The Pennsylvania State University

The Graduate School

Department of Biology

CONTRIBUTION OF TRANSPOSABLE ELEMENTS TO GENOMIC NOVELTY: A COMPUTATIONAL APPROACH

A Thesis in

Biology

by

Valer Gotea

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2007

The thesis of Valer Gotea was reviewed and approved* by the following:

Wojciech Makałowski
Associate Professor of Biology
Thesis Advisor
Chair of Committee

Stephen W. Schaeffer
Associate Professor of Biology

Kateryna D. Makova
Assistant Professor of Biology

Piotr Berman
Associate Professor of Computer Science and Engineering

Douglas R. Cavener
Professor of Biology
Head of the Department of Biology

*Signatures are on file in the Graduate School

ABSTRACT

Transposable elements (TEs) are DNA entities that have the ability to move and multiply within genomes, and thus have the ability to influence their function and evolution. Their impact on the genomes of different species varies greatly, yet they made an important contribution to eukaryotic genomes, including to those of vertebrate and mammalian species. Almost half of the human genome itself originated from various TEs, few of them still being active. Often times, TEs can disrupt the function of certain genes and generate disease phenotypes, but over long evolutionary times they can also offer evolutionary advantages to their host genome. For example, they can serve as recombination hotspots, they can influence gene regulation, or they can even contribute to the sequence of protein coding genes. Here I made use of multiple computational tools to investigate in more detail a few of these aspects. Starting with a set of well characterized proteins to complement inferences made at the level of transcripts, I investigated the contribution of TEs to protein coding sequences. I show that old TEs can indeed be found in functional proteins, albeit to a lesser extent than previously thought. I also investigated the protein coding potential of Alu elements, which are found in the alternatively spliced forms of many genes. No strong evidence could be found to support this hypothesis, but novel ways in which Alu elements can contribute to the human proteome are proposed. Complementing their protein coding potential, TEs were also shown to have a big potential to influence *cis* regulation of genes due to the composition of their sequence. This appears to be one important factor that determines their fate after insertion at new loci. A database, called the ScrapYard database (SYDB), was ultimately

created to provide an efficient tool in addressing various questions related to the presence of TE sequences in vertebrate transcripts. At the present time, when genomic and genetic data are generated at increased rates, important advancements in biological knowledge can only be made with the help of computational tools. This work provides an example of how computational biology can advance biological knowledge, by exploring current hypotheses and testing them with available data, and at the same time, proposing new ones that can be further tested experimentally.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

PREFACE

Generated by my interest in how transposable elements can influence the evolution of genomes, my research has expanded into several directions whose exploration would not have been possible without the direct help of two colleagues. Bartley G. Thornburg made crucial contributions to Chapter 4. After I extracted the promoter sequences to be investigated, he scanned them with RepeatMasker and MATCH for the presence of TE sequences and transcription factor binding sites, respectively. He also conducted the statistical analysis, tables 4.1, 4.2, and 4.3 being his direct work. I completed the work by summarizing the results in figures 4.1, 4.2, as well as by interpreting the data and by drawing conclusions in the context of current scientific literature. Vamsi Veeramachaneni made possible the existence of the ScrapYard database (SYDB) presented in Chapter 5. While I collected and prepared all the data for inclusion in a MySQL database, he designed and implemented the web interface of SYDB which allows public access to the data. I thank them both for their dedication and productivity with these projects.

ACKNOWLEDGEMENTS

Few personal endeavors require the knowledge, dedication, passion, support, and understanding of so many people than obtaining a Ph.D. degree does. Therefore, I will remain indebted to the following for their individual contributions to my success. To my advisor, Dr. Wojciech Makałowski, for inspiring and guiding me through the fascinating world of "junk" DNA. To Penn State's Worldwide Universities program, which supported my collaboration with Dr. Jordi Bella from University of Manchester, who helped me with the problem of protein homology modeling. To the members of my doctoral committee, Dr. Stephen Schaeffer, Dr. Kateryna Makova, and Dr. Piotr Berman, for their understanding and guidance. To my wife, Claudia Gotea, for her continuous support and inspiration in all the good and not so good moments, without whom this endeavor would have been much more difficult. And last, but not least, to my parents, Maria and Mircea Gotea, for their unconditional support and for guiding me towards becoming a scientist.

