and figs. S10 and S11). Only two known human selenoprotein genes were not identified by these procedures: The *SPS2* gene was absent in the human genome assembly, whereas the thioredoxin reductase 2 (TR2) gene contained a SECIS element with a thymidine preceding the quartet, a structure that does not correspond to other known SECIS elements.

The 24 mammalian selenoproteins were subsequently examined for the presence of homologs. This analysis identified a 25th human selenoprotein, designated glutathione peroxidase 6 (GPx6) (figs. S8, S10, and S11), a close homolog of plasma GPx3. GPx6 was not identified in the SECISearch-based computational screen, because its mouse and rat orthologs had Cys in place of Sec and the corresponding genes lacked SECIS elements. Rat GPx6 was previously cloned as rat odorant-metabolizing protein (*24*). Homology analyses revealed a "fossil," nonfunctional SECIS element in the 3′ UTR of the mouse GPx6 gene, which contained mutations that disrupted the quartet and secondary structure (Fig. 2A). We also cloned the gene encoding porcine GPx6 and found that it had a SECIS element and encoded a selenoprotein. These data revealed that Sec, which was initially present in the mammalian GPx family, was replaced by Cys in rodent genes for GPx6.

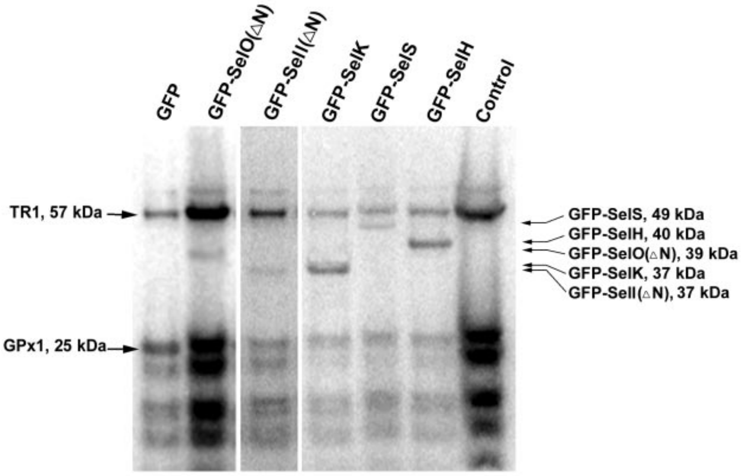


Fig. 3. Incorporation of selenium into newly identified mammalian selenoproteins. GFP-selenoprotein constructs were used for convenient visualization of signals, wherein the fusion proteins differed in size from endogenous selenoproteins. Also for convenient visualization, the N-terminal regions of SelO and SelI were deleted. After transfection into CV-1 cells, transfected and control cells were incubated with ^{75}Se[selenite] for 24 hours, the extracts were resolved by SDS–polyacrylamide gel electrophoresis, and the labeled selenoproteins were visualized with a PhosphorImager. Locations of transfected selenoproteins are indicated on the right, and locations of major endogenous selenoproteins (TR1 and GPx1) are on the left. The left lane (GFP) shows control transfection with GFP alone. The right lane (control) shows untransfected CV-1 cells. The five middle lanes show experiments with indicated selenoproteins. All five showed ^{75}Se-labeled bands of the size expected if TGA encoded Sec.

To estimate the number of false positives in the set of hits selected by SECISearch and SECISblastn, searches were performed using patterns that were complementary to the conserved SECIS sequences. The false positive rate with such patterns should be similar to that in the SECIS patterns, but the true positive rate with the complementary patterns should be zero. The difference between the number of SECIS candidates conforming to the major SECIS pattern, ATGA_AA_GA, and that of the complementary pattern corresponded approximately to the number of identified selenoprotein genes (Fig. 2B). Thus, the ability of our SECIS-based method to recognize known mammalian selenoproteins and to complete analyses of all other candidates indicates that all or almost all selenoproteins common to human and rodent genomes were identified by our procedures. In addition, neither the SECISearch analyses of human and mouse dbEST and pairwise searches of human/mouse genomes with altered SECIS patterns (*23*), nor the SECIS-independent searches for Sec/Cys pairs in homologous sequences (see below), revealed additional mammalian selenoproteins. The seven new human selenoproteins were either incorrectly predicted or not detected at all in Celera (*8*), National Center for Biotechnology Information (*7*), and Golden Path (*25*) human genome assemblies and annotations. In new as well as in known selenoproteins, Sec was located either upstream of an α helix or very close to the C terminus (Fig. 1B).

When the SECISearch-based method was applied to other eukaryotic genomes, we found neither selenoprotein genes nor Sec insertion machinery genes in yeast *Saccharomyces cerevisiae* or *Schizosaccharomyces pombe*, or in plant *Arabidopsis thaliana* genomes, whereas we could find only one and three already known selenoproteins in *Caenorhabditis elegans* and *Drosophila melanogaster* genomes, respectively (*26*) (fig. S12).

GPx6 and SelV were homologs of the previously characterized selenoproteins GPx1 and SelW, respectively, and shared a conserved Sec with these proteins. To validate the remaining five new selenoproteins, we demonstrated the incorporation of selenium into these proteins by metabolic ^{75}Se labeling of CV-1 cells that were transfected with selenoprotein constructs (Fig. 3). Analysis of the expression patterns of these selenoprotein genes revealed that SelH, SelI, SelO, SelS, and SelK mRNAs were present in a variety of tissues and cell types (*23*). However, the GPx6 mRNA was only detected in embryos and olfactory epithelium (Fig. 4A), and expression of SelV mRNA was restricted to testes (Fig. 4B), where it occurred in seminiferous

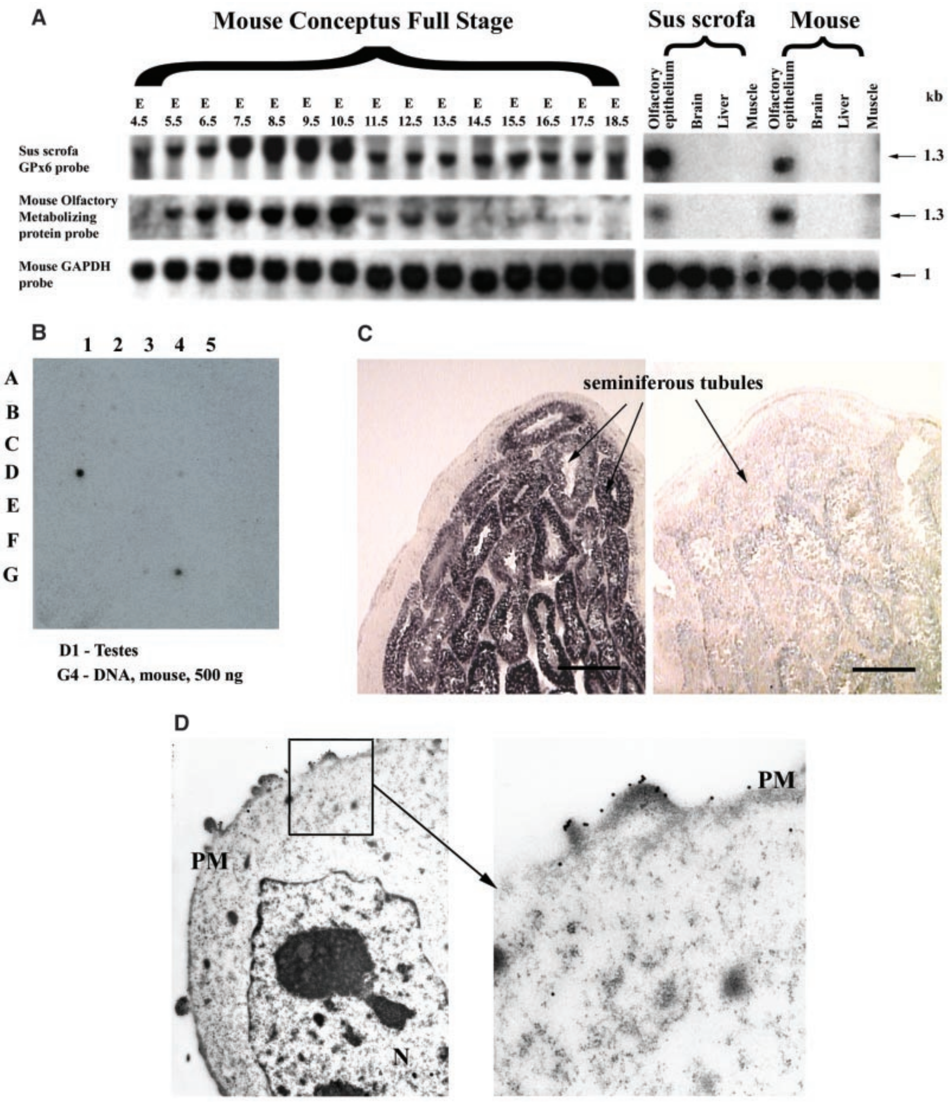


Fig. 4. Expression of mammalian selenoproteins. (**A**) GPx6 mRNA is expressed in embryos and olfactory epithelium. On the left, a mouse full-stage conceptus Northern blot (See-Gene, Del Mar, CA) was probed with pig GPx6, mouse GPx6, and glyceraldehyde-3-phosphate dehydrogenase cDNA probes. On the right, mRNA isolated from indicated mouse and pig tissues was probed as above. We observed no significant cross-hybridization with other GPx mRNAs, which also migrated differently than the 1.3-kb GPx6 mRNA on these northern blots. (**B**) SelV mRNA is expressed in testes. A mouse multiple-tissue blot was developed with a mouse SelV mRNA probe. Northern blots also revealed testes-specific expression (*23*). (**C**) In situ hybridization of SelV mRNA in seminiferous tubules. On the left, a SelV sense probe was used. On the right, a SelV antisense probe (control) was used. (**D**) SelS and SelK are plasma membrane proteins. A construct encoding SelS-GFP fusion protein was generated and transfected into NIH 3T3 cells, and the expressed protein was detected with antibodies to GFP by means of electron microscopy.

tubules (Fig. 4C). The secondary structure and protein organization predictions suggested that, like all previously characterized mammalian selenoproteins, GPx6, SelH, SelO, and SelV were globular proteins. However, SelK and SelS were predicted membrane proteins. We expressed fusions of SelK (*23*) and SelS (Fig. 4D) containing a C-terminal green fluorescent protein (GFP) tag in CV-1 cells and found that the fusion products did reside on the plasma membrane. Thus, SelK and SelS are the first known plasma membrane selenoproteins.

We next applied the Sec/Cys homology method to the human genome in two different ways. First, we predicted with geneid, and regardless of SECIS elements, all possible human genes that were interrupted by in-frame TGA codons. The predicted ORFs were extended from TGA to the next terminator signal and were analyzed by BLASTP and TBLASTN against all proteins predicted in completely sequenced eukaryotic genomes. This procedure was designed to identify sequences with homology in TGA-flanking regions, which either conserve TGA or replace TGA with TGC or TGT (Cyst codons). Second, we analyzed by TBLASTN all human proteins against all human expressed sequence tags to identify paralogs that contain TGA in place of a Cys codon. These two Sec/Cys homology approaches recognized the majority of selenoprotein genes that were found through SECIS elements but did not identify additional selenoproteins (*23*), providing additional evidence that all or virtually all mammalian selenoproteins have been identified in our work.

Dietary selenium plays an important role in cancer prevention (*27*), immune function (*28*), aging (*17*), male reproduction (*28*), and other physiological and pathophysiological processes (*29*). Selenoproteins are thought to be responsible for most biomedical effects of dietary selenium and are essential to mammals. Information on a set of human and mouse selenoproteins should provide the basis for future systematic analysis of mammalian selenoprotein functions.

References and Notes

1. A. Bock, *Biofactors* **11**, 77 (2000).
2. S. C. Low, M. J. Berry, *Trends. Biochem. Sci.* **21**, 203 (1996).
3. D. L. Hatfield, V. N. Gladyshev, *Mol. Cell. Biol.* **22**, 3565 (2002).
4. L. Cataldo *et al.*, *Mol. Reprod. Dev.* **45**, 320 (1996).
5. V. N. Gladyshev, K.-T. Jeang, T. C. Stadtman, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 6146 (1996).
6. M. J. Guimaraes *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 15086 (1996).
7. E. S. Lander *et al.*, *Nature* **409**, 860 (2001).
8. J. C. Venter *et al.*, *Science* **291**, 1304 (2001).
9. D. Behne *et al.*, *Biol. Trace Elem. Res.* **55**, 99 (1996).
10. R. F. Burk, K. E. Hill, *Bioessays* **21**, 231 (1999).
11. M. J. Berry *et al.*, *Nature* **353**, 273 (1991).
12. R. Walczak, E. Westhof, P. Carbon, A. Krol, *RNA* **2**, 367 (1996).
13. R. M. Tujebajeva *et al.*, *EMBO Rep.* **1**, 158 (2000).
14. D. Fagegaltier *et al.*, *EMBO J.* **19**, 4796 (2000).
15. P. R. Copeland *et al.*, *EMBO J.* **19**, 306 (2000).
16. G. V. Kryukov, V. M. Kryukov, V. N. Gladyshev, *J. Biol. Chem.* **274**, 33888 (1999).
17. M. J. Martin-Romeo *et al.*, *J. Biol. Chem.* **276**, 29798 (2001).
18. S. Castellano *et al.*, *EMBO Rep.* **2**, 697 (2001).
19. A. Lescure, D. Gautheret, P. Carbon, A. Krol, *J. Biol. Chem.* **274**, 38147 (1999).
20. Materials and methods are available as supporting material on *Science* Online.
21. G. V. Kryukov, V. N. Gladyshev, *Methods Enzymol.* **347**, 84 (2002).
22. K. V. Korotkov, S. V. Novoselov, D. L. Hatfield, V. N. Gladyshev, *Mol. Cell. Biol.* **22**, 1402 (2002).
23. G. V. Kryukov *et al.*, data not shown.
24. T. N. Dear, K. Campbell, T. H. Rabbitts, *Biochemistry* **30**, 10376 (1991).
25. J. W. Kent, D. Haussler, *Genome Res.* **11**, 1541 (2001).
26. G. V. Kryukov, V. N. Gladyshev, unpublished data.
27. G. F. Combs Jr., L. C. Clark, B. W. Turnbull, *Biofactors* **14**, 153 (2001).
28. F. Ursini *et al.*, *Science* **285**, 1393 (1999).
29. M. P. Rayman, *Lancet* **356**, 233 (2000).
30. We thank D. L. Hatfield for helpful discussions and Y. Zhou for assistance with microscopy. Supported by NIH GM61603 (to V.N.G.) and Ministerio de Ciencia y Tecnologia BIO2000-1358-C02-02 (to R.G.). S.C. is the recipient of a predoctoral fellowship from Generalitat de Catalunya.

Supporting Online Material

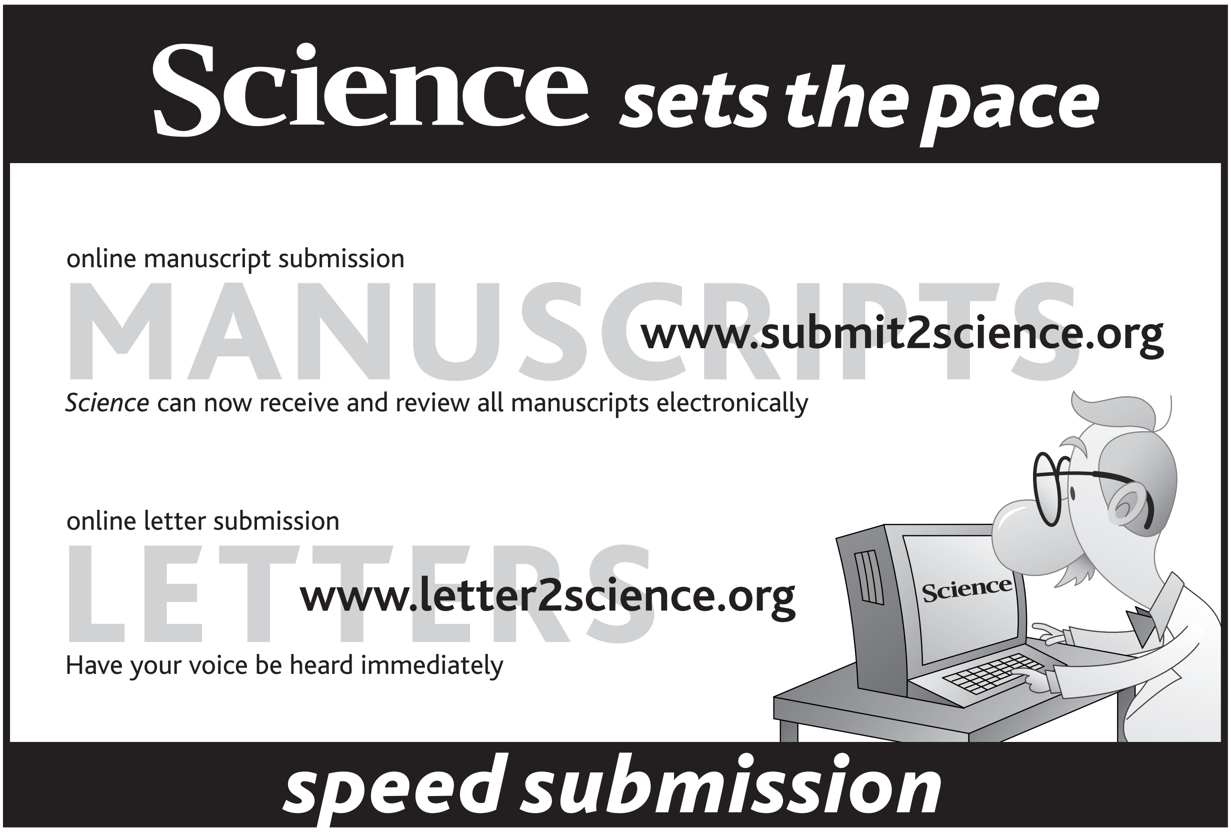
www.sciencemag.org/cgi/content/full/300/5624/1439/DC1

Materials and Methods

Figs. S1 to S13

References

14 February 2003; accepted 24 April 2003



3.3 Castellano et al., EMBO reports, 5, 71-77 (2004)

In this article, by means of a human-fish comparative gene prediction method, a new selenoprotein family, termed SelU, in the *Takifugu* genome was unveiled. In human, these proteins bear cysteine. This work, deserved a literature report at EMBO reports (Driscoll and Chavatte, 2004).

- Article: abstract, full text and pdf
- Supplementary material: additional data
- Datasets: data and software

3.4 Jaillon et al., Submitted (2004)

Finally, in this publication, we put at work the experience gained in the search of selenoprotein genes across eukaryotic sequences and we reannotate the compact genome of the puffer fish *T. nigroviridis*. The selenoproteome of *Tetraodon* consists of 19 selenoproteinfamilies, two more than in human. Besides 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. Only the selenoproteins section of the paper is shown in the next page.