Chapter 1

An Overview of Transposable Elements in Eukaryotic Genomes

"Transposable elements" (TEs) is a generic term that refers to DNA entities that have the ability to move or multiply within genomes generating self-copies interspersed with non-repetitive DNA. The term is often used with reference to copies of such elements that have lost the ability to move or multiply once inserted at a new genomic locus, thus "TE-derived sequences" or "transposed elements" would better describe the status of the sequence. A more general term, "interspersed repeats", can be used to refer to both active and non-active TEs.

1.1 Discovery of transposable elements

Originally labeled as "controlling units", and later "controlling elements" (McClintock 1956), TEs were discovered in a series of experiments on maize conducted in 1944 by Barbara McClintock. The discovery was followed by six years of additional experiments, and was finally published in 1951 (McClintock 1951). At the time, the concept of gene itself was under debate, and thus the new discovery was hard to assimilate in that scientific context. Writing about the reception of McClintock's talk at the 1951 Cold Spring Harbor Symposium, Evelyn-Fox Keller notes that the talk "was met with stony silence. With one or two exceptions, no one understood" (Keller 1983). Part of the confusion was due to McClintock's apparent siding with Richard Goldschmidt, who

opposed the generally accepted view of the gene as "a unitary entity that acted independently to produce a physiologically active molecule" (Comfort 1995). Based on his observations on *Bar eye* in Drosophila, Goldschmidt concluded that the unit of heredity was the chromosome, not the gene (Goldschmidt 1938, Dunn 1965, Comfort 1995). However, McClintock notes in her paper that the lack of use of the "*gene*" term in her discussion "does not imply a denial of the existence within chromosomes of units or elements having specific functions" (McClintock 1951). With previous evidence that the DNA was the carrier of genetic information (Avery*, et al.* 1944), it took a few more years for the structure of the DNA to be solved (Watson and Crick 1953) and the genetic code to be cracked (Nirenberg and Matthaei 1961, Nirenberg*, et al.* 1966), which paved the way for clarification of the "gene" concept.

The importance of the discovery of transposable elements, which appeared as repetitive elements in the maize genome, was acknowledged and reinforced after genomes of other organisms as diverse as fruit fly, frog, sea urchin, and human were shown to be "peppered" with repetitive sequences. For example, both the sea urchin and frog genomes were found to contain about 30% repetitive sequences interspersed with single copy sequences (Davidson*, et al.* 1973, Graham*, et al.* 1974), findings confirmed a quarter of a century later after the analysis of the genome sequences(Cameron*, et al.* 2000). In these genomes, short repetitive sequences (~0.4 kb) appeared interspersed with longer (~0.9 kb) single copy sequences. The fruit fly genome, however, revealed a different pattern of interspersion, with longer (~5 kb) repetitive sequences interspersed with even longer (~40 kb) unique DNA stretches (Manning*, et al.* 1975). By renaturation techniques, hydroxylapatite binding methods, and DNA hyperchromism, a pattern of

interspersion similar to that in frog and sea urchin genomes was found in the human genome in 1975 (Schmid and Deininger 1975). The same study revealed that repetitive sequences could be found in at least 80% of the human genome. Advances in DNA sequencing techniques have allowed for accumulation of DNA sequence data, and better estimates of genomes content in repetitive sequences. Using all human genomic GenBank entries, Smit estimated that a fraction as big as 35.5% of the human genome is made of interspersed repeats (Smit 1996). A few years later, the milestone initial sequencing of the human genome (Lander*, et al.* 2001, Venter*, et al.* 2001) revealed that more than half of the human genome is made of repetitive DNA.

Availability of entire genome sequences allows not only for a better estimate of genomic repeat content, but also for detailed analysis of the repeats themselves. Using genomic sequence, it was possible to determine structural characteristics of repetitive sequences, such as the length of complete transposable elements, the presence of open reading frames, the presence of terminal repeats and polyadenylation signals. It was also interesting to discover that some repeats, such as LTR retrotransposons, resemble very well retroviruses, but they only have the ability to reinsert into the genome from which they originate. The International Human Genome Sequencing Consortium (Lander*, et al.* 2001) defines five classes of repeat sequences: **1)** transposon-derived repeats, or interspersed repeats; **2)** processed pseudogenes, which are partially, thus inactive, retroposed copies of cellular genes; **3)** simple sequence repeats, consisting of direct repetitions of relatively short *k*-mers such as $(A)_n$, $(CA)_n$ or $(CGG)_n$; **4)** segmental duplications, defined as blocks of 10-300 kb copied from one genomic region to another; and **5)** blocks of tandemly repeated sequences, such as centromeres, telomeres, the short

arms of acrocentric chromosomes, and ribosomal gene clusters. As one would expect, estimates of number of members in each class varied with availability of sequence data. For example, the number of primate and rodent TE (sub)families deposited in Repbase increased from 284 and 157, respectively, in 1998 (Jurka 1998), to 443 and 252, respectively, two years later (Jurka 2000), and to 548 and 345, respectively, in the current Repbase release (12.06). It needs to be mentioned that the increase in number of (sub)family members is not only due to availability of sequence data, but also to advancement in sequence searching techniques, given that many repeats are very old, thus having high sequence divergence, and being hard to find with classical computational approaches (e.g. BLAST searches). Specific algorithms have been developed or adapted (Jurka*, et al.* 1996, Tulko*, et al.* 1997, Bao and Eddy 2002, McCarthy and McDonald 2003), and specialized software packages, such as RepeatMasker (http://repeatmasker.org) and Censor (Jurka*, et al.* 1996), are currently available for identifying various repetitive elements using libraries of consensus sequences. One should note that consensus sequences themselves were updated with sequences/domains previously unknown, mainly because most repeat copies were highly divergent. As an example, the consensus sequence of LINE2 was 2,977 bp long in the May 15, 2002 release of RepeatMasker libraries, and it is 3,314 bp long in the 9.07 release of Repbase.

The second class of repeated sequences mentioned above, the pseudogenes, is a special class, since they were the subject of Masatoshi Nei's study which recognized the importance of non-exonic sequences (Nei 1969). Referring to non-coding sequences a few years later, Suzumu Ohno coined the term "junk" DNA (Ohno 1972), which caught

on in the 1980s, and which extended to all non-coding sequences, obviously including all classes of repeat sequences mentioned above.

The subject of this dissertation is the first class of repetitive DNA, namely the transposable elements, transposon-derived repeats, or interspersed repeats, with focus on mammalian genomes, and especially on the human genome where TEs represent almost half of its sequence (Lander*, et al.* 2001).

1.2 Classification of transposable elements

Given the exponential growth of sequence data during the last few decades, finding and classifying TEs was a very fruitful enterprise. The steps for studying a new TE is described by Jurka (1994): identification of repeat copies, sequence alignment, classification into subfamilies (if applicable), and construction of consensus sequences. Consensus sequences are important because they represent the best available approximation of the original active TE that generated the copies (Jurka 1998). The relationship between similarities of individual repeats to perfect consensus sequence and similarities between repeats themselves is illustrated by Jurka (1994, 1998): repeats 37-52% similar to each other will be 55-70% similar to their consensus sequence. Approximate reconstruction of protein products encoded by the open reading frames contained by transposable elements (e.g. *Gag* and *Pol* proteins in LTR retrotransposons) is also possible from consensus sequences, and this usually provides a significant insight into the mechanism of transposition, and/or ways of distinguishing different classes of repeats. For example, the chicken genome contains about 30,000 copies of the chicken