- Database
- Datasets: data ans software

Analysis of the draft sequence of the compact *Tetraodon nigroviridis* genome provides new insights into vertebrate evolution

Olivire Jaillon,[1] Jean-Marc Aury,[1] Jean-Louis Petit,[1] Nicole Stange-Thomann,[2] Evan Mauceli,[2] Laurence Bouneau,[1] Frédrric Brunet,[3] Cécile Fischer,[1] Catherine Ozouf-Costaz,[4] Alain Bernot,[1] Sophie Nicaud,[1] David Jaffe,[2] Sheila Fischer,[2] Georges Lutfalla,[5] Carole Dossat,[1] Béatrice Segurens,[1] Corinne Dasilva,[1] Marcel Salanoubat,[1] Michael Levy,[1] Nathalie Boudet,[1] Sergi Castellano,[6] Véronique Anthouard,[1] Claire Jubin,[1] Vanina Castelli,[1] Michael Katinkam,[1] Benoît Vacherie,[1] Christian Biémont,[7] Zineb Skalli,[1] Laurence Cattolico,[1] Julie Poulain,[1] Simone Duprat,[1] Philippe Btottierm,[1] Jean-Pierre Coutanceau,[4] Jérôme Gouzy,[8] Gení Parra,[6] Guillaume Lardier,[1] Charles Chapple,[6] Kevin J. McKernan,[9] Paul McEwan,[9] Stephanie Bosak,[9] Jean-Nicolas Volff,[1]0 Roderic Guigó,[6] Mike Zody,[2] Jill Mesirov,[2] Kerstin Lindblasd-Toh,[2] Bruce Birren,[2] Chad Nusbaum,[2] Daniel Kahn,[8] Marc Robinson-Rechavi,[3] Vincent Laudet,[3] Vincent Schachter,[1] Francis Quetier,[1] William Saurin,[1] Claude Scarpelli,[1] Patrick Wincker,[1] Eric S. Lander,[2] Jean Weissenbach[1] and Hugues Roest Crollius[1]

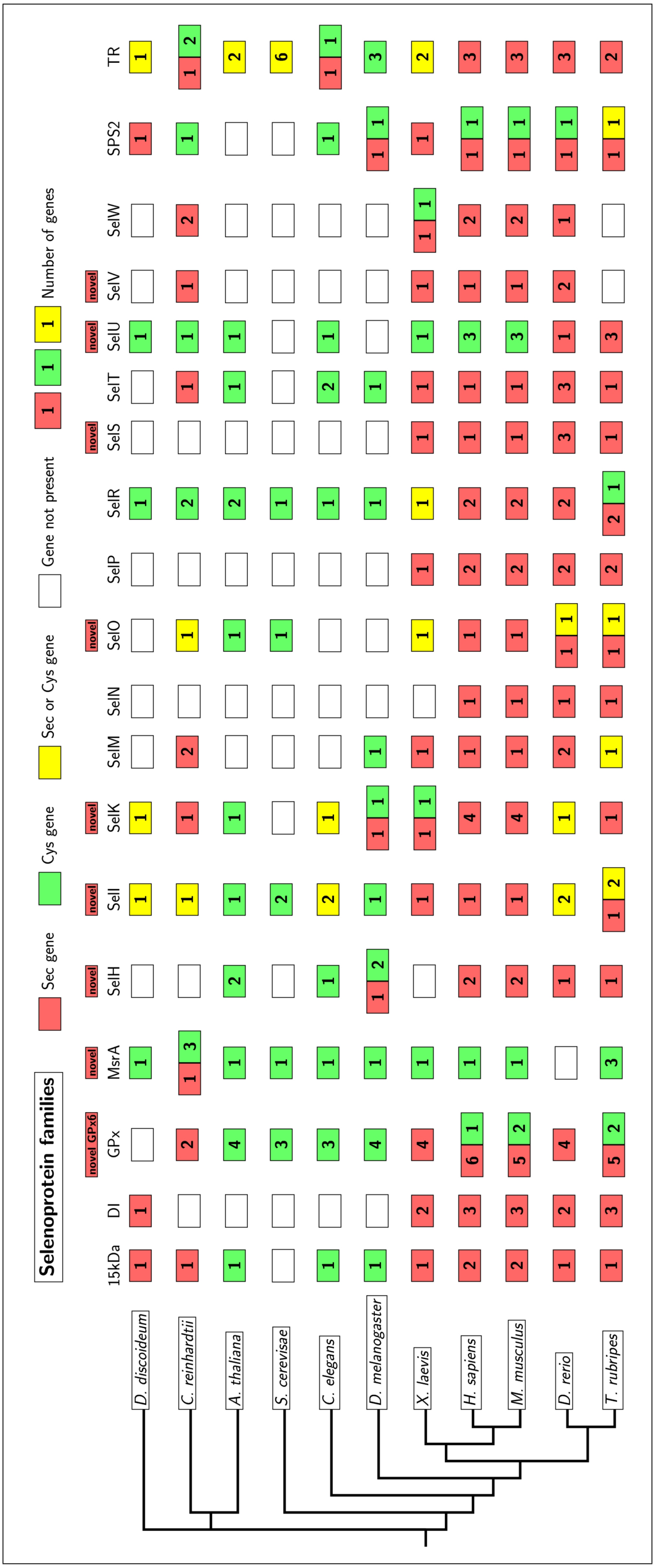
[1]Genoscope & CNRS UMR 8030, 2 rue Gaston Crémieux, 90057 Evry Cedex, France; [2]Whitehead Institute/MIT Center for Genome Research, 320 Charles Street, Cambridge, Massachusets 02141, USA; [3]Laboratoire de Biologie Moléculaire de la Cellule, CNRS UMR 5161, INRA UMR 1237, Ecole Normale Supérieure de Lyon, 46 allée d'Italie, 69364 Lyon Cedex 07, France; [4] Muséum National d'Histoire Naturelle, Département Systématique et Evolution, Service de Systématique Moléculaire, CNRS IFR 101, 43 rue Cuvier, 75231 Paris, France; [5]Défenses Antivirales et Antitumorales, CNRS UMR 5124, 1919 route de Mende, 34293 Montpellier Cedex 5, France; [6]Grup de Recerca en Informàtica Biomèdica, IMIM-UPF and Programa de Bioinformàtica i Genòmica (CRG), Barcelona, Catalonia, Spain; [7]CNRS UMR 5558 Biométrie et Biologie Evolutive, Université Lyon 1, 69622 Villeurbanne, France; [8]INRA-CNRS Laboratoire des Interactions Plantes Microorganismes, 31326 Castanet Tolosan Cedex, France; [9]Agencourt, Massachusetts, USA; [1]0 Biofuture Research ¡¡Evolutionary Fish Genomics¿¿, Physiologische Chemie I, Biozentrum, University of Wuerzburg, Am Hubland, D-97074 Wuerzburg, Germany.

***Tetraodon nigorviridis* is a freshwater pufferfish from South East Asia. It possesses the smallest known vertebrate genome, about 8 times smaller than the human genome. Here, we report a draft assembly and initial analysis of the sequecne of *Tetraodon*. The majority is assembled and mapped on 21 chromosomes, and the annotation identified about 28,000 genes and only 4,000 copies of transposable elements. The *Tetraodon* proteome and genome were compared to those of human, mouse and *Takifugu*, providing a global picture of genome organisation and divergence rates in DNA through the analysis fo two mammal and two fish genomes. The analysis of gene duplications within *Tetraodon* and with other fishes indicates that duplicate genes are created and deleted faster in fishes than in mammals. Type I cytokines and their receptors, selenoproteins and Hox genes were sistematically investigates and show interesting variations compared to other vertebrates. The sequence of *Tetraodon* thus reveals a compact and dynamic genome, and provides new resources for comparative genome anlysis in vertebrates.**

We examined in detail several gene families that represent important challenges for automatic annotation procedures or present a particular biological interest. The first is the family of selenoproteins which contain selenocysteine (Sec), the 21st amino acid, a rare cysteine analog where sulfur is replaced by selenium (Hatfield (ed.), 2001). Current automatic gene annotation methods are unable to identify selenoproteins because Sec is encoded by the UGA codon which usually represents a termination signal for translation and, therefore, confounds gene predictors. Selenoproteins also contain a specific RNA regulatory structure (SECIS) situated in their 3UTR which provides the required signal for the insertion of Sec at the TGA codon. Nineteen selenoprotein families have so far been identified in eukaryotic genomes, and most members also have Cys-containing homologs. Recent analyses indicate that the human selenoproteome has all but two known selenoprotein families (Kryukov et al., 2003), but that the Takifugu genome also possesses an additional SelU family (Castellano et al., in press), which exists in mammals in a Cys-containing form. To characterise the Tetraodon selenoproteome we first identified known selenoproteins and reannotated their partially predicted GAZE models by extending their coding gene structure. A modified version of geneid (Castellano et al., 2001) able to predict exons with an in-frame TGA in coordination with a downstream SECIS element was used and selenoprotein genes belonging to 18 distinct families, including SelU, were annotated (See Supplementary Information). We next searched the Tetraodon assembly and predicted genes for potential new selenoprotein families. One such candidate, currently under study, is widely distributed in, but restricted to, actinopterygians among vertebrates. Furthermore, it may be the first selenoprotein family without even a Cys-counterpart in mammals. The correct annotation of selenoproteins is important since they are thought to underlie the biological functions of selenium which is implicated in processes as diverse as male infertility, prevention of cancer and heart diseases, reduction of viral expression, ageing and immune function.

3.5 The eukaryotic selenoproteome

In the following figure, the distribution of eukaryotic selenoproteins and their Cys-homologs is depicted. Note that the species tree is only indicative of the relative phylogenetic position between organisms and no estimation should be made from branch lengths. On the other hand, selenoprotein families are ordered alphabetically and no inference is neither to be made from their arrangement or association.



Selenoprotein families

🟥 Sec gene 🟩 Cys gene 🟨 Sec or Cys gene □ Gene not present 🟥1 🟩1 🟨1 Number of genes

	15kDa	DI	novel GPx6 GPx	novel MsrA	novel SelH	novel SelI	novel SelK	SelM	SelN	novel SelO	SelP	SelR	novel SelS	SelT	novel SelU	novel SelV	SelW	SPS2	TR
D. discoideum	🟥1	🟥1	□	🟩1	□	🟨1	🟨1	□	□	□	□	🟩1	□	□	🟩1	□	□	🟥1	🟨1
C. reinhardtii	🟥1	□	🟥2	🟥1 🟩3	□	🟨1	🟥1	🟥2	□	🟨1	□	🟩2	□	🟥1	🟩1	🟥1	🟥2	🟩1	🟥1 🟩2
A. thaliana	🟩1	□	🟩4	🟩1	🟩2	🟩1	🟩1	□	□	🟩1	□	🟩2	□	🟩1	🟩1	□	□	□	🟨2
S. cerevisae	□	□	🟩3	🟩1	□	🟩2	□	□	□	🟩1	□	🟩1	□	□	□	□	□	□	🟨6
C. elegans	🟩1	□	🟩3	🟩1	🟩1	🟨2	🟨1	□	□	□	□	🟩1	□	🟩2	🟩1	□	□	🟩1	🟥1 🟩1
D. melanogaster	🟩1	□	🟩4	🟩1	🟥1 🟩2	🟩1	🟥1 🟩1	🟩1	□	□	□	🟩1	□	🟩1	□	□	□	🟥1 🟩1	🟩3
X. laevis	🟥1	🟥2	🟥4	🟩1	□	🟥1	🟥1 🟩1	🟥1	□	🟨1	🟥1	🟨1	🟥1	🟥1	🟩1	🟥1	🟥1 🟩1	🟥1	🟨2
H. sapiens	🟥2	🟥3	🟥6 🟩1	🟩1	🟥2	🟥1	🟥4	🟥1	🟥1	🟥1	🟥2	🟥2	🟥1	🟥1	🟩3	🟥1	🟥2	🟥1 🟩1	🟥3
M. musculus	🟥2	🟥3	🟥5 🟩2	🟩1	🟥2	🟥1	🟥4	🟥1	🟥1	🟥1	🟥2	🟥2	🟥1	🟥1	🟩3	🟥1	🟥2	🟥1 🟩1	🟥3
D. rerio	🟥1	🟥2	🟥4	□	🟥1	🟨2	🟨1	🟥2	🟥1	🟥1 🟨1	🟥2	🟥2	🟥3	🟥3	🟥1	🟥2	🟥1	🟥1 🟩1	🟥3
T. rubripes	🟥1	🟥3	🟥5 🟩2	🟩3	🟥1	🟥1 🟨2	🟥1	🟨1	🟥1	🟥1 🟨1	🟥2	🟥2 🟩1	🟥1	🟥1	🟥3	□	□	🟥1 🟨1	🟥2

Figure 3.1: The eukaryotic selenoproteome.

Methods

In this chapter, selected parts of the online Supplementary material of each article are given. The poor quality of the PDF documents offered by the journals has led to the extraction and recomposition of the text. Due to lenght constraints, figures have not been included but web links to the original supporting information are given instead (see below).

4.1 Castellano et al., EMBO reports, 2, 697-702 (2001)

Coding potential

Markov Models of order five are typically used to discriminate coding from non coding regions (Borodowsky and McInich, 1993; Guigó, 1999). In a Markov Model of order five of a nucleotide sequence, the probability of a given nucleotide at a particular position depends on the preceding five nucleotides (hence, the order five). Usually, the probabilities of the Markov Model are estimated separately from sets of coding and non coding sequences. Thus, for each hexamer h=s1s2s3s4s5s6, let E(h)=P(s6/s1s2s3s4s5) be the probability in coding sequences of nucleotide s6 given that s1s2s3s4s5 are the preceding nucleotides, and let I(h) be the same probability in non-coding sequences. Typically, the log-likelihood ratio L(h)=log(E(h)/I(h)) is computed. If P(s6/s1s2s3s4s5) is larger in coding sequences than in non-coding sequences, then L(h) is positive, otherwise it is negative. Then, given a nucleotide sequence S of length l, we compute the coding potential of S as:

Where si...j is the subsequence of S from positions i to j. L(S) is the logarithm of the ratio of the probability of S under the coding model (that is, assuming that S is a coding sequence), over the probability of S under the non coding model. L(S) tends to be positive in coding regions, and negative in non coding regions. Actually, L(S) is computed in somehow a more complicated way, since to compute the probability of the Markov chain S, the probability is required of the first five nucleotides in the sequence S (the so-called Initial probabilities I(S), versus the Transition Probabilities E(S)). Moreover, different Markov Models are usually computed for each different reading frame. See Borodovsky and McInich (1993) for details.

We have used Markov Models of order five to compute the coding potential L of the region comprised between the in frame TGA codon and the stop codon in selenoproteins, and of the region comprised between the stop codon TGA, and the next stop codon in frame in non selenoproteins. We hypothesized that the coding potential L will be in general much higher in selenoproteins than in no selenoproteins in this region, and therefore that its value can be used to distinguish between actual selenoproteins and false predictions in SECIS-positive nucleotide sequences. To test this hypothesis, three different sets of non-redundant human mRNAs from the 3' UTR database (Release 12.0, 09/1999, Pesole et al.) were extracted: 1) 3001 mRNAs with a UAG or UAA stop codon annotated; 2) 10 annotated selenoproteins; and 3) 1169 mRNAs with a UGA stop codon annotated. The following entries were discarded: partial, non standard, with alternative splicing, pseudogenes, predicted, artificial, viral, mitochondrial, histocompatibility related or with any problem in the CDS. The resulting sets of sequences can be obtained from http://genome.imim.es/datasets/spdroso2001

Prediction of Selenoproteins in nucleotide sequences

The method that we have developed relies in the correlated prediction of SECIS elements and of genes in which candidate exons with in frame TGA are allowed. SECIS elements are predicted first, and given as an input to the gene prediction program, which takes them into account to predict gene structures.

Given a query sequence (genomic or cDNA), first of all we predict SECIS elements using the pattern matching program PatScan with the SECIS pattern described below (`http://www-unix.mcs.anl.gov/compbio/PatScan/HTML/PatScan.html`). Only those predicted SECIS structures showing sufficient thermodynamic stability (as measured by the minimum free energy required to fold the structure) are further considered. Next, a modification (described below) of the program geneid (Guig et al., 1992; Parra et al., 2000) is used to predict genes along the query sequence. The program is able to predict genes in which exons may be interrupted by in frame TGA codons, as well as standard genes. Together with the query nucleotide sequence, the predicted SECIS elements positions are given as input to the geneid program, so that such genes are only predicted when a suitable SECIS appears upstream of the predicted gene at the appropriate distance. Thus, the coordinate prediction of SECIS and exons with in frame TGA codons decreases enormously the number of these elements that could actually occur in selenoproteins, leading to highly specific predictions of selenoprotein genes in nucleotide sequences (see Results). In what follows, we describe in some more detail the components of our method.

SECIS prediction

The program `patscan`, which searches protein or nucleotide sequences for instances of a pattern, is used to predict SECIS elements in nucleotide sequence. Eukaryotic SECIS structures fall into two slightly different classes, named forms 1 and 2. Form 2 elements present and additional helix due to base pairing in the apical loop. Thus, form 2 is an extension from form 1. When building a SECIS pattern is enough to model form 1 to be able to retrieve form 2 SECIS, though we may miss some form 2 specificity. In our case, because we favor sensitivity rather than specificity one model is sufficient. An input descriptor of the primary and secondary structure of the SECIS was built from 51 known SECIS (16 form 1 and 35 form 2) from 47 selenoproteins ranging from *Schistosoma mansoni* to human. Figure 5B shows the uniq resulting SECIS pattern. In order to assess the stability of the SECIS structure, the minimum free energy of all matching motifs was calculated with the `RNAfold` program (Viena RNA package) following the protocol of Kryukov et al. (1999). We estimate separately the free energies for Helix I plus internal loop and Helix II plus apical loop regions of the putative SECIS elements (see Figure 5A for SECIS structure). The cutoff parameters estimated by Kryukov et al. from 14 human SECIS were slightly changed from -7.4 to -7.5 kcal/mol for Helix I and internal loop, and from -11.0 to -10.0 kcal/mol for Helix II and apical loop. These values excluded the second SECIS of the four SelP proteins in our set of 47 selenoproteins.

Gene prediction

`geneid` is a program to predict protein coding genes in anonymous eukaryotic sequences designed with a hierarchical structure (see Parra et al., 2000, and the geneid documentation at `http://genome.imim.es/geneid` for details). Basically, it involves three steps:

1. prediction of sites. In the first step, start, stop codons, and splice sites are predicted and scored along the query sequence. Scores for potential sites are computed using a log-likelihood approach similar —albeit simpler— to the one described earlier. Essentially, given the sequence of a potential site, geneid computes the logarithm of the ratio of the likelihood of the sequence in a real site versus the likelihood of the sequence in a random site.

2. prediction of exons. In the second step, geneid builds all possible exons compatible with the predicted sites. Four types of exons are considered:

   (a) First, an ORF that begins with a start codon and ends with a donor site;

   (b) Internal, an ORF that begins with an acceptor site and ends with a donor site;

   (c) Terminal, an ORF that begins with an acceptor site and ends with a stop codon; and

(d) Single, an ORF that begins with a start codon and ends with a stop codon. It corresponds to intronless genes

ORFs are defined using the standard stop codons. Exons are scored as the sum of the log-likelihood scores of the exon defining sites, plus the log-likelihood ratio of a Markov model for coding DNA (exactly as the one described earlier) for the exon sequence. The resulting score can be assumed to be a log-likelihood ratio.

3. assembly of genes. From the set of predicted exons, geneid finally assembles the gene structure that maximizes the sum of the scores of the assembled exons. Note that the score of the resulting optimal structure can be assumed to be a log-likelihood ratio itself. geneid can predict multiple genes in both strands in the same query sequence.

   When assembling gene structures, geneid can take into account additional information about gene elements along the sequence. This information is provided externally, and may include previous knowledge about coding regions, or predictions obtained by other programs. Is in this way, that predicted SECIS elements can be introduced into the gene predictions.

   To be assembled into a gene structure, predicted exons and other genomic elements provided to geneid must conform to a number of user-defined biological constraints, such as frame compatibility, minimum and maximum distance between consecutive elements, and the order in which different genomic elements can be chained. All this rules are stated in the gene model, which is specified externally (see Parra et al., 2000, and the geneid documentation at `http://genome.imim.es/geneid` for details).

Prediction of selenoprotein genes

We have modified slightly `geneid` in order to include the possibility of predicting selenoproteins. Essentially, the codon TGA has been obviated as a stop codon when building First and Internal exons. When building Terminal exons, or intronless genes (Single exons), both the exons terminating at the codon TGA, and the exons extending beyond this codon to the next stop codon in frame have been considered.

The gene assembly algorithm has been modified to register during gene construction the incorporation of exons interrupted by codons TGA in frame, so that genes containing such exons are only predicted when an appropriate SECIS element is found at the right distance. In this way, the modified version of geneid is able to predict, at the same time, both standard genes and selenoprotein genes.

Prediction of protein secondary structure

The crystal structure of an eukaryotic selenocysteine, the bovine glutathione peroxidase, has been resolved at 0.2 nm resolution (Epp et al., 1983). The catalytic site of this enzyme is characterized by a beta-sheet—turn—alpha-helix structural motif, with the selenocysteine residue lying within the turn. Secondary structure predictions around the selenocysteine residue of most known selenoproteins, obtained using the program `predator` (Frishman and Argos, 1997), essentially conformed to this structure (data not shown). The same structure is predicted in dSps2, and dSelM, while dSelG lacked enough sequence context. However, a different structural motif (beta-sheet—turn—beta-sheet with the selenocysteine residue next to the first sheet) is predicted in the case of the fourth potential selenoprotein. We assumed thus, the fourth prediction to be a false positive.

Prediction of selenoproteins in the sequence of the *Drosophila* genome

`geneid` was used to scan the *Drosophila* genome for potential selenoproteins. The 19 large scaffolds of the genome summing up 115229998bp and with 13329 genes annotated were used (Adams et al., 2000). When using geneid, the restriction was enforced that selenoprotein genes could not be further than 500bp upstream from a predicted SECIS. geneid took about 45 minutes to scan this genome in a Pentium III processor running at 550 MHz. geneid predicted 12,194 genes; this number of genes comes from using the geneid parameters for sensitivity and specificity that achieved the highest accuracy when tested on the Adh region (Parra et al., 2000). Because we are interested in properly predicting a small number of genes, the non-standard selenoprotein genes, while keeping false positive to a minimum, accuracy must be the highest possible regardless of the total number of genes predicted. Predicting more genes, therefore, decreases the overall quality.

4.2 Kryukov et al., Science, 300, 1439-1443 (2003)

Databases

The 08/06/01 ”GoldenPath” draft assembly of the human genome that was masked for repetitive elements with RepeatMasker was used in the present study. Mouse and rat genome shotgun sequencing data, completely and incompletely sequenced genomes of eukaryotes, archaea and bacteria, and EST databases were obtained from NCBI, TIGR or sources indicated in the text.

SECIS-based identification of selenoprotein genes in eukaryotic genomes

SECISearch 2.0: genome-wide identification of SECIS elements

`SECISearch 2.0` can identify candidate SECIS elements in nucleotide sequence databases on the basis of their primary sequences, secondary structures and predicted free energy criteria. This program has major improvements over its initial version16 both at the level of individual modules and the overall composition of the program. An on-line version of `SECISearch` (Supplementary figure 13) is available at `http://genome.unl.edu/SECISearch.html` and allows a user to choose among three patterns of different stringency and manually adjust free energy parameters. Several fine structural filters are also optional.