repeat 1 (CR1), of which only a small fraction are full length copies. The consensus sequence deduced from these full length copies showed that the CR1 repeat encodes a reverse transcriptase-like sequence, which was used to infer that CR1 elements are related to LINE1 elements (Burch*, et al.* 1993). However, the CR1 repeats can be easily distinguished from LINE1 sequences based on the deduced sequence of the protein encoded by ORF2 (Poulter*, et al.* 1999). Another interesting example is the Maui repeat found in the genome of the Japanese pufferfish, which has an ORF2 protein that contains an N-terminal endonuclease domain, a conserved RT domain, and a C-terminal domain of unknown function. Based on this ORF2 protein it has been determined that Maui repeats are most closely related to chicken and turtle CR1 elements (Kajikawa*, et al.* 1997, Poulter*, et al.* 1999). Therefore, the proteins encoded by various TEs not only offer clues about the mechanism of their mobility, but are also an important criterion for their classification.

There are several accounts of TE classification and general overview in the literature (Finnegan 1989, McClure 1991, Smit 1996, Capy 1998, Lander*, et al.* 2001, Deininger and Batzer 2002, Deininger*, et al.* 2003, Kazazian 2004, Capy 2005). While some classifications only deal with reverse-transcription carrying elements (McClure 1991), other works provide a more detailed overview of all types of transposable elements, with insights into the mechanism of mobility, relationships between different classes, and effects on host genomes (Craig*, et al.* 2002). With the recent completion of several mammalian genome sequences, such as human, mouse, and macaque, we now also have detailed accounts of TE occurrences in those species (Lander*, et al.* 2001, Waterston*, et al.* 2002, Han*, et al.* 2007). TEs account for ~45% of the human genome,

thus being the most important class of repetitive sequences (Lander, *et al.* 2001, Makalowski 2001). One of the first classifications characterized TEs as either a) Class I elements that transpose by reverse transcription of an RNA intermediate, or b) Class II elements that "transpose directly from DNA to DNA" (Finnegan 1989). With further subdivisions, this is the classification agreed upon by most scientists.

1.2.1 Retrotransposons

Retrotransposons are mobile elements that multiply using an mRNA intermediate, which is reverse transcribed and reintegrated into the genome. Retrotransposons can be further divided into two classes: LTR retrotransposons, and non-LTR retrotransposons.

1.2.1.1 LTR Retrotransposons

LTR retrotransposons are characterized by the presence of long terminal repeats (LTR) at both ends, and similarity to retroviruses. Both exogenous retroviruses and LTR retrotransposons contain a *gag* gene, that encodes a viral particle coat, and a *pol* gene, that encodes a reverse transcriptase, a ribonuclease H, and an integrase, which provide the enzymatic machinery for reverse transcription from RNA and integration into the host genome. Reverse transcription occurs within the viral or viral-like particle (GAG) in the cytoplasm, and it is a complicated multi-step process (Voytas and Boeke 2002). Unlike LTR retrotransposons, exogenous retroviruses contain an *env* gene, which encodes an envelope that facilitates their migration to other cells. LTR retrotransposons either lack or