Both `SECISearch 2.0` and its on-line version contain three modules. The first module is based on the PatScan program (`http://www-unix.mcs.anl.gov/compbio/PatScan/HTML/patscan.html`) and searches for RNA structures that match the SECIS element primary sequence and the secondary structure consensus. The second module, based on the RNAfold program from Vienna RNA package (`http://www.tbi.univie.ac.at/~ivo/RNA`) (S1), predicts secondary structure and calculates free energy for the entire SECIS element and separately for its core structure composed of Quartet, Helix II and Apical loop. The program imposes three constrains: 1) pairing of the Quartet nucleotides; 2) the presence of an unpaired nucleotide in the 5′ proximal position to the Quartet; and 3) the presence of two unpaired nucleotides that correspond to the AA motif in the Apical loop or bulge in the SECIS element consensus. Predicted RNA structures, whose calculated free energies are above thresholds determined from the analysis of known SECIS elements, are excluded from further analysis by the second module. The third module of SECISearch 2.0 imports SECIS candidates that are generated by the first two modules and filters out structures that possess features not found in any known eukaryotic SECIS elements. Specifically, these fine structural requirements remove SECIS candidates that 1) are Y-shaped, 2) contain ¿2 adjacent unpaired nucleotides among 7 nucleotides in Helix II that are proximal to the Quartet, 3) contain ¡8 base pairs in the SECIS segment composed of Helix II and Apical loop; and 4) contain ¿2 unpaired nucleotides on the 5′ side than on the 3′ side. For convenient visualization and examination of the data, we developed an RNAnice program, which can draw SECIS elements in proper orientation, with annotation and highlighted features.

Eukaryotic SECIS elements are usually classified as Type I and Type II structures (S2). Type I SECIS elements have fully unpaired Apical loop, whereas Type II SECIS elements possess an additional minihelix within the Apical loop. Both structures are interconvertible by mutations in the minihelix (S2) and do not differ in their predicted free energy values16. `SECISearch 2.0` is able to identify both SECIS types using the same set of parameters. `SECISearch 2.0` parameters were tuned using a set of 75 eukaryotic SECIS elements that were extracted from non-redundant and EST databases. This set included SECIS elements from all previously known human and mouse selenoprotein genes and also contained 37 SECIS elements from 11 other species.

SECISblastn: analysis of evolutionary conservation of predicted candidate SECIS elements

Since SECIS elements are essential for Sec insertion (and therefore for selenoprotein function), they are subject to natural selection pressure. We have found that not only secondary structures of SECIS elements are conserved, but that allknown human SECIS elements exhibit nucleotide sequence similarity to SECIS elements in orthologous mammalian selenoprotein genes. In contrast, non-orthologous SECIS elements have no detectable sequence homology. This finding allowed us to greatly reduce the number of false positives by requiring that each human candidate SECIS element should have a homologous SECIS element in rat, mouse or both rat and mouse genomes.

blast (S3) databases were generated from human sets of candidate SECIS elements generated by SECISearch and the mouse and rat sets of SECIS candidates were searched against these databases using SECISblastn. This BLASTN-based program has been optimized for comparison of short segments of 3'-UTR regions (cost to open a gap is 3, cost to extend a gap is 1, reward for nucleotide match is 2, and low complexity sequence filtering with dust is off). Mouse or rat candidate SECIS elements were discarded if no hits in the human database were found with an expectation value below 1e-10 (this threshold was determined from homology analyses of known SECIS elements in human and mouse orthologous selenoprotein genes). SECISblastn allowed more than 100-fold reduction in the number of false positives.

Shotgun redundancy filter

Intrinsic redundancy of mouse and rat shotgun genome sequence data resulted in redundancy of the set of identified putative SECIS elements and was removed by the redundancy filter that was developed using String::Approx Perl module for approximate string matching. All candidate SECIS elements in the mouse and rat sets with identity of ¿95% (measured as Levenshtein edit distance) to each other were replaced by first representative hits.

Human contamination filter

Our preliminary searches indicated that the current rat and mouse shotgun sequence data are contaminated with human sequence entries. To remove human sequences, we utilized a ”cleaning” procedure each rodent shotgun sequence entry that contained a putative SECIS element was compared with non-masked human genome using BLASTN program. Entries with ¿96% homology in regions longer than 500 nucleotides were removed from further analysis, and those that produced hits with a length l and identity level I were removed from further analysis if I exceeded l*(1.142-0.005769*l). The fact that no known selenoprotein genes were lost during this procedure suggested the legitimate choice of criteria. A set of human candidate SECIS elements that corresponded to the remaining mouse and rat hits was extracted for further analysis of upstream genome regions with the geneid program.

geneid: a gene structure prediction program

geneid is a program that predicts protein coding genes in anonymous eukaryotic sequences (S4; program documentation is available at http://genome.imim.es/geneid). We have modified geneid for predicting selenoprotein genes. The new version of the program recognizes TGA as both a stop codon and as a sense codon for selenocysteine. Thus, coding exons with in-frame TGA can be reliably predicted as long as they maintain high coding potential in sequences downstream of the TGA. In a single prediction on a given genome, the modified version of geneid is able to predict both standard genes and selenoprotein genes. For each candidate human SECIS element, flanking 1 Mb sequence regions on each side were extracted. Selenoprotein gene prediction was performed, admitting genes interrupted by in-frame TGA codons with an additional requirement that SECIS structures be located less than 6,000 nucleotides downstream of the predicted stop codons.

Mammalian Selenoprotein Gene Signature: analysis of evolutionary conservation of predicted selenoprotein genes

Mammalian Selenoprotein Gene Signature (MSGS) is a set of criteria that describe features common to mammalian (and possibly eukaryotic) selenoprotein genes20:

1. TGA-encoded Sec should be conserved and Sec-flanking protein sequences should be homologous for mammalian orthologous selenoprotein genes;

2. The SECIS element should be conserved and located in the 3-UTRs of mammalian orthologous selenoprotein genes; and
3. Distinct Cys- and/or Sec-containing homologs should exist, i.e., the occurrence of genes containing a Cys codon in place of TGA (or occurrence of distinct homologous genes that conserve TGA).

Predicted amino acid sequences of `geneid`-predicted selenoproteins were analyzed for the presence of paralogs in the human genome and homologs in other species in non-redundant and EST databases with `blast` programs. Six predicted selenoproteins that had both selenoprotein homologs (which contained SECIS elements) and cysteine-containing homologs (which had no SECIS elements) were considered to be true positives (GPx6, SelH, SelK, SelO, SelS and SelV). The remaining new selenoprotein, SelI, had no cysteine homologs, but its orthologs in frogs, fish and other mammals had SECIS elements (Supplementary figure 8). Thus, this protein was also classified as true positive.

Identification of human selenoprotein genes by searching for Sec/Sec and Sec/Cys pars in homologous sequences (SECIS-independent methods)

Comparative selenoprotein gene prediction

SECIS-independent selenoprotein gene searches were performed on the 08/06/01 ”GoldenPath” human genome assembly. The procedure employed was based on identification of in-frame TGA codons regardless of the presence of downstream SECIS elements, therefore addressing the issue of non-canonical SECIS elements in the human genome. This procedure also addressed the issue of potential occurrence of selenoproteins specific to the human genome. In the SECIS-independent searches for new selenoproteins, sensitivity was preferred to specificity, thus the chance of missing yet unknown selenoproteins was minimized. The ab initio gene prediction yielded 50,126 potential human genes, of which 27,605 had a TGA in-frame. This latter set included 21 out of 24 true selenoprotein genes that were identified by the SECISearch/SECISblastn/geneid/MSGS procedure. The set of 27,605 genes was further analyzed as follows:

1. The human 27,605 sequences were analyzed by BLASTP against a corresponding set of Takifugu rubripes proteins interrupted by TGA codons. The genome of this puffer fish (10/25/01 JGI draft assembly) encodes selenoprotein homologs of all 25 human selenoproteins, although the number of proteins in each selenoprotein family is different between human and puffer fish genomes. The ab initio geneid analysis of the puffer fish genome yielded 33,126 genes, of which 28,603 had a TGA in-frame, including 16 true selenoproteins corresponding to all but three human families. Human and fish proteins were then analyzed to identify potential human-fish selenoprotein orthologs containing in-frame TGA codons. This analysis identified 351 candidate orthologs;
2. The 27,605 human sequences were analyzed by BLASTP against a set of predicted Takifugu rubripes standard proteins. The ab initio geneid analysis of the puffer fish genome yielded 41,127 standard genes. Human and fish proteins were then analyzed to identify potential human-fish selenoprotein orthologs containing cysteine in fish. This analysis identified 296 candidate orthologs;
3. The sequences of these two sets of human candidate selenoproteins (351 + 296) were analyzed by BLASTP and TBLASTN against several completely sequenced eukaryotic genomes as well as against proteins predicted in these genomes (Drosophila melanogaster, Caenorhabditis elegans, Saccharomyces cerevisae and Arabidopsis thaliana). The incompletely sequenced genomes were also analyzed (Mus musculus, Xenopus laevis and Danio rerio) to identify sequences with homology in TGA-flanking regions, containing either TGA (Sec codon) or TGT or TGC (Cys codons) in place of TGA. This analysis resulted in 32 human selenoprotein candidates with selenoprotein counterparts in fish and 58 human selenoprotein candidates with cysteine counterparts in fish; and
4. After filtering proteins that had been previously characterized, the set contained only known selenoproteins and 12 other candidates. However, comparisons of these twelve sequences with corresponding EST sequences discarded potential in-frame TGAs due to either 1) predicted gene structure incompatible with the exonic structure of identical ESTs; or 2) TGA codon not supported

by ESTs sequences (therefore, these were probable sequencing errors which produced false TGA codons in place of correct cysteine codons). Thus, SECIS-independent searches did not add new human selenoproteins to the set of selenoprotein predicted by the SECIS-dependent prediction.

Selenoprotein homology search: cysteine homolog approach

80% (20 out of 25) human selenoproteins have known homologs that contain cysteine in place of selenocysteine. Therefore, cysteine-containing homologs of most mammalian selenoproteins are likely already annotated in public databases and canbe used to unveil their selenoprotein counterparts, providing a third independent approach to selenoprotein identification. 29,076 standard human genes (Ensembl protein annotation on the 12/22/01 "GoldenPath" draft assembly) were analyzed by TBLASTN against all human ESTs (EMBL, Rel. 69). This set contained seven cysteine paralogs of known selenoprotein families: GPx (ENSP00000229441, ENSP00000262661, ENSP00000296734, ENSP00000244392), SelR (ENSP00000286571, ENSP00000277598) and SelW (ENSP00000269578).

In order to pinpoint novel human selenoproteins the following procedure was carried out: 1) selection of Ensembl proteins with at least 5 human ESTs containing a TGA codon in place of a given cysteine position; and 2) selection of Ensembl proteins with an unknown or unclear function that might correspond to a selenoprotein. The final set contained only the seven paralogs of already known human selenoproteins.

A similar procedure was carried out for 4,380 potential novel human proteins obtained from sgp2 predictions (S6). sgp2 is a program to predict genes by comparing anonymous genomic sequences from two different species. It combines `tblastx` (WU-Blast), a sequence similarity search program, with geneid, an ab initio gene prediction program. In this way, 4,380 new human proteins with a reliable mouse ortholog were obtained. Because of the novelty of these sequences, not many ESTs may be available. For this reason, proteins with as less as 2 human ESTs containing a TGA codon in place of a given cysteine position were selected for analysis. Four human candidates were further studied, though given the high error rate in EST sequencing, these proteins had low supporting evidence. No other homology support was found in screened genomes, and these ESTs were considered to have sequencing errors. Therefore, no novel human selenoproteins were discovered by this approach.

The overall data from the independent approaches (SECIS prediction, in-frame TGA prediction and Sec/Cys homology approaches) argue that we have identified all or almost all selenoprotein genes in the human genome. Thus, the remaining in-frame TGA codons may be interpreted as terminator signals.

4.3 Castellano et al., EMBO reports, 5, 71-77 (2004)

Gene prediction

geneid is a program to predict protein coding genes in anonymous eukaryotic sequences designed with a hierarchical structure (see Parra et al. 2000, and the geneid documentation at http://genome.imim.es/geneid for details).

Basically, gene prediction involves three main steps:

1. prediction of sites. That is, start (ATG), stop (TAA,TAG and TGA) and splice signals (GT and AG) that define potential exon boundaries. When predicting selenoproteins the TGA site is allowed two contrasting meanings, stop and selenocysteine codon (Castellano et al.,2001). Position Specific Scoring Matrices are used to predict splice sites and start codons. Thus, predicted sites are scored as the log-likelihood ratio of the site sequence under the site model and under the random model.

2. prediction of coding exons. geneid builds all possible exons compatible with the predicted sites and scores them according to the scores of the exon defining sites and to a coding potential function. The coding function reflects the species-specific bias in the usage of codons in protein coding regions. In geneid, a Markov Model of order five trained in known species-specific coding exons is used. These models have been typically applied to discriminate coding from non coding regions (Borodovsky and McIninch, 1993; Guig, 1999).

   We had previously shown that the region comprised between the in-frame TGA codon and the stop codon in selenoproteins bears the codon bias characteristic of protein coding regions, whereas the region comprised between the stop codon TGA, and the next stop codon in-frame in non-selenoproteins do not castellano: 2001a, as otherwise expected. Therefore, coding potential is in general much higher in selenoproteins than in no selenoproteins in this region, and this value can be used to distinguish between actual selenoproteins and false positive predictions.

3. assembly of genes. From the set of predicted exons, geneid assembles the gene structure that maximizes the sum of the scores of the assembled exons. When assembling gene structures, geneid can take into account additional information about gene elements along the sequence. This information is provided externally, and may include previous knowledge about coding regions, or predictions obtained by other programs. It is in this way, that predicted SECIS elements can be introduced into gene predictions (Castellano et al., 2001)

On the other hand, to be assembled into a gene structure, predicted exons and other genomic elements provided to geneidmust conform to a number of user-defined biological constraints, such as frame compatibility, minimum and maximum distance between consecutive elements, and the order in which different genomic elements can be chained. All this rules are stated in the gene model, which is specified externally. When predicting selenoproteins the model may specify that predicted genes with TGA in-frame interrupted exons are only allowed when a suitable SECIS element has been predicted within a given range of nucleotides of the predicted gene stop codon (Castellano et al., 2001).

Prediction of standard genes in the human and fugu genomes

Gene structure prediction using geneid was done in the human and fugu genomes to predict standard genes.

Human genome

geneid was ran on the August 6, 2001 Golden Path assembly (release hg8) of the Homo sapiens genome (http://genome.cse.ucsc.edu/). 42357 genes were predicted.

Fugu genome

geneid was ran on the October 25, 2001 Joint Genome Institute (JGI, release 1.0) assembly of the Takifugu rubripes genome (http://www.jgi.doe.gov/). This initial assembly provides short contigs, but the gene compactness of the fugu genome makes gene prediction feasible. 41127 genes were predicted.

Prediction of selenoprotein genes in the human and fugu genomes

As indicated above, we have modified slightly geneid in order to include the possibility of predicting selenoproteins. Essentially, the codon TGA can be understood both as stop and selenocysteine codon when building exons. Therefore, geneid is able to predict, at the same time, both standard genes and selenoprotein genes.

In contrast to the method presented in (Castellano et al., 2001), where candidate selenoprotein genes were predicted only when a suitable SECIS prediction was present at the appropiate downstream distance, here we introduce a SECIS independent gene prediction approach. Potential selenoprotein gene candidates are predicted regardless of the presence of a downstream SECIS structure. Gene predictions interrupted by in-frame TGA codons, are likely to occur only when the strong coding bias characteristic of coding regions is present across the in-frame TGA codon. However, SECIS independent selenoprotein prediction results in an overwhelming number of selenoprotein candidates, due to the additional number of exons predicted (those that contain a TGA in-frame), which decrease accuracy of final gene structures. Consequently, in the approach presented here, a different biological contraint is used. A comparative protocol is followed, in such a way, that homology assessments at the protein level (see below) take place of SECIS restriction.

Known selenoproteins: human and fugu genomes

Known selenoprotein genes were mapped in both, human and fugu genomes through blat (http://genome.cse.ucsc.edu/) and BLAST (Altschul et al., 1997) searches.

23 known human selenoprotein genes belonging to 15 different families (known at that time) were mapped onto the human genome. The modified geneid version was used to predict them and sensitivity of the program was assessed. 20 out of 23 selenoprotein genes were properly predicted. Only SelK, SelT and SelS genes were not predicted as selenoproteins.

22 known fugu selenoprotein genes belonging to 14 different families were mapped onto the fugu genome (SelW gene was not found in this genome). The modified geneid version was used to predict them and sensitivity of the program was assessed. 18 out of 22 selenoprotein genes were properly predicted. Only SelK, SelH, SelS and SelM genes were not predicted as selenoproteins.

In conclusion, 1) both genomes, as shown by the mapping of all but one fugu selenoprotein gene, are complete enough to run a gene prediction program on them; and 2) the modified geneid program is able to predict most selenoprotein genes without the SECIS constraint. Sensitivity (that is, predicting only as non-selenoprotein genes non-selenoprotein genes. Sn ¿80% in both genomes) is sufficient to make reasonable the prediction of novel selenoprotein genes in the human and fugu genomes.

In addition, the same seventeen (out of 22 common selenoprotein genes mapped on both genomes. Sn ¿75%) are properly predicted in the two genomes. This fact, makes also reasonable the asumption of, by means of a comparative approach between genomes, true selenoprotein genes can be pinpoint from false positive predictions.

Potential selenoproteins: human genome

The modified version of geneid able to predict TGA in-frame genes was run on the August 25, 2001 Golden Path assembly of the H. sapiens genome. 27605 selenoprotein genes and 21603 standard genes were predicted. The modified version of geneid yields, in a single gene prediction, standard genes and potential TGA in-frame genes. This set of standard genes was discarded because gene structures are more reliably retrieved from standard geneid (see Prediction of standard genes in the human and fugu genomes) and selenoprotein gene prediction is intended only to provide genes bearing a TGA in-frame.

On the other hand, the set of potential selenoprotein genes is, in number, more than half of the total standard genes predicted by the standard geneid program. In other words, specificity (that is, predicting as selenoproteins only real selenoproteins) of the modified version of geneid able to predict TGA in-frame genes is extremely low at the level of sensitivity demanded (see above). Reasons for this are 1) coding potential, despite higher and positive in coding open reading frames (ORFs), can not discriminate as well when admitting a stop codon (TGA) in-frame. Many genes add short ORFs after a real stop codon (TGA), having that untranslated regio a low, but positive, coding potential; and 2) geneid parameters of the modified version, are slightly bias to include TGA in-frame exons. In this way, and because our aim is finding novel selenoprotein families, we minimize the chance of missing yet unknown selenoproteins by overpredicting them. False positive predictions are removed at later stages (see below).

Potential selenoproteins: fugu genome

A modified version of geneid able to predict TGA in-frame genes was run on the October 25, 2001 JGI assembly of the *Takifugu* genome (http://www.jgi.doe.gov/). 28603 selenoprotein genes and 4523 standard genes were predicted. Same considerations, as for gene prediction in the human genome, apply to gene prediction in the fugu genome (see above).

Comparison of human and fugu standard protein and selenoprotein sets

Selenoprotein families can have cysteine-homologs in the same or different genomes, but the Sec/Cys pattern for novel selenoproteins is unknown. Distribution of homologs can help to pinpoint selenoproteins and, in consequence, we introduced a protocol to predict and compare both types of genes.

Given human and fugu selenoprotein and standard gene complements we do the following set of intra and inter-genomic comparisons, at the protein level with blastp (query sequences were not filtered for low compositional complexity and a expectation value of 1e-10 was used. Stop codons in BLOSUM62 matrix were treated as cysteines), to reproduce possible Sec/Cys distribution patterns:

1. Inter-genomic comparisons

   (a) Predicted human selenoproteins against predicted fugu selenoproteins (Sec/Sec)

   (b) Predicted human selenoproteins against predicted fugu standard genes (Sec/Cys) Predicted fugu selenoproteins against predicted human standard genes (Sec/Cys)

2. Intra-genomic comparisons

   (a) Predicted human selenoproteins against predicted human selenoproteins (Sec/Sec)

   (b) Predicted human selenoproteins against predicted human standard genes (Sec/Cys)

   (c) Predicted fugu selenoproteins against predicted fugu selenoproteins (Sec/Sec) Predicted fugu selenoproteins against predicted fugu standard genes (Sec/Cys)

   However, these two types of comparisons (inter and intra-genomic), are not processed in the same way. First and separately for each predicted human and fugu selenoprotein (27605 human and 28603 fugu proteins), all possible inter-genomic comparisons are computed to define potential selenoprotein pairs having selenocysteine in either human, fugu or, alternatively, in both genomes. The result is a collection (subset of initial human and fugu predicted selenoproteins) of individual human and fugu potential selenoproteins with orthology support. Some cases having only Sec-Sec support, some others having only Sec-Cys and the rest both of them. Second, and once putative ortholog pairs have already been selected, paralogy data, if exist, is included for each of them (previously calculated from intra-genomic comparisons). In this way, and because paralogy is not as informative as orthology (see below), potential selenoprotein orthologs between human and fugu define pairs of putative selenoprotein families, and paralogs add additional support to them.