contain a remnant of an *env* gene (Kazazian 2004), limiting their insertion capabilities to the originating genome. This would suggest that they originated in exogenous retroviruses by losing the *env* gene. However, there is evidence that would suggest the contrary, given that LTR retrotransposons can acquire the *env* gene and become infectious entities (Malik*, et al.* 2000). In mammalian genomes only the vertebrate specific endogenous retroviruses (ERVs) appear to have been active, in spite of a larger variety of LTR elements. Currently, most of the LTR sequences (85%) are found only as isolated LTRs, with the internal sequence having been lost by homologous recombination between the flanking LTRs (Ono 1986, Leib-Mosch*, et al.* 1993, Lander*, et al.* 2001), but HERV-K elements are thought to have been active in the human genome after the spit from chimpanzee (Lower*, et al.* 1996, Costas 2001). It is interesting to note that LTR retrotransposons target their reinsertion to specific genomic sites, often around genes, with putative important functional implications for host genomes (Kazazian 2004). For example, Ty3 elements of *Saccharomyces cerevisiae* are very successful when inserted within 750 bp upstream of *Pol*III transcribed genes. Lander *et al.* (2001) estimate that 450,000 LTR copies make up about 8% of our genome.

1.2.1.2 Non-LTR Retrotransposons

Non-LTR retrotransposons are missing the long terminal repeats, and have, instead, a poly-A tail at the 3' end. Several clades or superfamilies (about 10) of non-LTR retrotransposons have been defined in eukaryotes, based on phylogenetic studies using the sequences of proteins encoded by each element (Malik*, et al.* 1999). The best

represented non-LTR retrotransposons are **LINEs** (**L**ong **IN**terspersed **E**lements), which comprise about 21% of our genome (about 850,000 copies) (Lander*, et al.* 2001). Among these, the best described element is the LINE1 (L1) non-LTR retrotransposon. A full copy of L1 is about 6 kb long, contains a *Pol*II promoter, and two ORFs. The function of the ORF1 protein is unclear, but it is known that it contains non-specific Zn-finger, leucine zipper, and coiled-coil motifs (Holmes*, et al.* 1992, Dawson*, et al.* 1997, Haas*, et al.* 1997, Kajikawa*, et al.* 1997, Poulter*, et al.* 1999). The second ORF encodes an endonuclease, which makes a single stranded nick in the genomic DNA, and a reverse transcriptase, which uses the nicked DNA to prime reverse transcription of LINE RNA from the 3' end (Lander*, et al.* 2001). Reverse transcription is often unfinished, leaving behind fragmented copies of LINE elements, most of the L1-derived repeats being short, with an average size of 900 bp. LINE1 is the only LINE element still active in the human genome (Lander*, et al.* 2001). In the human genome there are two other LINE-like repeats, L2 and L3, that are distantly related to L1. They are part of the CR1 clade (Kapitonov and Jurka 2003), which has members in various organisms, such as fruit fly, mosquito, zebrafish, pufferfish, turtle, and chicken. L2 and L3 LINEs are very old repeats, thus with highly divergent copies, and consensus sequences that have not been completed. Because they encode their own retrotransposition machinery, LINE elements are regarded as autonomous retrotransposons.

SINEs, or **S**hort **IN**terspersed **E**lements, are another important class of non-LTR retrotransposons, evolved from RNA genes, such as 7SL, and tRNA genes. By definition, they are short elements, up to 1,000 bp long, and are non-autonomous elements, not encoding their own retrotranscription machinery. The outstanding member of this class

from the human genome is the Alu repeat, with more than one million copies in the human genome. It contains a cleavage site for the *Alu*I restriction enzyme, which gave it its name (Houck*, et al.* 1979). The consensus Alu sequence is about 300 bp long, and is formed by two unequal monomers (each about 18% different, the right monomer being longer) connected by an A-reach segment, a possible remnant of ancestral monomeric state. A 3' poly-A tail is characteristic for SINEs. Alus, as well as all other SINEs, lack coding capacity. A 2 bp mutation caused the formation of a *Pol*III promoter in the left monomer (Quentin 1994), but Alus can be transcribed by *Pol*II as well, when interspersed with genes (Schmid 1998, Li and Schmid 2004). They are thought to be retrotransposed by the L1 machinery (Dewannieux*, et al.* 2003, Schmid 2003), even tough they do not share the 3' end with LINE1. Alu elements are specific to primates, but they have a rodent relative, the B1 element, which is a monomeric repeat, also derived from the 7SL RNA gene (Kriegs*, et al.* 2007). Another type of SINE retrotransposons is the Mammalian wide Interspersed Repeat (MIR) element, which similarly to Alu elements, is non-autonomous. However, unlike Alu elements, MIRs share their 3' end 50 nt with L2 repeats which are though to have facilitated their mobility (Smit 1996). MIRs spread before eutherian radiation, copies being retrieved from marsupials and monotremes (Jurka*, et al.* 1995). SVA elements are a recent addition to the SINE group of non-autonomous retrotransposons, with their name indicating the fact that their sequence is a composite made of **S**INE-R (Ono*, et al.* 1987), **V**NTR (variable number of tandem repeats), and **A**lu-like sequences as originally described by Shen *et al.* (1994). SVA elements are thought to be currently active in the human genome, being mobilized *in trans* by L1 elements (Ostertag*, et al.* 2003), just as in the case of Alu elements. In

fact, the autonomous L1 LINEs, and non-autonomous Alu and SVA elements, are the only non-LTR retrotransposons known to be active in the human genome.

One should note that "SINES" and "LINES" were terms introduced in 1982 by Maxine F. Singer as denominations for Short and Long Interspersed Repeated Sequences in mammalian genomes, respectively (Singer 1982a, 1982b). The terms were attractive and caught on in the scientific community, but because there was no direct correspondence between these apparent acronyms, various interpretations were given, such as Short/Long Interspersed Nuclear Elements (Hassoun*, et al.* 1994, Craig 2002), or more simply, Short/Long INterspersed Elements (Deininger*, et al.* 1992).

1.2.2 DNA Transposons

DNA transposons are TEs that move in the host genome through a DNA intermediate. They are well represented in bacteria by Insertion Sequences (IS), but are present in many other genomes, including insects, worms, and human (Kazazian 2004). They have terminal inverted repeats, and encode a transposase that binds near the inverted repeats and mediates mobility through a "cut and paste" mechanism (Smit 1996, Lander*, et al.* 2001). This process is not usually a replicative one, unless the gap caused by excision is repaired using the sister chromatid (Smit 1996). When inserted at the new location, the transposon is flanked by small gaps, which, when filled by host enzymes, cause duplication of sequence at target site. These are called target site duplications (TSDs), and their length is characteristic for particular transposons (Reiss*, et al.* 2003). Based on sequence similarity of the transposase, eukaryotic DNA transposons fall into

two classes: the *Ac/hobo* class, characterized by 8 bp TSD, and the *Tc1/mariner* class, characterized by TA dimmer duplication (Smit 1996). The former is also referred to as the hAT superfamily, named for ***h****obo* element from *Drosophila melanogaster*, **A**c/Ds elements described by Barbara McClintock in corn, and ***T****am3* element from snapdragon (Robertson 2002). Members of both classes have been found in the human genome (Smit 1996, Lander*, et al.* 2001). It is estimated that DNA transposons make up about 3% of the human genome. Miniature Inverted-repeat Transposable Elements (MITEs) is a class of non-autonomous DNA transposons that were first described in maize (Bureau and Wessler 1992). They are short (100-500 bp) repeats with no coding potential, but which present terminal repeats and exhibit target site preference (TAA for *Stowaway*, and TA for *Tourist* elements) (Feschotte*, et al.* 2002). More recently, new classes of elements that move through DNA intermediates have been described. *Helitron* transposons are TEs that employ a rolling-circle mode of transposition (Kapitonov and Jurka 2001). They were found in plants, fungi, invertebrates, and fish, and recently also in the genome of the little brown bat, *Myotis lucifugus*, the first mammal known to contain *Helitron* transposons (Pritham and Feschotte 2007). *Mavericks* are the most recently described and the larget known (15-30 kb long) type of TEs (Feschotte and Pritham 2005, Pritham*, et al.* 2007). They are sometimes called *Polintons*, and their transposition through a double-stranded DNA intermediate is facilitated by a set of proteins that includes a protein-primed DNA polymerase B, retroviral integrase (c-integrase), cysteine protease, and ATPase (Kapitonov and Jurka 2006). The sequence and structure of *Mavericks* are similar to those of the atypical *Tlr1* elements, described in the ciliate *Tetrahymena thermophila* (Wells*, et al.* 1994). Because of their peculiar features, *Helitron* and *Maverick*