The rational behind this approach is that intra-genomic comparisons are false positive prone. Because of genome organization, where genes duplicate and may conserve sequence and gene structure, a false positive prediction in a genome (that is a gene with an incorrect TGA in-frame) may appear several times. Posterior comparisons would regard this gene as apotential selenoprotein family. However, this contingency is much more unlikely between genomes. The TGA (which is a false codon for Sec) may not be conserved and, at the same time, coding potential may be different (which can make that exon not to be included into predicted gene structure).

This procedure is consistent with the fact that human and fugu have all known selenoprotein families in Sec or Cys form. Therefore, we expect to predict a potential selenoprotein or cysteine homolog gene in both genomes and, at the same time, we use paralog information (too noisy by itself). Finally, human and fugu uniq selenoproteins, that have been treated independently up to now, are collapsed when define the same human-fugu or fugu-human pair (that is, query and subjectare the same but inverted).

Results were the following, 1) 368 human selenoprotein - fugu selenoprotein pairs (including 17 known human-fugu selenoprotein pairs); 2) 296 human selenoprotein - fugu cysteine homolog pairs; and 3) 216 fugu selenoprotein - human cysteinehomolog pairs. Note that Sec-Sec pairs may also have Sec-Cys homologs, though are included only in the Sec-Sec division.

3. Conservation around the selenocysteine amino acid

   Selected ortholog pairs were further analyzed to assess protein sequence conservation around the selenocysteine amino acid. A block of 20 amino acids (10 at each side of the Sec residue aligned to either Sec or Cys) was checked for havingat least 4 similar residues (according to BLOSUM62 matrix) on both parts. In order to gain sensitivity, when there were less than 10 residues on one, or both, sides the conservation assessment was skiped on that side(s). When applied, all known human and fugu selenoprotein pairs were recovered.

   The results of this filtering step were the following, 1) 49 human selenoprotein - fugu selenoprotein pairs (including 17 known human-fugu selenoprotein pairs); 2) 58 human selenoprotein - fugu cysteine homolog pairs; and 3) 26 fugu selenoprotein - human cysteine homolog pairs.

Search for homologs

In order to further validate the resulting human-fugu pairs, we undertook an exhaustive search against a number of databases of known coding sequences (proteins and ESTs) and several partial and full-length genomes. This approach should elicit real selenoprotein genes along with their Sec/Cys eukaryotic distribution. Each human and fugu selenoprotein member of potential pairs was studied.

International Protein Index

The International Protein Index (IPI, human version 2.0) (http://www.ebi.ac.uk/IPI/) is a protein database that provides a minimally redundant yet maximally complete set of human genes and proteins. IPI is assembled from human protein sequence information taken from the following 5 data sources: 1) SWISS-PROT; 2) TrEMBL; 3) Ensembl (http://www.ensembl.org); 4) RefSeq NPs; and 5) RefSeq XPs. This database was used to discard sequences highly similar to known proteins with functions apparently unrelated to those of selenoproteins.

In this way, blast searches against the IPI database narrowed the number of potential pairs, that is containing unknownproteins, to 1) 21 human selenoprotein - fugu selenoprotein pairs (including 17 known human-fugu selenoprotein pairs); 2) 9 human selenoprotein - fugu cysteine homolog pairs; and 3) 2 fugu selenoprotein - human cysteine homolog pairs.

Genomes

The following completely sequenced genomes from 1) Drosophila melanogaster; 2) Caenorhabditis elegans; 3) Saccharomicescerevisae; 4) Schizosaccharomyces pombe; 5) Plasmodium falciparum; and 6) Arabidopsis thaliana were queried by TBLASTN to identify sequences with homology in TGA-flanking region, containing either TGA (Sec codon) or TGT or TGC (Cys codons)in place of TGA. BLASTP searches against proteins annotated in these genomes were also carried out to identify cysteine-containing homologs. At the same time, partial sequenced genomes from 1) Mus musculus; 2) Xenopus laevis; 3) Danio rerio; 4) Dictyostelium discoideum; and 5) Chlamydomonas reinhardtii were also screened in the same way. These searches, allowed screening for new homolog sequences and reconstruction of Sec/Cys distribution across the eukaryotic lineage.

ESTs

NCBI EST database (dbEST, build of April 15, 2002) was queried to 1) check consistency of human and fugu genomic sequence at the Sec/Cys region; and 2) search for novel homologs for members of the 14 potential selenoprotein pairs and 3) define Sec/Cys distribution across the eukaryotic lineage.

Blast searches against dbEST discarded pairs with either 1) predicted gene structure incompatible with the exonic structure of identical EST sequences; or 2) TGA selenocysteine codon not supported by corresponding EST sequences, therefore,presumedly a genomic sequence error. This filtering step, apart from known human and fugu selenoproteins, resulted in two pairs containing both fugu selenoproteins and human cysteine homologs.

On the other hand, several Sec and Cys-containing SelU homologs were found (see below).

cDNAs

The TIGR collection of transcripts (cDNAs and ESTs, `http://www.tigr.org`) was screened to search for SelU orthologs. In this way, a cysteine-containing homolog was found for zebrafish (*Danio rerio*, TC173888) and japanese medaka (*Oryzias latipes*, TC21944).

Paralogs

The four sequences of the predicted two pairs, accounting for two fugu selenoproteins and two human cysteine homologs, were globally aligned with clustalw (Thompson et al., 1994). Their alignment clearly showed that, on basis of sequence similarity, they belong to the same protein family. This fact reinforced the likelihood of them belonging to a real selenoprotein family.

On the other hand, further TBLASTN searches were done against the human and fugu genomes to unveil unpredicted paralogous sequences. BLASTP searches against annotated proteins in these genomes were also accomplished. An additional fugu selenoprotein member of the SelU family and a human cysteine-homolog belonging also to this familiy were found.

Search for prokaryotic homologs

Fugu SelUa and human ENSG00000122378 proteins were blasted against 246 bacterial and 18 archaeal genomes available at NCBI (`http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi`). TBLASTN and BLASTP programs, against proteins from 177 annotated genomes, were used. No significant hits were found.

SelU distribution across the eukaryotic lineage

Searches above yielded SelU homologs all across the eukaryotic lineage. They can be divided into (common name given when known):

Sec-containing homologs were found in:

Fish: fugu (*Takifugu rubripes*), zebrafish (*Danio rerio*), japanese medaka (*Oryzias latipes*), catfish (*I. punctatus*), rainbow trout (*Oncorhynchus mykiss*), carp (*Cyprinus carpio*), three spined stickleback (*Gasterosteus aculeatus*)

Birds: chicken (*Gallus gallus*)

Echinoderms: sea urchin (*Strongylocentrotus purpuratus*)

Green algae: *Chlamydomonas reinhardtii*

Diatoms: *Thalasiosira pseudonana*

Cys-containing homologs were found in:

Mammals: human (*Homo sapiens*), mouse (*Mus musculus*), rat (*Rattus norvegicus*), pig (*Sus scrofa*), cow (*Bos taurus*), dog (*Canis canis*), rabbit (*Oryctolagus cuniculus*).

Fish: fugu (*Takifugu rubripes*), zebrafish (*Danio rerio*), japanese medaka (*Oryzias latipes*)

Amphibians: frog (*Xenopus laevis*), frog (*Silurana tropicalis*)

Tunicates: *Ciona intestinalis*

Arthropods: (insects): silkworm (*Bombix mori*)

Nematodes: *Caenorhabditis elegans*, *Caenorhabditis briggsae*, *Ancylostoma ceylanicum*, *Parastrongyloides trichosuri*, *Strongyloides stercoralis*, *Pristionchus pacificus*, *Toxocara canis*

Land plants: sweet orange (*Citrus sinensis*), barrel medic (*Medicago truncatula*), cabernet sauvignon (*Vitis vinifera*), sunflower (*Helianthus annuus*), barley (*Hordeum vulgare*), onion (*Allium cepa*), rape (*Brassica napus*), european aspen (*Populus tremula*), pepper (*Capsicum annuum*), sorghum (*Sorghum bicolor*)

Green algae: *Chlamydomonas reinhardtii*

Slime molds: *Dictyostelium discoideum*

Arg-containing homologs were found in:

Nematodes: *Strongyloides ratti*

No homologs were found in (complete genome sequence):

Arthropods (insects): fly (*Drosophila melanogaster*), mosquito (*Anopheles gambiae*)

Yeast: bakers yeast (*Saccharomices cerevisae*), fissions yeast (*Schizosaccharomyces pombe*)

Apicomplexa: malaria parasite (*Plasmodium falciparum*)

Prediction of protein secondary structure

The crystal structure of an eukaryotic selenocysteine, the bovine glutathione peroxidase, has been resolved at 0.2 nm resolution (Epp et al., 1983). The catalytic site of this enzyme is characterized by a beta-sheetturnalpha-helix structural motif, with the selenocysteine residue lying within the turn. Secondary structure predictions around the selenocysteine residue of most known selenoproteins, obtained using the program Predator (Frishman and Argos, 1997; Castellano et al., 2001), essentially conformed to this structure (data not shown). Fugu SelU selenoproteins also stick to this pattern when predicted with the `predator` program.

Prediction of SECIS elements

SECIS elements were predicted in selected selenoprotein genes with the SECISearch program (Kryukov et al., 2003). This program is available as a web server resource at http://genome.unl.edu/SECISearch.html. Given that predictions are only done in short genomic regions, false positive are not a concern, therefore a loose SECIS pattern can be used to permit identification of SECIS variants. The whole range of SECIS patterns provided by SECISearch were used. However, only canonical SECIS were found in T.rubripes (fugu, puffer fish), D. rerio (zebrafish) and G.gallus (chicken).

Search for fossil SECIS

Annotated UTR regions were extracted from Ensembl (www.ensembl.org) for human, mouse and rat SelU homologs. The IDs forthe three sets of SelU orthologous genes are: 1) ENSG00000122378, ENSMUSG00000021792, ENSRNOG00000011140; 2) ENSG00000157870, ENSMUSG00000029059, ENSRNOG00000013468; 3) ENSG00000158122, ENSMUSG00000021482, ENSRNOG00000018886. However, mostof these annotated UTRs were uncomplete. Possibly, because of the lack of EST sequences. In addition, UTR regions for SelU Cys-homologs from Takifugu rubripes, Danio rerio, Oryzias latipes, Xenopus laevis, Ciona intestinalis, Caenorhabditis elegans, Caenorhabditis briggsae and Dictyostelium discoideum were extracted from the TIGR collection of transcripts (cDNAs and ESTs, http://www.tigr.org) and, if needed, from the original genomic sequence.

In these UTR regions two analysis were performed:

Fish and chicken SECIS sequences were blasted against these UTRs in the search for similarity. No significant hits were found. However, while SECIS elements share a high degree of sequence identity among mammals (Kryukov et al., 2003), this is not necessarily the case for functional and vestigial SECIS between, for example, fish, chicken and mammalian SECIS. SECISearch was run on these UTRs with canonical and non-canonical patterns. No hits were found. Furthermore, the program PatScan (Dsouza et al., 1997) was used to run even more degenerated patterns. However, matches were unclear. Specially, because no similar hits were found between human and rodent UTRs.

In any case, the lack of a potential fossil SECIS does not yet discard the hypothesis of a Sec to Cys mutation, becausethe UTRs under study could have accumulated enough mutations to fade the SECIS phylogenetic signal.

In addition, SECIS similarity searches were run on the whole TIGR collection of transcripts (cDNAs and ESTs, http://www.tigr.org). The rational behind this was, again, to find vestigial SECIS elements through sequence similarity. In the hope that they are still recognizable, that is, change from Sec to Cys is either quite recent or the mutation rate is low enough, we could expect still some phylogenetic footprint. However, because even functional SECIS diverge, a negative result is likely and, at the same time, inconclusive respect to clarify evolutionary events. Searches were done on the Eukaryotic Gene Ortholog (EGO) database at TIGR. It is a collection of partial and full length cDNAs from 61 different eukaryotic organisms. Again, results were not convincing.

4.4 Jaillon et al., Submitted (2004)

Selenoproteins

Selenoproteins are proteins that incorporate selenium in the form of selenocysteine, the 21st amino acid (Hatfield (ed.), 2001), and they are widely distributed across the eukaryotic lineage in a family and taxa-specific fashion. In order to describe the Tetraodon selenoproteome (set of selenoproteins), we devised a two step protocol. First, a parallel annotation pipeline to reannotate known selenoprotein genes and second, a search for novel selenoprotein families in the Tetraodon sequence data. Finally, a variety of independent bioinformatics methods based on gene and SECIS prediction, together with comparative genomics approaches, were applied (Kryukov et al., 2003).

All these complementary approaches recognized the majority of known selenoprotein genes and have identified at least one promising novel selenoprotein candidate in the Tetraodon genome. Further computational and experimental analyses are pending. The Tetraodon selenoproteome consists of 18-19 distinct selenoprotein families. One of them, SelU, has Sec in fishes but Cys in mammals (Castellano et al., in press) and the putative novel one, also has Sec in fishes but no gene counterpart in other vertebrates. In conclusion, the Tetraodon and Takifugu genomes recapitulate all (15kDa, DI, GPx, SelH, SelI, SelK, SelM, SelN, SelO, SelP, SelR, SelS, SelT, SelU, SelV, SelW, SPS2, TR) but one (MsrA) of the 19-20 eukaryotic selenoprotein families and we hold the remaining gene models to be free of a recoded TGA codon.

Predicted GAZE gene models with high homology to human selenoproteins were reannotated to include the selenocysteine TGA codon in the ORF. A variety of methodologies were used to build the best possible gene models. First, when a full-length Tetraodon cDNA was available, the TGA-containing ORF was mapped into the genome to define the gene exon-intron structure with the spidey program (Wheelan et al., 2001). Second, if a cDNA sequence were not available (or not complete), theability of geneid to predict genes having a TGA in-frame and its capacity of handling partial homology data was used. Finally, when needed, the program genewise was also used to align the set of human selenoproteins against the Tetraodon genome.

We started the search of novel selenoprotein genes by running a coordinated prediction of SECIS elements and genes interrupted by in-frame TGA codons on both the genome and the cDNA sequence data. Such genes, however, can be predicted onlywhen a putative SECIS, whose position along the genome is input (GFF file) into geneid during gene prediction, exists at the right distance (no more than 1000 nt downstream). On the genome, the SECISearch program (Kryukov et al., 2003) predicted 2138 SECIS elements that resembled the standard SECIS secondary structure basepairing (Dsouza et al., 1997), were thermodynamically stable (Zuker and Stiegler, 1981) and had homology to the Takifugu genome (over 75% identity), of which geneid (Castellano et al., 2001) only paired 138 with a gene having a TGA in-frame.

These proteins were further analyzed by comparative genomics. In short, we searched for protein sequence alignments with conservation around Sec-Sec or Sec-Cys pairs, as suggestive of selenoprotein function. The underlying assumption is that sequence conservation in regions flanking a UGA codon strongly argues for protein coding function across the codon. Predicted protein sequences were therefore blasted against a variety of genomic and transcript sequences from a wide range of eukaryotic organisms. In addition, a block of 20 amino acids (10 on each side of the Sec residue aligned to either Sec or Cys) was checked for having at least 5 similar residues in both regions and proteins with high homology to well characterized proteins (human IPI, version 2.24) functionally unrelated to selenoproteins were discarded. No new selenoprotein families were unveiled.

A similar prediction was carried out on the cDNA sequences. SECIS and gene prediction are more accurate on transcript sequences because real SECIS elements only exist in UTRs and the lack of introns facilitates the finding of the right ORF. 245 potential SECIS were predicted, and gene prediction yielded 25 genes with a TGA in-frame. Comparative analyses were ran on this set and only one protein has turned out to be a putative novel selenoprotein family, and is now under further investigation. Interestingly, within vetebrates it is widely distributed in, but resticted to, actinopterygians in either Sec or Cys form. If it is indeed a bona fide selenoprotein, this family demonstrates the discrete and taxa-specific distribution of selenoprotein families in eukaryotes and, remarkably, could be the first selenoprotein without a Cys-containing gene ortholog in mammals and other vertebrates.

In addition, we also predicted genes interrupted by an in-frame TGA codon irrespective of the presence of SECIS structures on the genomic and cDNA data. In this way, selenoprotein genes with altered SECIS structures may be found. However, no novel selenoprotein candidates arised from this approach.

Finally, we made use of the possibility of having Cys-containing selenoprotein homologs among the predicted GAZE gene models. We tried to identify paralogs and orthologs in fishes that contain TGA in place of a Cys codon. No uncharacterized selenoproteins were found.

Annotation of genes

The annotation process refers to the description and location of genes and other biologically relevant features on a genomic sequence. That is, to define the particular genomic coordinates (nucleotide position along a DNA sequence) of the biological element of interest e.g. a gene or a promoter element. In the *Tetraodon* genome project, genes are annotated in the following way:

1. Three vertebrate genomes (human, mouse and *Takifugu*) had been totally or partially sequenced prior to *Tetraodon*, providing a catalogue of vertebrate predicted protein coding genes to guide annotation of those present in this pufferfish genome. This data was exploited using Exofish, a tool that identifies evolutionary conserved regions (ecores) with a high specificity. In particular, Human and mouse IPI proteins were compared to the Tetraodon assembly using Exofish.

2. Human and mouse IPI proteins that matched with Exofish were also aligned on *Tetraodon* using genewise.

3. The Exofish tool was then used to compare the entire human and mouse genomic sequences to the Tetraodon genome. Additional ecores, not identified through proteome comparisons, were found.

4. The Exofish tool was also used to compare the *Takifugu* and *Tetraodon* genome assemblies. This step had a higher sensitivity to detect genes than its equivalent with mammalian sequences.

5. To further increase the the possibility of identifying *Tetraodon* genes that are not conserved in any of the aforementioned genomes, and to refine annotation of predicted homologs the ends of about 155,000 *Tetraodon* cDNA clones from 7 different tissues were sequenced.

6. In addition, two ab initio gene prediction programs were used. genscan and geneid were trained on manually annotated *Tetraodon* genes and provided additional and/or complementary gene predictions.

7. Finally, all these predicted sequence segments were combined with the GAZE algorithm to provide the final annotation of predicted gene models on the *Tetraodon* genome sequence.

However, this genome annotation pipeline did not annotate correctly selenoprotein genes.

Reannotation of selenoproteins

The selenoprotein reannotation protocol is the following:

1. Map all known human selenoprotein families onto the predicted gaze genes (tblastn against virtual cDNAs) to find selenoprotein gene models that need reannotation.

2. Map all known human selenoprotein families onto the genomic sequence (tblastn against scaffolds) to search for additional selenoprotein homologs not predicted.

3. Collect the gff annotation for each gaze gene model likely to be a selenoprotein.

4. Map collected gaze genes corresponding to selenoproteins to experimental transcript sequences (cDNA or ESTs).

5. Predict selenoprotein genes on cDNA data with geneid and genewise (human selenoproteins are used as homologs).

6. Predict selenoprotein genes on genomic data with spidey (using Tetraodon transcripts. UTRs can be defined in this way) and genewise (predicted Tetraodon selenoproteins in cDNAs and human ones are used as homologs)

7. Predict SECIS elements on cDNA and ESTs data with SECISearch 2.0 and then map them back to the genome with blastn.

8. Modify gff annotation for each gaze gene model based on new gene structures predicted and include the SECIS information.

4.5 The eukaryotic selenoproteome

Human selenoproteins were mapped on several organisms with `blast`. The protocol for each genome was the following:

1. Collect all genomic an transcript sequences; and
2. `tblastn` all human selenoproteins against these data; and
3. Establish for each independent hit its Sec, Cys or unknown nature;
4. Add species to the selenoprotein distribution table at the correct phylogenetic position.

Discussion

This chapter owes its overall structure to the recent review of our work and others by Driscoll and Chavatte (2004). Although unpublished data is also discussed, as indicated appropriately. First, a convenient discussion of the previous knowledge on selenoproteins and on computational techniques of relevance to this work is presented. This includes the analysis of genes, proteins and SECIS sequences we performed before carrying out any genome-wide search for selenoproteins. Second, the novel computational methods, developed by us and others, for the finding of selenoprotein are outlined and compared, their strengths and weaknesses discussed and the results from their application addressed. Third, the most up-to-date distribution of selenoprotein genes and their Cys-containing homologues across the eukaryotic lineage is discussed. Finally, descriptive, but consequential, insights into the biology and evolution of selenoproteins are debated. A few speculations on the implication of selenoproteins and the UGA codon in the evolution of the genetic code are lightly touched in here.

Historically, selenoproteins have been identified by purifying a protein and cloning its cognate cDNA and, while effective, this approach is time-consuming. More recent *in silico* analysis have shown a greater potential to exhaustively identify selenoprotein genes and, notably, to keep up with the growing pace of new genome releases. [1] This shift towards the computational identification of selenoprotein genes, has resulted in the rapid accumulation of functionally uncharacterized selenoproteins.[2] Moreover, the role of the majority of previously recognized selenoproteins is yet unclear or poorly known and, therefore, the field faces the challenging task of determining the biological functions of this set of aged and newfound selenoproteins.

5.1 Eukaryotic selenoproteins

selenoprotein genes have several characteristics that can only be considered eccentric for the mainstream bioinformatician. Therefore, computational approaches have had to deal with selenoprotein peculiarities but, at the same time, have made use of the fact that these genes are, otherwise, archetypal in most of their features. This claim arises from the comprehensive study of the identified selenoproteins at the gene, protein and SECIS level. On one hand, the main nonstandard characteristics of selenoproteins are the existence of a UGA codon (a stop signal in the canonical genetic code) in-frame of the coding sequence and a folded RNA secondary structure in its 3'UTR of regulatory function.[3] In consequence, *ad hoc* bioinformatics approaches had to be developed. Needless to say at this stage that the nonstandard use of gene defining signals compounds selenoprotein identification.

As a matter of fact, and in relation to the atypical use of the UGA codon, current computational gene prediction programs at the base of genome annotation projects, invariably rely on the standard stop codons TAA, TAG and TGA to identify open reading frames (ORFs). More specifically, to define single (intronless) or terminal coding exons. Under such an assumption, selenoprotein genes, in which TGA does not necessarily imply termination of translation will be incorrectly predicted by these programs. While no exhaustive list is intended, well-known examples of eukaryotic *ab initio* gene prediction are the HMM-based `genescan` (Burge and Karlin, 1997), `fgenesh` (Salamov and Solovyev, 2000), `genie`

[1] In the time of this work, the fly, mosquito, human, chimpanzee, mouse, rat, *Arabidopsis*, *Takifugu*, *Tetraodon*, chicken, worm, slime mold, yeast and other eukaryotic genomes have been released. In addition, transcript data exist for a myriad of species.

[2] 11 novel selenoprotein families (half of the known families) have been identified by computational means in the last 5 years.

[3] Although, it is now well documented that the presence of primary and secondary structure signals with a control function is shared by a growing number of genes. Specially, in the case of genes subjected to recoding.

(Kulp et al., 1996) and the hierarchically-designed geneid (Parra et al., 2000), the gene prediction tool elaborated in our group. Without exception, all these programs would need recoding to admit a UGA codon within a predicted ORF and, while this technical modification may be feasible and even uncomplicated for some of the algorithms, the handling of, first, the drop in the regular accuracy and, second, the overwhelming number of predicted exons with a UGA codon in-frame, may result laborious and unsuccessful. Other more advanced software which, besides of gene defining signals, takes into account similarity with either DNA or protein sequences as slam (Alexandersson et al., 2003; Cawley et al., 2003), twinscan (Korf et al., 2001; Wu et al., 2004), genomescan (Yeh et al., 2001), sgp1 (Wiehe et al., 2001) and sgp2 (Parra et al., 2003; Guigó et al., 2003) suffer of the same frustrating problem: a UGA codon signals only protein termination no matter sequence conservation extends across it. In consequence, none of the aforementioned programs were appropriate for the goal we had in mind, that is to say, the prediction of genes with a UGA codon in-frame by bioinformatics means.

A different approach for gene prediction is the one based purely on sequence similarity.[4] Thus, one could ask about the suitability of this method to our problem. Particularly, because in relation to the comparative gene prediction methods mentioned above, this approach works conversely in that exon-defining signals and ORFs are used in the final steps of the algorithm. First, sequences are aligned (either protein/DNA or transcript/DNA) and, subsequently, splice sites are checked to adjust exon boundaries[5]. These procedure may make the existence of an underlying in-frame TGA codon irrelevant and, therefore, be an approach suitable for the prediction of selenoprotein genes. This ORF-independence should be obvious for the transcript/genomic alignments because UTRs do not necessarily keep an open reading frame[6] and, thus, could help to identify exons, coding and non-coding, having stop codons in-frame. Differentially, for protein/genomic alignments, the inclusion of termination codons within a coding exon may be more difficult. However, this is easier than expected because the UGA codon is not use as a gene defining-signal but, more or less, as any other codon and can be considered as just a genomic sequencing error (which may let the UGA to be aligned). In any case, the assumption behind these homology-based methods is that sequence similarity can delineate gene structures (coding and noncoding regions) regardless of biologically meaningful signals. That is, without the understanding of the meaning of the sequence. Though, this may lead to rough gene models in need of subsequent refinement.

In this respect, there are several well-known programs to either align transcript to protein data to genomic sequences. A first and basic tool is blast (Altschul et al., 1997). This program, unaware of exons, will provide a initial description of the gene structure, but with blurred exon edges. However, the alignment of stop codons[7] is possible and valuable. Another of such algorithms is the blat program (Kent, 2002), a tool which performs rapid mRNA/DNA and cross-species protein alignments, and outperforms blast in speed an accuracy. blat could be considered a sort of strict blast[8] with splice site consensus matching (GT/AG). This program, while perfect for RNA/DNA alignment, is troublesome when it comes proteins with stop codons in-frame. In this respect, other well-known program to align protein data to genomic sequences is genewise (Birney and Durbin, 2000). This tool, is able to admit a UGA codon in-frame of an alignment[9], as long as it is well centered between a stretch of highly conserved sequence. In other words, it assumes that, to obtain the best and longest possible alignment, a stop codon in-frame needs to be included and read-through. This makes the algorithm robust to small sequencing errors and, indirectly, to selenoprotein gene structure prediction. Thus, in principal, given a selenoprotein sequence we can build its coding genomic structure and, maybe, search for additional genes.

On the other hand, the alignment of transcript sequences to their original genomic context can be achieved, besides the blast and blat programs, by means of the est2genome (Mott, 1997) and spidey (Wheelan et al., 2001) programs, among others. In this case, a DNA-DNA alignment that

[4] Although, models of the exon-defining signals can be also used here to improve the finding of the correct exon-intron boundaries.

[5] This is an oversimplification of the algorithms behind.

[6] A concept in fact functionally meaningless in this region, besides for the so-called upstream ORF (uORF) which control the rate and pace in which the ribosome translates the main ORF in some mRNAs.

[7] Usually translated as stars (*) or X if masked.

[8] Although, blast builds an index of the query sequence and then scans linearly through the database and blat, reversely, builds an index of the database and then scans linearly through the query sequence.

[9] Though, it outputs a clear but sardonic message in your terminal with no awareness of selenoprotein genes:"Got a stop codon in the middle of a translation. Yuk¡‘

includes the recoded UGA codon will be easily obtained and the output exons will be ready for further inspection, that is, to determine them as coding or non-coding (UTR). Again, it is in the step of finding the ORF where is likely that the UGA codon confound us, but this time we will know with higher precision the exonic structure. Hence, intronic and intergenic UGAs will no longer be a problem as in *ab initio* gene prediction. In addition to the straightforward alignment of a transcript or protein sequence with the genomic locus that originated it, these programs have the capacity to align homologous sequences (proteins or transcripts) with other genomic regions. This is of interest because permits, making use of sequence conservation at the DNA or protein level, to find related genes in the same or different genomes. Back to the original question raised here, is time to know about the suitability of these programs for finding selenoprotein genes. Apparently, they are better tailored than the *ab initio* and comparative gene prediction programs for this task. However, different selenoprotein families do not share sequence similarity, ruling out the identification of novel selenoproteins families through homology searches. Such homology, if detectable, is distant and will be a matter of discussion below as it poses interesting evolutionary issues on the raising of Sec-containing proteins. In fact, and besides having in common the use of the selenocysteine residue, selenoproteins are unrelated in the sense that they do not form an homogenous or close monophyletic family. Nevertheless, these tools can be very useful to build the gene structure of known selenoprotein genes or as an alternative way to post-process such data (finding additional homologues, checking for supporting evidence for a TGA in-frame and others). A methodological conclusion, so far, is that available tools were of limited use for the prediction of novel selenoprotein genes and, in consequence, alternative algorithms that incorporate selenoproteins characteristics were in need. The identification of new selenoprotein families will retract us to the classical but modified *ab initio* gene finding.

As for the second selenoprotein peculiarity, the SECIS recoding signal, a smart and successful work had already been done towards its molecular description. In a series of “back to back” papers in the late nineties, in Berry’s and Krol’s lab the SECIS RNA secondary structure was figured out. Its consensus structure is composed of an initial helix and internal loop, followed by a second helix containing non-Watson-Crick base pairs UGAN....NGAN (the SECIS core or quartet), an unpaired A preceding the quartet, and an unpaired AA motif in the apical loop that ranges from the core 11 to 12 nucleotides (Walczak et al., 1996; Berry et al., 1997; Walczak et al., 1998). In addition, SECIS elements are divided into two classes, named form 1 and form 2, the latter having an additional small stem-loop at the end of the apical loop (Grundner-Culemann et al., 1999). However, no functional differences have been observed between the two classes. This initial consensus definition of the SECIS stem-loop, though correct for most selenoprotein genes, presents some deviations in later discovered SECIS structures. These are 1)the unpaired adenine is replaced by guanine in four SECIS elements (Buettner et al., 1999; **?**); and 2) conserved adenosines in the apical loop are replaced by cytidines in the human SelM SECIS structure (Korotkov et al., 2002). Thus, SECIS diversity is still under research.

The existence of such a secondary structure can be used as a signature for eukaryotic selenoprotein identification. This reasonable hypothesis, however, relies on the accuracy of the computational methods capable of dealing with bi-dimensional biological patterns. In an enlightening article, Dandekar and Hentze (1995), described the main computational steps towards the identification functional secondary structures on transcript sequences. In there, the problem is stated into three steps: 1) the coding of the known secondary structure in a pattern-oriented language parseable by computers; 2) the screening of DNA/RNA sequence data for instances of such a pattern. At this level, it means just to find stretches of sequence able to basepair in the pattern-defined way and, depending on the primary sequence conservation and the complexity of the structure, the number of false positive matches is overwhelming and meaningless; and 3) the assessment of such potential secondary structures from a more biological standpoint. Thermodynamic measures to compute the stability of the structure and comparative approaches to evaluate the sequence conservation of the signal across homologues are valuable possibilities.

The first and second step can be treated together. Programs exist that read their own pattern-defining language as input to then scan for matching structures on a sequence database. At this point, however, no biological conclusions can be inferred because the majority of predicted secondary structures are not functional. In other words, they just happen to be short stretches of sequence able to basepair in the specified way. This implies the need of an additional method to discriminate potentially functional structures from fake ones. As mentioned above, one such method is to evaluation of the thermodynamic stability of the predicted structures. Given that the majority SECIS structures fold with a significant decrease of the Gibbs free energy, potential SECIS can be reduced by several fold by this procedure. In addition, homology searches can play an outstanding role here. As discussed below, SECIS elements are highly conserved in sequence between close species (eg. human-mouse, *Takifugu-Tetraodon*), thus, the pinpointing of those predicted and conserved through species also decreases greatly the number of potential hits.

On the other hand, and as mentioned above, selenoproteins are as any other gene or protein in their remaining characteristics. One such standard feature of importance for the gene prediction purpose, is the typical usage of codons found in selenoprotein coding sequences (see Methods). We have shown that selenoprotein genes are prototypical in this regard and, of greater biological and computational implications, we proved that the region comprised between the recoded UGA codon and the real stop signal exhibits a similar codon usage bias (Castellano et al., 2001). This is of outstanding interest because permits, to a degree under discussion below, the computational read-through of only those UGA codons across which a coding potential function measures equivalently as coding. One should expect a sharp decrease in coding bias when entering a non-coding region. Let's say when bypassing a real stop codon and entering the 3′ UTR.

Another normal feature is the nonrestrictive situation of the Sec residue within the amino acid chain or the Sec codon in the gene structure, as it happens for the majority of the twenty standard amino acids and other codons. However, selenoproteincan be classified into two groups according to Sec location. One group includes proteins containing Sec in the N-terminal portion of short domains. These proteins are largely alpha-beta proteins, and Sec is often located in the loop between a beta-strand and an alpha-helix according to secondary structure predictions. In these regard, we have carried out the prediction of the secondary structures of all known human selenoprotein with the `predator` program (data not shown) and use these results as a signature to filter out false positive hits.

The second group of eukaryotic selenoprotein is characterized by the presence of Sec in the C-terminal region , usually three or less residues from the real stop codon (eg. thioredoxin reductase). This terminal situation of the selenocysteine residue can be troublesome for our approach, because the lack of a long ORF after the Sec codon compounds the accuracy of gene prediction programs. Moreover, short but coding ORFs can have negative coding potential measures, which are neither statistically nor biologically significant. To face this problem, the modified version of the `geneid` program used to predict exons with in-frame UGAs, does not compute the coding potential for exons up to 9 nucleotides. This modification resulted in an increase of sensitivity in the prediction of such genes.

As mentioned before, when selenoprotein families are compared, besides having in common a Sec residue in their protein sequence, do not have sequence homology, similar structures or related functions (Hatfield and Gladyshev, mini-review). Thus, the finding a of novel selenoprotein gene does not necessarily aids the search for other selenoprotein families. However, it raises interesting questions about the introduction of the UGA codon in the genetic code and, therefore, the introduction of Sec into proteins (see below). However, conservation within families across the whole eukaryotic domain is high, which provides a comparative method to test selenoprotein predictions.

5.2 Computational methods

Any computational method to be developed has to take into account one or more of these gene-defining characteristics. However, and as we will see, the ill-defined nature of these specific signals (or our limitations to biologically and computationally represent them) has lead to a progressive coordination of techniques to better analyze the ever-growing sequence data with a reasonable Sn and Sp.

First computational attempts to find novel selenoprotein genes were developed and published independently by two research groups. Kryukov, Kryukov, and Gladyshev (1999) developed a computer program, `SECISearch 1.0`, able to identify selenoprotein genes by recognizing SECIS elements on the basis of their primary and, above all, secondary structure. This program nicely joined two existing ones: 1) `patscan`, a pattern-matching program in C (see Introduction); 2) `RNAfold`, a C program for secondary structure prediction and free energy evaluation from the `Vienna RNA package` (see Introduction); and add 3) `RNAnice`, a perl module to visualize predicted SECIS structure. Modules were linked through perl scripting and interaction with the software was done on a web-based interface. When SECISearch was applied to search the human dbEST, two new mammalian selenoproteins, designated SelT and SelR, were identified.

At the same time, Lescure et al. (1999) presented a similar computational screen for selenoproteins on a similar set of target sequences. Their strategy relied on the pattern search program `rnamot` but it did not included any thermodynamic assessment, which lead to a greater experimental verification task. In this case, three novel mammalian selenoproteins, named SelX, SelN and SelZ (with two alternative splice forms, SelZf1 and SelZf2), were uncovered.

At that time, the existence of mammalian selenoprotein other than those previously characterized was predicted by works based on selenium labeling experiments in rats (Behne et al., 1996), but did not lead to their sequence identification. These experimental results and the availability of databases of transcript sequences, stimulated the development of novel computational algorithms for selenoprotein detection. The originality of these seminal approaches resided in taking advantage of the obligatory presence of the SECIS element in the selenoprotein mRNAs, instead of focusing on the amino acid sequence of the protein. This was a wise strategy because, as already suggested by data available at that time, selenoprotein families do not have sequence homology, similar structures or related functions. Therefore, methods such as `blast` or `fasta` were inappropriate. Furthermore, although the SECIS hairpin is well conserved at the secondary structure level, that is, the overall structure of helices and loops, the extent of primary sequence conservation is rather poor. Again, searches based on pure sequence similarity of the SECIS element were not useful to unveil unknown selenoprotein genes from the anonymous sequence data. Consequently, these authors introduced the still *de facto* standard method to search for SECIS structures.

However, a clear disadvantage of this approach is the inability to detect genes that contain SECIS signals other than the specified in the derived pattern. That is to say, that secondary structures that substantially differ from known ones are overlooked. As mentioned above, the extent of diversity in SECIS structures is still under research. In addition, the low specificity of SECIS searches produces a large number of predictions when applied to eukaryotic genomes. Therefore, this single approach is impractical when applied to complex genomic sequences.

5.3 The fly selenoproteome

After such an analysis of selenoprotein from the biological and computational point of view, we decided to attempt the screening of the *D. melanogaster* genome. As mentioned above, no selenoproteins were known in this model organism at that time. The main difference between a genome sequence and a database of ESTs, is that, while in size can be similar, the latter is highly redundant. This ensures that the number of unique hits, for example predicted SECIS, is reasonably low. At least, after the post-processing steps to validate them (see above). On the contrary, we realized that the number of potential SECIS structures in the fly genome exceed by far our ability to process them. Logically, we knew that the majority of them were biologically nonfunctional and that lied in unsuitable genomic regions as intergenic, intronic, exonic (coding), 5'UTR, RNA coding genes, opposite strand to protein coding genes sequences.

In the lack of a perfect annotation of the *D. melanogaster* gene complement, and in accordance to our experience, we decided to couple the prediction of genes with that of SECIS elements. In such a way, that all the predicted elements, and that includes besides standard exons, exons with a TGA in-frame and SECIS elements compete for a place in the final set of predictions. In addition, a rule was established so no gene with a TGA in-frame could be predicted unless there was a potential SECIS structure at the right distance downstream of it.

It should be noted here, that the prediction of selenoprotein genes relies on but aims in an opposite direction than the general prediction of genes in a genome of interest. While any *ab initio* gene prediction pipeline tries to maximize the overall accuracy of the proposed gene models (with no declared preference for any particular type), the prediction of selenoproteins demands focusing the accuracy on such nonstandard genes. With this in mind, the optimal prediction of selenoprotein genes is an trade off between a decreased general gene prediction accuracy and a greater precision on genes with a TGA in-frame. Furthermore, the partial prediction of selenoprotein genes, as long as it includes the Sec residue and the region around, is, for us, perfectly valid. When the goal is to uncover novel selenoprotein genes, previously missed and difficult to predict, the description of the exact gene structure can await posterior refinement.

This coordination of genes and SECIS elements has, as a main advantage, the mutual exclusion of these features along the genome and the gene prediction step. That is, if a false positive SECIS is predicted in an intergenic region where no gene is found, that SECIS element is not included. Or, if the SECIS overlaps a correct exon... And so on.

Independently, a similar work was carried out in Gladyshev's lab. The main difference with our approach was that they made use of the available annotation of the standard genes in this organism. In other words, they coordinated the presence of a predicted SECIS element downstream of an annotated gene, which, in case of a selenoprotein had to be mis-predicted. This approach is risky, because depending on the characteristics of the selenoprotein (eg. position of the Sec residue), the gene may not be predicted at all. Their screening resulted in two novel selenoprotein families, which happened also to be the SelH and SelK families. However, we cannot yet claim that the fly selenoproteome consists of only of three selenoprotein families accounting for ? genes. Two reasons for this. First, experiments of Se labeling in this organism show a band with no associated selenoprotein and second the use of only two similar methods to screen this genome. Further analysis are needed.

5.4 The Human selenoproteome

The size of the human genome is about 25 times larger than the fly one (3000 Mb Vs. 120 Mb, respectively), while having, maybe, only twice the number of genes [10]. This single fact is of many consequences in the biological and the computational analysis of the human genome. First, a longer but less gene-rich genome sequence means longer intergenic regions and, in the case of human, also means longer introns. This is a major drawback for our approach, because gene prediction programs performed substantially worse on such genomes, which have more room for errors. With a reduced accuracy on standard genes, the introduction of an additional exon-defining signal can be disastrous. For the prediction of SECIS signals, with such a poor sequence identity, the number of false positive predictions can be overwhelming. In conclusion, the *in silico* techniques successfully used in the screening of the fly genome, were of limited application in the larger and more complex mammalian genomes.

To our aid, it comes old and new ideas. First, and as stated above, we do not long for the highest overall accuracy to obtain a correct prediction of selenoprotein genes. Though, it is valuable to keep the drop as low as possible and, above all, to do it correctly when it comes to selenoproteins. Second, comparative genomics has not been, so far, exploited to its full possibilities. Besides helping to discard predicted peptides by checking conservation beyond the in-frame Sec codon comparing to genomic and transcript sequences, the SECIS element itself can be used. This was based on our observation that human, mouse and rat SECIS elements in orthologous selenoprotein genes exhibited detectable sequence similarity. The genome of the mouse and the rat were under heavy sequencing at that time, and a substantial number of shotgun reads were therefore available. Therefore, we input to `geneid` only those SECIS sequences validated in the rodent genomes. In addition, we predicted with `geneid`, and regardless of SECIS elements, all possible human genes that were interrupted by an in-frame TGA codon. These proteins were analyzed by `blastp` and `tblastn` against all eukaryotic proteins and full genome sequences available at that time. This procedure was designed to identify sequences with homology in TGA-flanking regions, which either conserve TGA or replace TGA with TGC or TGT (Cys codons). Finally, we introduced and independent and conversed approach that makes use of the habitual presence of Cys-containing homologues of selenoprotein genes. We analyzed by `tblastn` all human standard proteins against all human ESTs to identify paralogs that contain TGA in place of a Cys codon.

The use of these three independent approaches makes a difference with the fly analysis. While, we cannot claim in the fly that all selenoproteins have been found, we believe that in human and rodent genomes all or almost all selenoproteins have been identified in this work.

[10] The issue of the number of genes in the human and others genomes is controversial. For a revision see ? and references therein. It is also advisable to frequently consult the major genome browsers, Ensembl, NCBI and Golden Path, for the most up-to-date estimations of such numbers.