transposons are sometimes classified as neither DNA transposons nor retrotransposons, but as TEs of "other type".

1.3 The Impact of Transposable Elements on Evolution of Host Genomes

TEs are an important component of eukaryotic genomes, representing as much as almost half of the human genome (Lander*, et al.* 2001). The TE content of some grass species, such as maize and rice, is even higher, reaching 70% of the genome sequence (Wessler 2006). Different species differ by their genomic TE content, and also by the type of TEs found in their genomes. For example, the yeast *Saccharomyces cerevisiae* contains only LTR retrotransposons (*Ty* elements), while the biggest fraction of mammalian TEs is made of non-LTR retrotransposons (Wessler 2006). Many species, however, such as *Arabidopsis thaliana* among plants, and *Fugu rubripes* among vertebrates, have very compact genomes, with a very low TE content. This raised many questions about why other genomes have such a high TE content, and ultimately about the role of TEs. Originally, TEs, as any other non-exonic sequence, were perceived as "junk" (Ohno 1972), selfish, or parasite DNA (Doolittle and Sapienza 1980, Orgel and Crick 1980, Hickey 1982). This view of uselessness was consequently challenged when TEs were seen as "seeds of evolution" (Brosius 1991), or as a "genomic scrap yard" (Makalowski 1995) from which genomic novelties could arise. A great deal of attention was dedicated to ways in which TEs have been influencing the evolution of host genomes. The interest has been well rewarded, as indeed, TEs can impact their hosts in many ways (Makalowski 2000).

One way in which TEs can influence the evolution of their host genomes is by favoring recombination events. In some instances, such TE-mediated recombinations could lead to disease phenotypes (Kazazian 1998, Deininger and Batzer 1999). In others they can contribute to the expansion of certain gene families, such as the human glycophorin family that evolved through several duplication steps that involved recombination between Alu elements (Makalowski 2000). Alu-Alu mediated recombination events are also thought to have favored a significant fraction of recent segmental duplication that have occurred in the last 40 million years of human evolution (Bailey*, et al.* 2003). Such duplication can have a significant impact, because they can promote further rearrangement through their own misalignment (Eichler 2001) and subsequent non-allelic homologous recombination that can lead to the formation of genomic regions of complex architecture (Stankiewicz and Lupski 2002).

TEs can also provide different genomic motifs, such as transcriptional regulatory elements (Vidal*, et al.* 1993, Ferrigno*, et al.* 2001, Landry*, et al.* 2001, Jordan*, et al.* 2003, van de Lagemaat*, et al.* 2003, Thornburg*, et al.* 2006), poly-A signals (Paulson*, et al.* 1987), and even sequences to protein coding regions, mainly through the means of alternatively spliced exons (Nekrutenko and Li 2001, Sorek*, et al.* 2002, Lorenc and Makalowski 2003, Cordaux*, et al.* 2006, Gotea and Makalowski 2006). The influence of TEs on *cis*-regulation of gene expression by providing transcription factor binding sites has been intensively studied, but more recently a new regulatory mechanism has emerged. Alternatively spliced exons that carry an in frame premature termination codon are used to down regulate the amount of protein products of certain genes by being targets of the non-sense mediate decay (NMD) pathway (Lareau*, et al.* 2007, Ni*, et al.*

2007). Many TEs can provide alternatively spliced exons with premature STOP codons, especially Alu elements (Sorek*, et al.* 2002), thus a potential role in regulating gene expression post-transcriptionally needs to be taken seriously into consideration.