5.5 The Takifugu selenoproteome

In another twist on the application of comparative genomics to predict selenoprotein genes, we made use of two genomes, human and *Takifugu*, to pinpoint novel selenoprotein genes. This fully comparative approach was necessary if a SECIS-independent search was to be done. As mentioned above, the SECIS element can be used as a constraint to the prediction of in-frame TGAs but otherwise limiting the possible variation of this RNA secondary structure. Here, comparisons between genomes act as an alternative constraint. In brief, a genome-wide prediction of genes with in-frame TGA codons in the two species was followed by the comparison of prediction between genomes. Those alignments in which conservation in regions flanking the TGA codon was suggestive of selenocysteine coding function were kept and further analyzed. This resulted in a novel selenoprotein family, termed SelU, which showed and interesting pattern of distribution among eukaryotes (see below). However, one worrying limitation of this approach is that it is based on the conservation of the Sec residue context. Therefore, sporadic selenoproteins that are unique to a particular species, or those in which Sec is the last or penultimate amino acid, might be missed. This first objection could seem unimportant at first sight, but our subsequent studies on the distribution of selenoproteins across the eukaryotic domain showed it to be pertinent.

5.6 The Tetraodon selenoproteome

Usual computational tools lack the ability to correctly assign UGA function. Consequently, there are numerous examples of misinterpretations of UGA codons as both Sec codons and terminators, including annotations of the human genome where no selenoproteins have been correctly predicted. Within the *Tetraodon* genome sequencing effort, two goals in relation to selenoproteins were defined. First the annotation (or re-annotation) of known selenoproteins. Second, the finding of novel selenoproteinfamilies which this genome could encompass.

The gene annotation protocol in this genome (see Methods), included the use of two standard *ab initio* gene prediction programs, `genscan` and `geneid`, the use of homology data such as cDNAs, ESTs and protein sequences and the comparison between genomes through the `exofish` procedure. At the end of the day, all this information (annotated features in the genome) was input to the `gaze` system, which is able to combined it and output the final best gene models. None of these methods contemplates without change the possibility of having exons with a TGA in-frame. Therefore, this protocol could not face the existence of exons with a TGA in-frame and selenoprotein genes were misannotated, that is to say, partially predicted or even not predicted at all. In addition, The SECIS element was also overlooked. To overcome this problem, we devised a reannotation protocol (see Methods). 18 known selenoprotein families were reannotated and, additionally, a novel selenoprotein, termed SelJ, was found. Its biological function is yet unclear.

As a methodological conclusion, it can be said that the identification of novel selenoprotein genes in eukaryotes is a difficult task. Accordingly, over the last five years, computational methodologies have become more and more complex. *In silico* approaches, which started in a simple but, at the same time, complicated search for SECIS elements in transcript databases, have been refined to face the advent of full genome sequences. In this transition, a number of complementary an independent methods have been developed. First, methods that coordinate the prediction of SECIS elements with the prediction of either standard genes or genes with a TGA codon in-frame were successfully applied (REF, REF). Second, truly SECIS-independent algorithms were designed and, finally, comparative genomic approaches have become increasingly important. It is the interaction of these independent methods the clue for a complete description of the eukaryotic selenoproteomes.

5.7 Selenoprotein distribution

When one is confronted to the analysis of the distribution of the eukaryotic selenoproteins, that is to say, the pattern of occurrences of Sec-containing proteins and their Cys-and-others homologues across the eukaryotic lineage, a dual axis of protein families and organisms is to be considered and plotted (Figure ?). The main reason for this, is the discrete and scattered distribution of each particular selenoprotein family over the eukaryotic genomes, which makes necessary to carefully describe each family in each species, somehow, separately and with no generalities applied.

In Figure ?, such description is graphically displayed. It shows not only the presence or absence of a gene in a given genome, but the Sec or Cys-nature of such gene. To our knowledge, it is the most comprehensive chart describing eukaryotic selenoproteomes available. However, to interpret this plot, several disclaiming issues are to be bear in mind. They are 1) only a few eukaryotic organisms are included, and they represent a wide phylogenetic range but, obviously, a tiny fraction of the eukaryotic lineage; 2) the full genomic sequence is not available for all of them, which could limit the finding of some protein families (though, available transcript data has been added to this search to prevent missing gene information); 3) vertebrate and specially mammalian genomes have been surveyed, experimentally and computationally, much extensively than others, which may account for the overrepresentation of selenoprotein families in these taxa; 4) yellow boxes indicate the existence of an homologue gene but of an unknown Sec or Cys nature; and 5) for the species shown, genes involved in the metabolic process of Se and Sec have been reported for all but the land plant *A. thaliana* and the fission yeast *S. cerevisae*. The SPS2 family would be an example of this.

A rapid look to Figure ?, could provide wrong conclusions. As mentioned, while it is blatantly clear that, currently, vertebrates accumulate the majority of selenoprotein families, it is also true that other genomes have just started to be analyzed. Thus, other selenoprotein genes may exist.

This hypothesis is reinforced by some recent discoveries: 1) the methionine-S-sulphoxide reductase (MsrA) occurs as a selenoprotein in *Chlamydomonas reinhardtii*, a green algae, but has Cys in vertebrates (including mammals) and other invertebrates (REF); 2) the SelU family is a true selenoprotein in fish, birds, echinoderms, green algae and diatoms, while is Cys-containing in mammals, land plants, arthropods, worms, amphibians, tunicates and slime molds. Apparently, yeast and flies (among arthropods) lack proteins of this family(**?**); 3) the SelJ family, which is widely distributed in, but restricted to, actinopterygians among vertebrates. Furthermore, it may be the first selenoprotein family without even a Cys-counterpart in mammals; and 4) a glutathione peroxidase homologue (GPx6) was reported to have Sec in humans and pigs, but Cys in rodents(**?**)

These observations stand for a high diversity in the short and long phylogenetic scale, and suggests that the usage of Se Vs. S, or that of Sec Vs. Cys, is of an unordered and scattered nature. We therefore anticipate, that other taxa-specific selenoproteins probably exist.

5.8 Selenoprotein evolution

Theory holds that the number of selenoprotein genes increases from less to more complex genomes. In other words, it has longer been assumed that mammals and other close vertebrates recapitulate the eukaryotic selenoproteome. This is may not be true. This assumption was reasonable at the time that all selenoprotein families were always found as true Sec-containing proteins in mammalian genomes and as punctual Sec or Cys-genes in others. The analysis of the *Drosophila* genome also backed this scenario. The two novel selenoprotein families found in this genome were shown to be also Sec-proteins in mammals and other vertebrates, while had Cys in other invertebrate species. However, as summarized above, the new findings strongly argue for a no restrictive and scattered distribution of selenoprotein genes. Therefore, each family may have particular history, a successive chain of events which have lead to the discrete and individual arrangement of Sec and Cys-homologs across genomes and species.

5.9 Closing remarks

In this chapter, our contribution to the understanding of the selenium-dependent world in eukaryotes has been presented in detail and debated accordingly. As closing remarks, only highlight that the research described here is due to a bunch of computational and experimental biologists and demonstrates the power of the combined *in silico*, *in vitro* and *in vivo* approaches toward a better understanding of living systems.

Conclusions

In short, the research presented here has contributed to:

1. The computational prediction of 9 novel selenoprotein families on several eukaryotic organisms[1]: 1) the SelH and SelK families in *D. melanogaster*; 2) the SelI, SelO, SelS, SelV and GPx6 in *H. sapiens*; 3) the SelU family in *T. rubripes*; and, finally, 4) the SelJ family in *T. nigroviridis*. These descriptive results have set the current count of eukaryotic selenoproteins to 20 distinct families; [2]

2. The development of several independent computational methods to identify selenoprotein genes: 1) the prediction of SECIS elements through RNA secondary structure pattern matching coupled with thermodynamic stability assessments; 2) the prediction of genes having a TGA codon in-frame with a dedicated gene prediction program (a modification of `geneid`); and 3) the prediction of standard genes (Cys-containing genes) with `geneid` and the subsequent search of selenoprotein homologs with sequence similarity tools (`blast`). The combined use of these methods and the additional application of comparative genomics strategies has been shown to be an unvaluable tool for identifying new selenoprotein genes;[3]

3. The definition of the selenoproteome in the genomes analyzed. Our different methodologies, when applied altogether, not only provide yet unknown selenoproteins but reasonably ensure that all, or almost all, such genes have been found in a genome;[4]

4. The annotation and reannotation of genomes. While selenoprotein genes are misannotated in the majority of sequenced genomes, we have already started collaborations to correct this problem. The first example is the analysis of the *Tetraodon* genome;[5]

5. The better knowledge of the selenoprotein distribution throughout the eukaryotic domain. We have shown that mammals have no special privilege regarding selenoproteins. On the contrary, the eukaryotic selenoproteome is a mosaic, with taxa-specific selenoprotein genes distributed around in a family and organism-dependent fashion and with no necessary increase in the number of selenoprotein genes from low complexity eukaryotes to vertebrates;[6]

6. The better knowledge of the evolution of selenoproteins. This distribution is suggestive of a particular and independent history for each family and taxa. In this scenario, and probably mediated by an ongoing evolutionary process of Sec/Cys interconversion dependent as much as on functional constraints as on contingent events, the existence of sporadic selenoproteins that are unique to a particular species (or group of species) is reasonable.

[1] With a subsequent positive experimental validation.

[2] They are: 15kDa, MsrA, DI (DI1, DI2, DI3), GPx (GPx1, GPx2, GPx3, GPx4, GPx6), SelH, SelI, SelJ, SelK, SelM, SelN, SelO, SelP, SelR, SelS, SPS2, SelT, TR (TR1, TR2, TR3), SelU, SelV and SelW. Based on functional criterias, some families have also been divided into subfamilies as indicated in parenthesis.

[3] Because of the size and complexity of the genomes analyzed and the limitations of the methods, a reasonable trade off between Sn and Sp was only achieved when coordinating two or more of these approaches.

[4] Although this set of methods aims at the exhaustive characterization of eukaryotic selenoproteomes, it is certainly possible that punctual selenoprotein genes are missed. However, recognition of the majority of known selenoproteins by these procedures and the independence between approaches suggest that all or virtually all selenoproteins have been found in some of the genomes analyzed. Furthermore, SECIS-independent approaches are able to detect genes that contain a noncanonical SECIS element.

[5] We are now working in chicken and would like to reannotate genomes available at `ENSEMBL`.

[6] To clarify this point, more genomes widespread across the eukaryotic lineage need to be scanned.

Speculations

This chapter gathers, under the warning epigraph of speculative hypothesis on selenoproteins, the most provocative and controversial issues in this field that merit a section to itselfs. For the sake of completeness, these questions are addressed in the light of the findings presented here, though no conclusive statements are whatsoever intended.

Termination codons the UAA, UAG and UGA stop codons are completely or partially (dually) recoded in many genomes. This make sense because these codons are used in a single position in proteins, thus, punctual recoding may be easier to achieve. Does recoding events need a scarcely used codon? is decoding competition with release factors less problematic than with tRNAs?

Amino acidic codons so far, only the GUG (Val), UUG (Leu) and AUU (Ile) have been found to be redefined to function as start codons. These codons, when at internal positions code for the amino acids indicated in parenthesis, but when they function as an initiator they specify methionine. However, other amino acidic codons are seldom used in some genomes and could be the target of a recoding mechanism. Therefore, are there other amino acidic codons dually coded?

New amino acids one can think that the number of *standard* amino acids[1], which has just jumped from 20 to 22, is possibly increasing in the near future. Will they be part of a complete or dual decoding?

A step in evolution dual decoding is yet a poorly understood mechanism and its relation to the general evolution of the genetic code and the raising of new variants unclear. Is dual decoding an intermediate step towards the complete recoding of a codon? a remnant of an ancient codon meaning? or an independent process designed to introduce diversity to protein sequences and functions? the study of selenoprotein genes advocates for the latter possibility.

[1] Understood here as those amino acids which are incorporated cotranslationally by means of an specific codon-anticodon pair, in contrast to those produced by post translational modifications.

Epilogue

In[2] some recent textbooks, non-canonical genetic codes, such as UGA (Trp) in *Mycoplasma capricolum* and UAR (Gln) in ciliated protozoans, are dealt with simply as exceptions. Similarly, and in these same texts, dual-non-canonical decoding, such as UGA (Sec) in selenoproteins and maybe in UAG (pyrrolysine) in *Methanosarcina barkeri*, are considered no more than mere particularities of some protein families. Such treatment gives the impression that nearly all organisms use both the canonical genetic code and the canonical single decoding rules. But is this true? Despite an enormous diversity of organisms, all of which are derived from a single group of ancestors, until recently, we have to admit, only a handful of organisms has been examined genetically.

While the existence of non-canonical genetic codes is now generally accepted in both mitochondria and nuclear genomes, at least within the field of the genetic code evolution, the idea of a widespread dual-decoding of codons is still provocative and controversial. This is understandable because data is scarce and, objectively, it may not lead to more than to a testimonial use of dual decoding rules in a few protein families (with either use of one of the traditional twenty amino acids or others). However, it is my impression that these two phenomena, the existence of complete and partial deviations to the standard genetic code, are intimately related. The link, being the evolving (and gradual) nature of the genetic code and the unnecessary concept of an ultimate set of *optimal* decoding rules.[3]

The crux of the matter is whether the dual decoding will be shown to be a general mechanisms, affecting many protein families in many organisms, or it will just become a biological curiosity. No definitive answer can be given now. However, the availability of more and more genomes and the development of experimental and computational techniques as the ones presented here, may shed light to this problem in the years to come.

In any case, one can wonder about the relation between this mechanism and the evolution of the standard genetic code and its variants. The case of selenoproteins suggest an autonomous mechanism to increase protein variety, rather than a temporal and intermediary situation between a canonical meaning an a noncanonical one. However, no possibility can yet be ruled out.

[2] Epilogue adapted from “The Evolution of the Genetic code” by S. Osawa. Oxford Press (1995).

[3] This final stage in biology is otherwise meaningless and misleading.

List of abbreviations

For the three-letter abbreviations of amino acids, see Appendix C.1 in 81.

A adenine

aa amino acid

C cytosine

cDNA complementary deoxyribonucleic acid

CDS coding sequence

DI deiodinase

DNA deoxyribonucleic acid

EST expressed sequence tag

G guanine

GPx glutathione peroxidase

mRNA messenger ribonucleic acid

MsR methionine sulforeductase

nt nucleotide

ORF open-reading frame

PDF portable document format

RNA ribonucleic acid

SECIS selenocysteine insertion sequence

Sp specificity

SPS selenophosphate synthetase

Sn sensitivity

T thymine

TR thiorredoxin reductase

tRNA transfer ribonucleic acid

U uracil

uORF upstream open-reading frame

URL universal resource locator

UTR unstranslated region

Glossary

This is a compilation of concepts and their definitions as understood and used throughout these pages and these years.[1]

amino acidic codons those that code for an amino acid. This nomenclature is prefered to the widespread "sense" codon term

complex eukaryotes those species with a more intricate and difficult to comprehend biology (eg. humans are more complex than yeast). This nomenclature is prefered to the widely used "higher eukaryotes". The same commentary for "Lower eukaryotes".

Homology refers to genes (or any other biological feature) that are related through a common evolutionary ancestor. Orthology, paralogy and xenology are homology subtypes. That is, they define a specific type of relationship between genes over space and time.

Mosaic made of many distinct element, which may behave differently in relation to biological features (eg. genomic regions have different G+C content or mutation rate). In selenoproteins Se distribute in a mosaic fashion across species.

non-amino acidic codons those that do not code for an amino acid. Also called termination codons. This nomenclature is prefered to the widespread "nonsense" codon term

Orthology refers to genes that are present in different organisms and have evolved from a common ancestral gene by speciation.

Paralogy refers to genes that are present in the same organism or in different organisms and have evolved from a common ancestral gene by a gene duplication event. If this gene duplication event took place before a speciation event, these are paralogous genes in different genomes.

Sec/Cys (or Cys/Sec) either Sec or Cys

Sec-Cys (or Cys-Sec) Alignment of Sec with Cys

Synteny originally in the same string (strand), but here refers to homologous regions in two different genomes with conserved gene type and order.

[1] With no intention of proselitism.

A genetic code chart

A	Ala	Alanine	**GCA** GCC **GCG** GCU
C	Cys	Cysteine	**UGC** UGU
D	Asp	Aspartic acid	**GAC** GAU
E	Glu	Glutamic acid	**GAA** GAG
F	Phe	Phenylalanine	**UUC** UUU
G	Gly	Glycine	**GGA** GGC **GGG** GGU
H	His	Histidine	**CAC** CAU
I	Ile	Isoleucine	**AUA** AUC **AUU**
K	Lys	Lysine	**AAA** AAG
L	Leu	Leucine	**UUA** UUG **CUA** CUC **CUG** CUU
M	Met	Metionine	**AUG**
N	Asn	Asparagine	**AAC** AAU
P	Pro	Proline	**CCA** CCC **CCG** CCU
Q	Gln	Glutamine	**CAA** CAG
R	Arg	Arginine	**AGA** AGG **CGA** CGC **CGG** CGU
S	Ser	Serine	**AGC** AGU **UCA** UCC **UCG** UCU
T	Thr	Threonine	**ACA** ACC **ACG** ACU
V	Val	Valine	**GUA** GUC **GUG** GUU
W	Trp	Tryptophan	**UGG**
Y	Tyr	Tyrosine	**UAC** UAU

Table C.1: The standard genetic code

The GNU General Public License

Version 2, June 1991

59 Temple Place - Suite 330, Boston, MA 02111-1307, USA

Preamble

The licenses for most software are designed to take away your freedom to share and change it. By contrast, the GNU General Public License is intended to guarantee your freedom to share and change free software—to make sure the software is free for all its users. This General Public License applies to most of the Free Software Foundation's software and to any other program whose authors commit to using it. (Some other Free Software Foundation software is covered by the GNU Library General Public License instead.) You can apply it to your programs, too.

When we speak of free software, we are referring to freedom, not price. Our General Public Licenses are designed to make sure that you have the freedom to distribute copies of free software (and charge for this service if you wish), that you receive source code or can get it if you want it, that you can change the software or use pieces of it in new free programs; and that you know you can do these things.

To protect your rights, we need to make restrictions that forbid anyone to deny you these rights or to ask you to surrender the rights. These restrictions translate to certain responsibilities for you if you distribute copies of the software, or if you modify it.

For example, if you distribute copies of such a program, whether gratis or for a fee, you must give the recipients all the rights that you have. You must make sure that they, too, receive or can get the source code. And you must show them these terms so they know their rights.

We protect your rights with two steps: (1) copyright the software, and (2) offer you this license which gives you legal permission to copy, distribute and/or modify the software.

Also, for each author's protection and ours, we want to make certain that everyone understands that there is no warranty for this free software. If the software is modified by someone else and passed on, we want its recipients to know that what they have is not the original, so that any problems introduced by others will not reflect on the original authors' reputations.

Finally, any free program is threatened constantly by software patents. We wish to avoid the danger that redistributors of a free program will individually obtain patent licenses, in effect making the program proprietary. To prevent this, we have made it clear that any patent must be licensed for everyone's free use or not licensed at all.

The precise terms and conditions for copying, distribution and modification follow.

TERMS AND CONDITIONS FOR COPYING, DISTRIBUTION AND MODIFICATION

0. This License applies to any program or other work which contains a notice placed by the copyright holder saying it may be distributed under the terms of this General Public License. The “Program”, below, refers to any such program or work, and a “work based on the Program” means either the Program or any derivative work under copyright law: that is to say, a work containing the Program or a portion of it, either verbatim or with modifications and/or translated into another language. (Hereinafter, translation is included without limitation in the term “modification”.) Each licensee is addressed as “you”.

   Activities other than copying, distribution and modification are not covered by this License; they are outside its scope. The act of running the Program is not restricted, and the output from the Program is covered only if its contents constitute a work based on the Program (independent of having been made by running the Program). Whether that is true depends on what the Program does.

1. You may copy and distribute verbatim copies of the Program’s source code as you receive it, in any medium, provided that you conspicuously and appropriately publish on each copy an appropriate copyright notice and disclaimer of warranty; keep intact all the notices that refer to this License and to the absence of any warranty; and give any other recipients of the Program a copy of this License along with the Program.

   You may charge a fee for the physical act of transferring a copy, and you may at your option offer warranty protection in exchange for a fee.

2. You may modify your copy or copies of the Program or any portion of it, thus forming a work based on the Program, and copy and distribute such modifications or work under the terms of Section 1 above, provided that you also meet all of these conditions:

   (a) You must cause the modified files to carry prominent notices stating that you changed the files and the date of any change.

   (b) You must cause any work that you distribute or publish, that in whole or in part contains or is derived from the Program or any part thereof, to be licensed as a whole at no charge to all third parties under the terms of this License.

   (c) If the modified program normally reads commands interactively when run, you must cause it, when started running for such interactive use in the most ordinary way, to print or display an announcement including an appropriate copyright notice and a notice that there is no warranty (or else, saying that you provide a warranty) and that users may redistribute the program under these conditions, and telling the user how to view a copy of this License. (Exception: if the Program itself is interactive but does not normally print such an announcement, your work based on the Program is not required to print an announcement.)

   These requirements apply to the modified work as a whole. If identifiable sections of that work are not derived from the Program, and can be reasonably considered independent and separate works in themselves, then this License, and its terms, do not apply to those sections when you distribute them as separate works. But when you distribute the same sections as part of a whole which is a work based on the Program, the distribution of the whole must be on the terms of this License, whose permissions for other licensees extend to the entire whole, and thus to each and every part regardless of who wrote it.

   Thus, it is not the intent of this section to claim rights or contest your rights to work written entirely by you; rather, the intent is to exercise the right to control the distribution of derivative or collective works based on the Program.

   In addition, mere aggregation of another work not based on the Program with the Program (or with a work based on the Program) on a volume of a storage or distribution medium does not bring the other work under the scope of this License.

3. You may copy and distribute the Program (or a work based on it, under Section 2) in object code or executable form under the terms of Sections 1 and 2 above provided that you also do one of the following:

(a) Accompany it with the complete corresponding machine-readable source code, which must be distributed under the terms of Sections 1 and 2 above on a medium customarily used for software interchange; or,

(b) Accompany it with a written offer, valid for at least three years, to give any third party, for a charge no more than your cost of physically performing source distribution, a complete machine-readable copy of the corresponding source code, to be distributed under the terms of Sections 1 and 2 above on a medium customarily used for software interchange; or,

(c) Accompany it with the information you received as to the offer to distribute corresponding source code. (This alternative is allowed only for noncommercial distribution and only if you received the program in object code or executable form with such an offer, in accord with Subsection b above.)

The source code for a work means the preferred form of the work for making modifications to it. For an executable work, complete source code means all the source code for all modules it contains, plus any associated interface definition files, plus the scripts used to control compilation and installation of the executable. However, as a special exception, the source code distributed need not include anything that is normally distributed (in either source or binary form) with the major components (compiler, kernel, and so on) of the operating system on which the executable runs, unless that component itself accompanies the executable.

If distribution of executable or object code is made by offering access to copy from a designated place, then offering equivalent access to copy the source code from the same place counts as distribution of the source code, even though third parties are not compelled to copy the source along with the object code.

4. You may not copy, modify, sublicense, or distribute the Program except as expressly provided under this License. Any attempt otherwise to copy, modify, sublicense or distribute the Program is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

5. You are not required to accept this License, since you have not signed it. However, nothing else grants you permission to modify or distribute the Program or its derivative works. These actions are prohibited by law if you do not accept this License. Therefore, by modifying or distributing the Program (or any work based on the Program), you indicate your acceptance of this License to do so, and all its terms and conditions for copying, distributing or modifying the Program or works based on it.

6. Each time you redistribute the Program (or any work based on the Program), the recipient automatically receives a license from the original licensor to copy, distribute or modify the Program subject to these terms and conditions. You may not impose any further restrictions on the recipients' exercise of the rights granted herein. You are not responsible for enforcing compliance by third parties to this License.

7. If, as a consequence of a court judgment or allegation of patent infringement or for any other reason (not limited to patent issues), conditions are imposed on you (whether by court order, agreement or otherwise) that contradict the conditions of this License, they do not excuse you from the conditions of this License. If you cannot distribute so as to satisfy simultaneously your obligations under this License and any other pertinent obligations, then as a consequence you may not distribute the Program at all. For example, if a patent license would not permit royalty-free redistribution of the Program by all those who receive copies directly or indirectly through you, then the only way you could satisfy both it and this License would be to refrain entirely from distribution of the Program.

   If any portion of this section is held invalid or unenforceable under any particular circumstance, the balance of the section is intended to apply and the section as a whole is intended to apply in other circumstances.

   It is not the purpose of this section to induce you to infringe any patents or other property right claims or to contest validity of any such claims; this section has the sole purpose of protecting the integrity of the free software distribution system, which is implemented by public license practices. Many people have made generous contributions to the wide range of software distributed through that system in reliance on consistent application of that system; it is up to the author/donor to decide if he or she is willing to distribute software through any other system and a licensee cannot impose that choice.

This section is intended to make thoroughly clear what is believed to be a consequence of the rest of this License.

8. If the distribution and/or use of the Program is restricted in certain countries either by patents or by copyrighted interfaces, the original copyright holder who places the Program under this License may add an explicit geographical distribution limitation excluding those countries, so that distribution is permitted only in or among countries not thus excluded. In such case, this License incorporates the limitation as if written in the body of this License.

9. The Free Software Foundation may publish revised and/or new versions of the General Public License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns.

   Each version is given a distinguishing version number. If the Program specifies a version number of this License which applies to it and "any later version", you have the option of following the terms and conditions either of that version or of any later version published by the Free Software Foundation. If the Program does not specify a version number of this License, you may choose any version ever published by the Free Software Foundation.

10. If you wish to incorporate parts of the Program into other free programs whose distribution conditions are different, write to the author to ask for permission. For software which is copyrighted by the Free Software Foundation, write to the Free Software Foundation; we sometimes make exceptions for this. Our decision will be guided by the two goals of preserving the free status of all derivatives of our free software and of promoting the sharing and reuse of software generally.

No Warranty

11. Because the program is licensed free of charge, there is no warranty for the program, to the extent permitted by applicable law. Except when otherwise stated in writing the copyright holders and/or other parties provide the program "as is" without warranty of any kind, either expressed or implied, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose. The entire risk as to the quality and performance of the program is with you. Should the program prove defective, you assume the cost of all necessary servicing, repair or correction.

12. In no event unless required by applicable law or agreed to in writing will any copyright holder, or any other party who may modify and/or redistribute the program as permitted above, be liable to you for damages, including any general, special, incidental or consequential damages arising out of the use or inability to use the program (including but not limited to loss of data or data being rendered inaccurate or losses sustained by you or third parties or a failure of the program to operate with any other programs), even if such holder or other party has been advised of the possibility of such damages.

End of Terms and Conditions

Appendix: How to Apply These Terms to Your New Programs

If you develop a new program, and you want it to be of the greatest possible use to the public, the best way to achieve this is to make it free software which everyone can redistribute and change under these terms.

To do so, attach the following notices to the program. It is safest to attach them to the start of each source file to most effectively convey the exclusion of warranty; and each file should have at least the "copyright" line and a pointer to where the full notice is found.

> one line to give the program's name and a brief idea of what it does.
> Copyright (C) yyyy name of author
>
> This program is free software; you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation; either version 2 of the License, or (at your option) any later version.
>
> This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details.
>
> You should have received a copy of the GNU General Public License along with this program; if not, write to the Free Software Foundation, Inc., 59 Temple Place - Suite 330, Boston, MA 02111-1307, USA.

Also add information on how to contact you by electronic and paper mail.

If the program is interactive, make it output a short notice like this when it starts in an interactive mode:

> Gnomovision version 69, Copyright (C) yyyy name of author
> Gnomovision comes with ABSOLUTELY NO WARRANTY; for details type 'show w'.
> This is free software, and you are welcome to redistribute it under certain conditions; type 'show c' for details.

The hypothetical commands `show w` and `show c` should show the appropriate parts of the General Public License. Of course, the commands you use may be called something other than `show w` and `show c`; they could even be mouse-clicks or menu items—whatever suits your program.

You should also get your employer (if you work as a programmer) or your school, if any, to sign a "copyright disclaimer" for the program, if necessary. Here is a sample; alter the names:

> Yoyodyne, Inc., hereby disclaims all copyright interest in the program
> 'Gnomovision' (which makes passes at compilers) written by James Hacker.
>
> signature of Ty Coon, 1 April 1989
> Ty Coon, President of Vice

This General Public License does not permit incorporating your program into proprietary programs. If your program is a subroutine library, you may consider it more useful to permit linking proprietary applications with the library. If this is what you want to do, use the GNU Library General Public License instead of this License.

Human genome and human rights

Introduction

The Universal Declaration on the Human Genome and Human Rights, which was adopted unanimously and by acclamation by the General Conference of UNESCO at its 29th session on 11 November 1997, is the first universal instrument in the field of biology. The uncontested merit of this text resides in the balance it strikes between safeguarding respect for human rights and fundamental freedoms and the need to ensure freedom of research.

Together with the Declaration, UNESCO's General Conference adopted a resolution for its implementation, which commits States to taking appropriate measures to promote the principles set out in the Declaration and encourage their implementation.

The moral commitment entered into by States in adopting the Universal Declaration on the Human Genome and Human Rights is a starting point, the beginning of international awareness of the need for ethical issues to be addressed in science and technology. It is now up to States, through the measures they decide to adopt, to put the Declaration into practice and thus ensure its continued existence.

***

The General Conference,

Recalling that the Preamble of UNESCO's Constitution refers to "the democratic principles of the dignity, equality and mutual respect of men", rejects any "doctrine of the inequality of men and races", stipulates "that the wide diffusion of culture, and the education of humanity for justice and liberty and peace are indispensable to the dignity of men and constitute a sacred duty which all the nations must fulfil in a spirit of mutual assistance and concern", proclaims that "peace must be founded upon the intellectual and moral solidarity of mankind", and states that the Organization seeks to advance "through the educational and scientific and cultural relations of the peoples of the world, the objectives of international peace and of the common welfare of mankind for which the United Nations Organization was established and which its Charter proclaims",

Solemnly recalling its attachment to the universal principles of human rights, affirmed in particular in the Universal Declaration of Human Rights of 10 December 1948 and in the two International United Nations Covenants on Economic, Social and Cultural Rights and on Civil and Political Rights of l6 December 1966, in the United Nations Convention on the Prevention and Punishment of the Crime of Genocide of 9 December 1948, the International United Nations Convention on the Elimination of All Forms of Racial Discrimination of 21 December 1965, the United Nations Declaration on the Rights of Mentally Retarded Persons of 20 December 1971, the United Nations Declaration on the Rights of Disabled Persons of 9 December 1975, the United Nations Convention on the Elimination of All Forms of Discrimination Against Women of 18 December 1979, the United Nations Declaration of Basic Principles of Justice for Victims of Crime and Abuse of Power of 29 November 1985, the United Nations Convention on the Rights of the Child of 20 November 1989, the United Nations Standard Rules on the Equalization of Opportunities for Persons with Disabilities of 20 December 1993, the Convention on the Prohibition of the Development, Production and Stockpiling of Bacteriological (Biological) and Toxin Weapons and on their Destruction of 16 December 1971, the UNESCO Convention against Discrimination in Education of 14 December 1960, the UNESCO Declaration of the Principles of International Cultural Co-operation of 4 November 1966, the UNESCO Recommendation on the Status of Scientific Researchers of 20 November 1974, the UNESCO Declaration on Race and Racial Prejudice of 27 November 1978, the ILO Convention (N 111) concerning Discrimination in Respect of Employment and Occupation of 25 June 1958 and the ILO Convention (N 169) concerning Indigenous and Tribal Peoples in Independent Countries of 27 June 1989,

Bearing in mind , and without prejudice to, the international instruments which could have a bearing on the applications of genetics in the field of intellectual property, inter alia the Bern Convention for the Protection of Literary and Artistic Works of 9 September 1886 and the UNESCO Universal Copyright Convention of 6 September 1952, as last revised in Paris on 24 July 1971, the Paris Convention for the Protection of Industrial Property of 20 March 1883, as last revised at Stockholm on 14 July 1967, the Budapest Treaty of the WIPO on International Recognition of the Deposit of Micro-Organisms for the Purposes of Patent Procedures of 28 April 1977, and the Trade Related Aspects of Intellectual Property Rights Agreement (TRIPs) annexed to the Agreement establishing the World Trade Organization, which entered into force on 1st January 1995,

Bearing in mind also the United Nations Convention on Biological Diversity of 5 June 1992 and emphasizing in that connection that the recognition of the genetic diversity of humanity must not give rise to any interpretation of a social or political nature which could call into question "the inherent dignity and (...) the equal and inalienable rights of all members of the human family", in accordance with the Preamble to the Universal Declaration of Human Rights,

Recalling 22 C/Resolution 13.1, 23 C/Resolution 13.1, 24 C/Resolution 13.1, 25 C/Resolutions 5.2 and 7.3, 27 C/Resolution 5.15 and 28 C/Resolutions 0.12, 2.1 and 2.2, urging UNESCO to promote and develop ethical studies, and the actions arising out of them, on the consequences of scientific and technological progress in the fields of biology and genetics, within the framework of respect for human rights and fundamental freedoms,

Recognizing that research on the human genome and the resulting applications open up vast prospects for progress in improving the health of individuals and of humankind as a whole, but emphasizing that such research should fully respect human dignity, freedom and human rights, as well as the prohibition of all forms of discrimination based on genetic characteristics,

Proclaims the principles that follow and adopts the present Declaration.

A Human dignity and the human

Article 1

The human genome underlies the fundamental unity of all members of the human family, as well as the recognition of their inherent dignity and diversity. In a symbolic sense, it is the heritage of humanity.

Article 2

(a) Everyone has a right to respect for their dignity and for their rights regardless of their genetic characteristics.

(b) That dignity makes it imperative not to reduce individuals to their genetic characteristics and to respect their uniqueness and diversity.

Article 3

The human genome, which by its nature evolves, is subject to mutations. It contains potentialities that are expressed differently according to each individual's natural and social environment including the individual's state of health, living conditions, nutrition and education.

Article 4

The human genome in its natural state shall not give rise to financial gains.

B Rights of the persons concerned

Article 5

(a) Research, treatment or diagnosis affecting an individual's genome shall be undertaken only after rigorous and prior assessment of the potential risks and benefits pertaining thereto and in accordance with any other requirement of national law.

(b) In all cases, the prior, free and informed consent of the person concerned shall be obtained. If the latter is not in a position to consent, consent or authorization shall be obtained in the manner prescribed by law, guided by the person's best interest.

(c) The right of each individual to decide whether or not to be informed of the results of genetic examination and the resulting consequences should be respected.

(d) In the case of research, protocols shall, in addition, be submitted for prior review in accordance with relevant national and international research standards or guidelines.

(e) If according to the law a person does not have the capacity to consent, research affecting his or her genome may only be carried out for his or her direct health benefit, subject to the authorization and the protective conditions prescribed by law. Research which does not have an expected direct health benefit may only be undertaken by way of exception, with the utmost restraint, exposing the person only to a minimal risk and minimal burden and if the research is intended to contribute to the health benefit of other persons in the same age category or with the same genetic condition, subject to the conditions prescribed by law, and provided such research is compatible with the protection of the individual's human rights.

Article 6

No one shall be subjected to discrimination based on genetic characteristics that is intended to infringe or has the effect of infringing human rights, fundamental freedoms and human dignity.

Article 7

Genetic data associated with an identifiable person and stored or processed for the purposes of research or any other purpose must be held confidential in the conditions set by law.

Article 8

Every individual shall have the right, according to international and national law, to just reparation for any damage sustained as a direct and determining result of an intervention affecting his or her genome.

Article 9

In order to protect human rights and fundamental freedoms, limitations to the principles of consent and confidentiality may only be prescribed by law, for compelling reasons within the bounds of public international law and the international law of human rights.

C Research on the human genome

Article 10

No research or research applications concerning the human genome, in particular in the fields of biology, genetics and medicine, should prevail over respect for the human rights, fundamental freedoms and human dignity of individuals or, where applicable, of groups of people.

Article 11

Practices which are contrary to human dignity, such as reproductive cloning of human beings, shall not be permitted. States and competent international organizations are invited to co-operate in identifying such practices and in taking, at national or international level, the measures necessary to ensure that the principles set out in this Declaration are respected.

Article 12

(a) Benefits from advances in biology, genetics and medicine, concerning the human genome, shall be made available to all, with due regard for the dignity and human rights of each individual.

(b) Freedom of research, which is necessary for the progress of knowledge, is part of freedom of thought. The applications of research, including applications in biology, genetics and medicine, concerning the human genome, shall seek to offer relief from suffering and improve the health of individuals and humankind as a whole.

D Conditions for the exercise of scientific activity

Article 13

The responsibilities inherent in the activities of researchers, including meticulousness, caution, intellectual honesty and integrity in carrying out their research as well as in the presentation and utilization of their findings, should be the subject of particular attention in the framework of research on the human genome, because of its ethical and social implications. Public and private science policy-makers also have particular responsibilities in this respect.

Article 14

States should take appropriate measures to foster the intellectual and material conditions favourable to freedom in the conduct of research on the human genome and to consider the ethical, legal, social and economic implications of such research, on the basis of the principles set out in this Declaration.

Article 15

States should take appropriate steps to provide the framework for the free exercise of research on the human genome with due regard for the principles set out in this Declaration, in order to safeguard respect for human rights, fundamental freedoms and human dignity and to protect public health. They should seek to ensure that research results are not used for non-peaceful purposes.

Article 16

States should recognize the value of promoting, at various levels, as appropriate, the establishment of independent, multidisciplinary and pluralist ethics committees to assess the ethical, legal and social issues raised by research on the human genome and its application.

E Solidarity and international co-operation

Article 17

States should respect and promote the practice of solidarity towards individuals, families and population groups who are particularly vulnerable to or affected by disease or disability of a genetic character. They should foster, inter alia, research on the identification, prevention and treatment of genetically-based and genetically-influenced diseases, in particular rare as well as endemic diseases which affect large numbers of the world's population.

Article 18

States should make every effort, with due and appropriate regard for the principles set out in this Declaration, to continue fostering the international dissemination of scientific knowledge concerning the human genome, human diversity and genetic research and, in that regard, to foster scientific and cultural co-operation, particularly between industrialized and developing countries.

Article 19

(a) In the framework of international co-operation with developing countries, States should seek to encourage measures enabling:

(i) assessment of the risks and benefits pertaining to research on the human genome to be carried out and abuse to be prevented;

(ii) the capacity of developing countries to carry out research on human biology and genetics, taking into consideration their specific problems, to be developed and strengthened;

(iii) developing countries to benefit from the achievements of scientific and technological research so that their use in favour of economic and social progress can be to the benefit of all;

(iv) the free exchange of scientific knowledge and information in the areas of biology, genetics and medicine to be promoted.

(b) Relevant international organizations should support and promote the initiatives taken by States for the above-mentioned purposes.

F Promotion of the principles set out in the Declaration

Article 20

States should take appropriate measures to promote the principles set out in the Declaration, through education and relevant means, inter alia through the conduct of research and training in interdisciplinary fields and through the promotion of education in bioethics, at all levels, in particular for those responsible for science policies.

Article 21

States should take appropriate measures to encourage other forms of research, training and information dissemination conducive to raising the awareness of society and all of its members of their responsibilities regarding the fundamental issues relating to the defence of human dignity which may be raised by research in biology, in genetics and in medicine, and its applications. They should also undertake to facilitate on this subject an open international discussion, ensuring the free expression of various socio-cultural, religious and philosophical opinions.

G Implementation of the Declaration

Article 22

States should make every effort to promote the principles set out in this Declaration and should, by means of all appropriate measures, promote their implementation.

Article 23

States should take appropriate measures to promote, through education, training and information dissemination, respect for the above-mentioned principles and to foster their recognition and effective application. States should also encourage exchanges and networks among independent ethics committees, as they are established, to foster full collaboration.

Article 24

The International Bioethics Committee of UNESCO should contribute to the dissemination of the principles set out in this Declaration and to the further examination of issues raised by their applications and by the evolution of the technologies in question. It should organize appropriate consultations with parties concerned, such as vulnerable groups. It should make recommendations, in accordance with UNESCO's statutory procedures, addressed to the General Conference and give advice concerning the follow-up of this Declaration, in particular regarding the identification of practices that could be contrary to human dignity, such as germ-line interventions.

Article 25

Nothing in this Declaration may be interpreted as implying for any State, group or person any claim to engage in any activity or to perform any act contrary to human rights and fundamental freedoms, including the principles set out in this Declaration.

***

Implementation of the Universal Declaration on the Human Genome and Human Rights

The General Conference,

Considering the Universal Declaration on the Human Genome and Human Rights, which was adopted on this eleventh day of November 1997,

Noting that the considerations formulated by the Member States at the time of the adoption of the Universal Declaration are relevant for the follow-up of the Declaration,

- Urges Member States:
  - in the light of the provisions of the Universal Declaration on the Human Genome and Human Rights, to take appropriate steps, including where necessary the introduction of legislation or regulations, to promote the principles set forth in the Declaration, and to promote their implementation;
  - to keep the Director-General regularly informed of all measures they have taken to implement the principles set forth in the Declaration;
- Invites the Director-General:
  - to convene as soon as possible after the 29th session of the General Conference an ad hoc working group with balanced geographical representation, comprised of representatives of Member States, with a view to advising him on the constitution and the tasks of the International Bioethics Committee with respect to the Universal Declaration and on the conditions, including the breadth of consultations, under which it will ensure the follow-up to the said Declaration, and to report on this to the Executive Board at its 154th session;
  - to take the necessary steps to enable the International Bioethics Committee to ensure the dissemination and follow-up of the Declaration, and promotion of the principles set forth therein;
  - to prepare for the General Conference a global report on the situation world-wide in the fields relevant to the Declaration, on the basis of information supplied by the Member States and of other demonstrably trustworthy information gathered by whatever methods he may deem appropriate;
  - to take due account, in the preparation of his global report, of the work of the organizations and agencies of the United Nations system, of other intergovernmental organizations, and of the competent international non-governmental organizations;
  - to submit his global report to the General Conference, together with whatever general observations and recommendations may be deemed necessary in order to promote the implementation of the Declaration

List of publications

Articles

Jaillon *et al.* (including S. Castellano and R. Guigó)
The initial analysis of the compact *Tetraodon nigroviridis* genome provides insight into vertebrate evolution
Submitted (2004)

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

G.V. Kryukov, S. Castellano, S.V. Novoselov, A.V. Lobanov, O. Zehtab, R. Guigó and V.N. Gladyshev
Characterization of Mammalian Selenoproteomes
Science, **300**, 1439-1443 (2003)

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

Book chapters

J.F. Abril, S. Castellano and R. Guigó
Comparative gene prediction in
Comparative Gene Prediction in Comparative Genomics:
A Guide to the Analysis of Eukaryotic Genomes
M.D. Adams (Ed.). Humana Press. *In press*.

Posters

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

S. Castellano, G.V. Kryukov, S.V. Novoselov, A.V. Lobanov, V.N. Gladyshev and R. Guigó
The Eukaryotic Selenoproteome

Translational Control, CSHL, New York (USA) (2002)

J.F. Abril, M. Albà, E. Blanco, M. Burset, F. Câmara, S. Castellano, R. Castelo,
O. Gonzalez, G. Parra and R. Guigó
Understanding the Eukaryotic Genome Sequence
Inaugural Symposium of the Center for Genomic Regulation, Barcelona, (Spain) (2002)

E. Blanco, G. Parra, S. Castellano, J.F. Abril, M. Burset, X. Fustero, X. Messeguer and R. Guigó
Gene Prediction in the Post-Genomic Era
ISMB, Copenhagen (Denmark) (2001)

S. Castellano, M. Morozova, M. Morey, M.J. Berry, F. Serras, M. Corominas and R. Guigó
Genome-wide Search for Selenoproteins in Eukaryotes
Genome Sequencing & Biology, CSHL, NY (USA) (2001)

J.F. Abril, E. Blanco, M. Burset, S. Castellano, X. Fustero, G. Parra and R. Guigó
Genome Informatics Research Laboratory: Main Research Topics
First Spanish Bioinformatic's meeting, Cartagena (Spain) (2000)

Contact information

Find below, in alphabetical order, the contact information for some of the authors of the research presented here:

Sergi Castellano
Ph.D.
Research Group in Genome Bioinformatics
Institut Municipal d'Investigació Mèdica
Dr. Aiguader, 80, 08003 Barcelona (Spain)
Phone: +34 93 2240891; Fax: +34 93 2240875
E-mail: scaste@imim.es
Web: http://genome.imim.es/˜scaste/

Montserrat Coromines
Professor
Departament de Genètica
Universitat de Barcelona
Diagonal, 645, 08028 Barcelona (Spain)
Phone: +34 93 403 70 03; Fax: +34 93 411 09 69
E-mail: mcorominas@ub.edu

Vadim N. Gladyshev
Associate Professor
Department of Biochemistry
N151 Beadle Center, University of Nebraska
Lincoln, NE 68588-0664 (USA)
Phone: (402) 472-4948; Fax (402) 472-7842
E-mail: vgladyshev1@unl.edu
Web: http://www.unl.edu/biochem/gladyshev

Roderic Guigó
Investigator
Research Group in Genome Bioinformatics
Institut Municipal d'Investigació Mèdica
Dr. Aiguader, 80, 08003 Barcelona (Spain)
Phone: +34 93 2240877; Fax: +34 93 2240875
E-mail: rguigo@imim.es
Web: http://genome.imim.es/˜rguigo/

Alain Krol
Directeur de Recherche CNRS
UPR 9002 du CNRS
Institut de Biologie Moleculaire et Cellulaire
15, Rue Rene Descartes, 67084 Strasbourg Cedex (France)

Phone: +33 (0)3 88 41 70 50; Fax: +33 (0)3 88 60 22 18
E-mail: A.Krol@ibmc.u-strasbg.fr
http://www-ibmc.u-strasbg.fr/upr9002/krol

Hugues Roest
Investigator
Genoscope
CNRS UMR 8030
2, rue Gaston Crémieux, 91057 Evry Cedex (France)
E-mail: hrc@genoscope.cns.fr

Florenci Serras
Professor
Departament de Genètica
Facultat de Biologia
Diagonal, 645, 08028 Barcelona (Spain)
Phone: +34 93 403 70 03; Fax: +34 93 411 09 69
E-mail: fserras@ub.edu

Curriculum Vitae

Sergi Castellano, born in 1975 in Esparreguera, Barcelona (Spain), graduated on 1999 in Biology (Bachelor's degree) by the Universitat de Barcelona (UB). He spent his last semester as undergraduate doing research at the School of Biological Sciences (University of Manchester) under the direction of Dr. Paul Higgs.

In late 1999, he worked several months as UNIX system administrator at the Informàtica de Recerca i Docència at the UB, taking care of the central UNIX servers of the Mathematics, Geology, History and Philosophy faculties.

From 2000 till mid 2004 he stayed as a *PhD* student, under supervision of Dr. Roderic Guigó within the Grup de Recerca en Informàtiva Biomèdica at the meta-institution formed by the soon merging Institut Municipal d'Investigació Mèdica (IMIM), Universitat Pompeu Fabra (UPF) and Centre de Regulació Genòmica (CRG). This new research center is to be called Parc de Recerca Biomèdica de Barcelona (PRBB).

After the completion of his *PhD*, he is soon moving for a short research stay to the group of Dr. Marla Berry at the University of Hawai'i at Mânoa. In early 2005, he is appointed for incorporation to a postdoc position in the lab of Dr. Sean Eddy at the Washington University (WashU).

Miscellanea

This thesis layout largely derives from the $\LaTeX\,2_{\varepsilon}$ template created by Robert Castelo (2002). However, it has been extensively modified and, maybe, improved. Here, I provide some comments on it and the source code for download.

Technical comments

This book was typeset with GNU `emacs` 21.3.1 in LaTeX mode and converted to PDF with `pdfTeX` 3.14159-1.10b. All running on a linux box with Red Hat 7.3 (Valhalla) Kernel 2.4.20-13bigmem.

This $\LaTeX\,2_{\varepsilon}$ thesis template

$\LaTeX\,2_{\varepsilon}$ is a document preparation system, powerful, robust and able to achieve professional results. However, the learning curve may be stiff, thus, an initial template is given here for your convenience.

Style file

R. Castelo wrote his own thesis style file (`mythesis.sty`) to handle fonts and control section title layout. It has been slightly modified to include expanded title font faces.

Makefile

This file (`Makefile`), also derived from R. Castelo's equivalent, is read to produce a PDF version of your thesis just by typing `make pdf` in the command line. It is aware of any change you make in any of the child directories and LaTeX files that compose your document. It also reruns itself to update the references section. It needs the `make` program in your system, though it is usually installed by default.

Preamble

This initial set of instructions to control the overall document processing is stored in its own file (`preamble.tex`), separately of the thesis text for the sake of clarity. In this, the style file is called.

Document

This body section of the document (`mythesis.tex`) is simply a call of the independent chapter files (`*.tex`) your thesis consist of (each of them nicely placed in its own directory). It also calls, at the very beginning, the preamble file.

Download

This template can be found at: `genome.imim.es/˜scaste/mythesis/`

Bibliography

J. Aaseth, M. Haugen, and O. Forre. Rheumatoid arthritis and metal compounds-perspectives on the role of oxygen radical detoxification. *Analyst*, 123:3–6, 1998.

M. Alexandersson, S. Cawley, and L. Pachter. SLAM: cross-species gene finding and alignment with a generalized pair hidden Markov model. *Genome Research*, 13:496–502, 2003.

S. F. Altschul, T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25: 3389–3402, 1997.

B. Alzina. *Cracterització i anàlisi funcional del gen patufet de Drosophila melanogaster*. PhD thesis, Departament de Genètica, Universitat de Barcelona, 1999.

M. J. Axley, A. Böck, and T. C. Stadman. Catalytic properties of an *Escherichia coli* formate dehydrogenase mutant in which sulfur replaces selenium. *Proceedings of the National Academy of Sciences USA*, 88:8450–8454, 1991.

C. Baron, E. Westhof, A. Böck, and R. Giege. Solution structure of selenocysteine -inserting tRNA^{Sec} from *Escherichia coli*. Comparison with canonical tRNA. *J. Mol. Biol.*, 231:274–292, 1993.

M. K. Baum, A. Campa, M. J. Miguez-Burbano, X. Burbano, and G. Shor-Posner. *Selenium: its molecular biology and role in human health*, chapter Role of selenium in HIV/AIDS, pages 247–255. Kluwer Academic Publishers, 2001.

M. A. Beck. *Selenium: its molecular biology and role in human health*, chapter Selenium as an antiviral agent, pages 235–245. Kluwer Academic Publishers, 2001.

D. Behne, A. Kyriakopoeulos, C. Weiss-Nowak, M. Kalcklosch adn C. Westphal, and H. Gessner. Newly found selenium-containing proteins in the tissues of the rat. *Biol. Trace. Elem. Res.*, 55:99–110, 1996.

M. J. Berry, G. W. Martin III, and S. C. Low. RNA and protein requirements for eukaryotic selenoprotein synthesis. *Biomed. Environ. Sci*, 10:182–189, 1997.

M. J. Berry, A. L. Mai, J.D. Kieffer, J. W. Harney, and P.R Larsen. Substitution of cysteine for selenocysteine in type i iodothyronine deiodinase reduces the catalytic efficiency of the protein but enhances its translation. *Endocrinology*, 131:1848–1852, 1992.

Marla J. Berry, Laila Banu, John W. Harney, and P.Reed Larsen. Functional characteritzation of the eukaryotic SECIS elements which direct selenocysteine insertion at UGA codons. *EMBO Journal*, 12: 3315–3322, 1993.

Marla J. Berry, Laila Banu, and P. Reed Larsen. Type I iodothyronine deiodinase is a selenocysteine-containing enzyme. *Nature*, 349:438–440, 1991.

E. Birney and R. Durbin. Using GeneWise in the *Drosophila* annotation experiment. *Genome Research*, 10: 547–548, 2000.

Christoph Buettner, John W. Harney, and Marla J. Berry. The *caenorhabditis elegans* Homologue of Thioredoxin Reductase Contains a Selenocysteine Insertion Sequence (SECIS) Element that Differs from Mammalian SECIS Elements but Directs Selenocysteine Incorporation. *Journal of Biological Chemistry*, 274:21598–21602, 1999.

C. B. Burge and S. Karlin. Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, 268:78–94, 1997.

M. Burset and R. Guigó. Evaluation of gene structure prediction programs. *Genomics*, 34:353–357, 1996.

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

R. Castelo. *The Discrete Acyclic Digraph Markov Model in Data Mining*. PhD thesis, Faculteit Wiskunde en Informatica, Universiteit Utrecht, April 2002.

S. Cawley, L. Pachter, and M. Alexandersson. SLAM web server for comparative gene finding and alignment. *Nucleic Acids Research*, 31:3507–3509, 2003.

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

G. F. Combs and L. Lu. *Selenium: its molecular biology and role in human health*, chapter Selenium as a cancer preventive agent, pages 205–217. Kluwer Academic Publishers, 2001.

J.E Cone, M. del Rio, J.N Davis, and T.C. Stadman. Chemical characterization of the selenoprotein component of clostridial glycine reductase: Identification of selenocysteine as the organoselenium moiety. *Proc. Natl. Acad. Sci.*, 73:2659–2663, 1976.

P.R. Copeland, J.E. Fletcher, B.A. Carlson, D.L. Hatfield, and D.M. Driscoll. A novel rna binding protein, sbp2, is required for the translation of mammalian selenoprotein mrnas. *EMBO Journal*, 19:306–314, 2000.

R. J. Coppinger and A. M. Diamond. *Selenium: its molecular biology and role in human health*, chapter Selenium deficiency and human disease, pages 219–233. Kluwer Academic Publishers, 2001.

T. Dandekar, editor. *RNA Motifs and Regulatory Elements*. Springer, 2002.

Thomas Dandekar and Mathias W. Hentze. Finding the hairpin in the haystack: searching for RNA motifs. *Trends in Genetics*, 11:45–50, 1995.

D. M. Driscoll and L. Chavatte. Finding needles in a haystack: *in silico* identification of eukaryotic selenoprotein genes. *EMBO reports*, 5:140–141, 2004.

D. Fagegaltier, N. Hubert, K. Yamada, T. Mizutani, P. Carbon, and A. Krol. Characterization of mselb, a novel mammalian elongation factor for selenoprotein translation. *EMBO Journal*, 19:4796–4805, 2000a.

D. Fagegaltier, A. Lescure, R. Walczak, P. Carbon, and A. Krol. Structural analysis of new local features in SECIS RNA hairpins. *Nucleic Acids Research*, 28:2679–2689, 2000b.

L. R: Flohé, Brigelius-Flohé, M. Maiorino, A. Roveri, J. Wissing, and F: Ursini. *Selenium: its molecular biology and role in human health*, chapter Selenium and male reproduction, pages 273–281. Kluwer Academic Publishers, 2001.

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

R. Guigó, E. T. Dermitzakis, P. Agarwal, C. P. Ponting, G. Parra, A. Reymond, J. F. Abril, E. Keibler, R. Lyle, C. Ucla, S. E. Antonarakis, and M. R. Brent. Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes. *Proc. Natl. Acad. Sci. USA*, 4:1140–1145, 2003.

D. L. Hatfield, V. N. Gladyshev, J. M. Park, S. I. Park, H. S. Chittum, J. R. Huh, B. A. Carlson, M. Kim, M. E. Moustafa, and B. J. Lee. *Comprehensive natural products chemistry*, volume 4, chapter Biosynthesis of selenocysteine and its incorporation into protein as the 21st aminoo acid, pages 353–380. Elsevier Science, Ltd., 1999.

J. Hathcock. Vitamins and minerals: Efficacy and safety. *Am. J. Clin. Nutr.*, 66:427–437, 1997.

M. Hirosawa-Takamori, H. R. Chung, and H. Jackle. Conserved selenoprotein synthesis is not critical for oxidative stress defence and the lifespan of *Drosophila*. *EMBO Reports*, 5:317–322, 2004.

N. Hubert, R. Walczak, C. Sturchler, E. Myslinski, C. Schuster, E. Westhof, P. Carbon, and A. Krol. RNAs mediating cotranslational insertion of selenocysteine in eukaryotic selenoproteins. *Biochimie*, 78:590–596, 1996.

Glover W. Martin III, John W. Harney, and Marla Berry. Functionality of mutations at conserved nucleotides in eukaryotic SECIS elements is determined by the identity of a single nonconserved nucleotide. *RNA*, 4:65–73, 1998.

W.J. Kent. BLAT –the BLAST-like alignment tool. *Genome Research*, 12:656–664, 2002.

J. Kohrle. The deiodinase family: selenoenzymes regulating thyroid hormone availability and action. *Cell. Mool. Life Sci.*, 57:1853–1863, 2000.

L. D. Koller and J. H. Exon. The two faces of selenium-deficiency and toxicity-are similar in animals and man. *Can. J. Vet. Res.*, 50:297–306, 1986.

Heike Kollmus, Leopold Flohé, and John E.G. McCarthy. Analysis of eukaryotic mRNA structures directing cotranslational incorporation of selenocysteine. *Nucleic Acids Research*, 24:1195–1201, 1996.

I. Korf, P. Flicek, D. Duan, and M. R. Brent. Integrating genomic homology into gene structure prediction. *Bioinformatics.*, 17 Suppl 1:S140–148, 2001.

K.V Korotkov, S.V. Novoselov, D.L. Hatfield, and V.N. Gladyshev. Mammalian Selenoprotein in Which Selenocysteine (Sec) Incorporation is Supported by a New Form of Sec Insertion Sequence Element. *Molecular and Cellular Biology*, 22:1402–1411, 2002.

K. Kose, P. Dogan, Y. Kardas, and R. Saraymen. Plasma selenium levels in rheumatoid arthritis. *Biol. Trace. Elem. Res.*, 53:51–56, 1996.

Gregory V. Kryukov, Valentin M. Kryukov, and Vadim N. Gladyshev. New Mammalian Selenocysteine-containing Proteins Identified with an Algorithm that Searches for Selenocysteine Insertion Sequence Elements. *Journal of Biological Chemistry*, 274(48):33888–33897, Nov 1999.

D. Kulp, D. Haussler, M. G. Reese, and F. H. Eeckman. A generalized hidden markov model for the recognition of human genes in DNA. In D. J. States, P. Agarwal, T. Gaasterland, L. Hunter, and R. Smith, editors, *Intelligent Systems for Molecular Biology*, pages 134–142, Menlo Park, California, 1996. AAAI press.

O. A. Levander and M. A. Beck. Interacting nutritional and infectious etiologies of Keshan disease. Insights from coxsackie virus B-induced myocarditis in mice deficient in selenium or vitamin E. *Biol. Trace. Elem. Res.*, 56:5–21, 1997.

Susan C. Low and Marla J. Berry. Knowing when not to stop: selenocysteine incorporation in eukaryotes. *Trends in Biochemical Sciences*, 21:203–208, 1996.

R. C. McKenzie, T. S. Rafferty, G. J. Beckett, and J. R. Arthur. *Selenium: its molecular biology and role in human health*, chapter Effects of selenium on immunity and aging, pages 257–272. Kluwer Academic Publishers, 2001.

M. Morey. *Selenoproteïnes i estrès oxidatiu a Drosophila: regulació negativa de la via Ras/MAPK i activació de l'apopotosi*. PhD thesis, Departament de Genètica, Universitat de Barcelona, May 2003.

N. Morozova, E. P. Forry, E. Shahid, A. M. Zavacki, J. W. Harney, Y. Kraytsberg, and M. J. Berry. Antioxidant function of a novel selenoprotein in *Drosophila melanogaster*. *Genes Cells*, 8:963–971, 2003.

R. Mott. EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *Computer Applications in the Biosciences*, 13:477–478, 1997.

S. Müller, J. Heider, and A. Böck. The path of unspecific incorporation of selenium in *Escherichia coli*. *Arch. Microbiol.*, 168:421–427, 1997.

S. Osawa. *Evolution of the Genetic Code*. Oxford University Press, Oxford, 1995.

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

G. Parra, E. Blanco, and R. Guigó. Geneid in *Drosophila*. *Genome Research*, 10:511–515, 2000.

M. F. Raisbeck, E. R. Dahl, D. A. Sanchez, E. L. Belden, and D. O'Toole. Naturally occurring selenosis in Wyoming. *J Vet. Diagn. Invest.*, 5:84–87, 1993.

A. A. Salamov and V. V. Solovyev. *Ab initio* gene finding in *Drosophila* genomic DNA. *Genome Research*, 10:516–522, 2000.

Qichang Shen, Jack L. Leonard, and Peter E. Newburger. Structure and function of the selenium translation element in the 3′-unstranslated region of human cellular glutathione peroxidase mrna. *RNA*, 1: 519–525, 1995.

Thressa C. Stadman. Selenocysteine. *Annual Review of Biochemistry*, 65:83–100, 1996.

C. Sturchler, E. Westhof, P. Carbon, and A. Krol. Unique secondary and tertiary structural features of the eukaryotic selenocysteine tRNA^{Sec}.

R.M. Tujebajeva, P.R. Copeland, X. Xu, B.A. Carlson, J.W. Harney, D.M. Driscoll, D.L. Hatfield, and M.J. Berry. Decoding apparatus for eukaryotic selenocysteine insertion. *EMBO reports*, 1:158–163, 2000.

Robert Walczak, Philippe Carbon, and Alain Krol. An essential non-Watson-Crick base pair motif in 3'UTR to mediate selenoprotein translation. *RNA*, 4:74–84, 1998.

Robert Walczak, Eric Westhof, Philippe Carbon, and Alain Krol. A novel RNA structural motif in the selenocysteine insertion element of eukaryotic selenoprotein mRNAs. *RNA*, 2:367–379, 1996.

S. J. Wheelan, D. M. Church, and J. M. Ostell. Spidey: a tool for mRNA-to-genomic alignments. *Genome Research*, 11:1952–1957, 2001.

T. Wiehe, S. Gebauer-Jung, T. Mitchell-Olds, and R. Guigó. SGP-1: Prediction and validation of homologous genes based on sequence alignments. *Genome Research*, 11:1574–1583, 2001.

J. Q. Wu, D. Shteynberg, M. Arumugam, R. A. Gibbs, and M. R. Brent. Identification of rat genes by TWINSCAN gene prediction, RT-PCR, and direct sequencing. *Genome Research*, 14:665–671, 2004.

R. F. Yeh, L. P. Lim, and C. B. Burge. Computational inference of homologous gene structures in the human genome. *Genome Research*, 11:803–816, 2001.

F. Zinoni, A. Birkmann, T. C. Stadman, and A. Böck. Nucleotide-sequence and expression of the selenocysteine-containing polypeptide of formate-dehydrogenase (formate-hydrogen-lyase-linked) from *Escherichia coli*. *Proc. Natl. Acad. Sci.*, 83:4560–4564, 1986.

Index