The current work is focused on two aspects of TE contribution to mammalian genomes. Chapters 2 and 3 are dedicated to the contribution of TEs to protein coding genes. Documenting cases of functional proteins with TE-encoded fragments is the subject of chapter two, which will consolidate previous evidence at the level of transcripts and conceptual translation. In chapter 3, I am trying to elucidate the coding potential of Alu elements, because they can be found in alternatively spliced transcripts of many genes but with no solid evidence for their protein coding potential. The high potential impact of TEs on *cis* regulation of gene expression is addressed in chapter 4. Chapter 5 is dedicated to the ScrapYard database, a collection of transcripts with annotated TE fragments that facilitates the study of TE contribution to mammalian transcriptomes and proteomes. Overall, this work advances our knowledge of the impact that TEs have on the evolution of genomes, and provides new insights into how TEs can generate genomic novelties.

Chapter 2

Contribution of Transposable Elements to Mammalian Proteomes

2.1 Introduction

As a major component of many eukaryotic genomes, TEs have had a major impact on their evolution (see Chapter 1.3). Of particular interest is the TE contribution to protein coding sequences, because they can directly influence the phenotype by altering protein sequences. This aspect was documented, however, only at the transcript level, and the presence of TE-encoded fragments was not confirmed at the protein level (Pavlicek, *et al.* 2002), which therefore required further clarifications.

Two decades ago, a few studies reported that some mRNAs contain TE cassettes in their coding regions (Lundwall, *et al.* 1985, Caras, *et al.* 1987, Brownell, *et al.* 1989) that sometimes resulted in disease phenotypes such as the gyrate atrophy of the choroid and retina (Mitchell, *et al.* 1991). These observations led to the hypothesis that, in other cases, TE exaptation could have neutral effects or even enhance fitness and, therefore, might increase protein variability with positive evolutionary consequences (Makalowski, *et al.* 1994). Since then, several studies discovered TE cassettes of many TE types in the coding region of many genes (Li, *et al.* 2001, Nekrutenko and Li 2001, Lorenc and Makalowski 2003). Despite a few reports of potentially functional proteins containing TE-encoded fragments (Gerber, *et al.* 1997, Hilgard, *et al.* 2002, Hoenicka, *et al.* 2002), there is no strong evidence that supports the existence of such proteins in vivo. The

presence of TE cassettes in transcripts does not guarantee their translation, because eukaryotic cells contain several quality control mechanisms that can initiate the degradation of the transcript and even of the protein product immediately after translation (Jacobson and Peltz 1996, Hilleren and Parker 1999, Wagner and Lykke-Andersen 2002). Moreover, even if translation occurs, the product might be non-functional and even 'mildly deleterious to the cell' (Lovell 2003). Consequently, overstatements such as 'many translated repetitive elements are found in proteins' (Deragon and Capy 2000) and 'pieces of TEs found in exons are translated in functional proteins' (Li*, et al.* 2001) are misleading in the absence of evidence at the level of functional proteins. Evolutionary implications of TE exaptation would be more profound if TE cassettes were present not only at the transcript level but also at protein level, given that 'proteins, rather than genes or mRNA, represent the key players in the cell' (Pradet-Balade*, et al.* 2001) by determining the cellular phenotype, and thus directly affecting fitness. Pavlicek et al. have argued that TE contribution to proteins can be reliably studied only with directly sequenced proteins or with proteins that have known three dimensional structures (Pavlicek*, et al.* 2002). Therefore, the contribution of TEs to the proteome needs to be confirmed (Lorenc and Makalowski 2003, Kazazian 2004).

2.2 Materials and Methods

2.2.1 Selection of Proteins to be Analyzed

For this analysis, I used only human proteins from the Protein Databank (PDB) (Kouranov*, et al.* 2006), for which three dimensional structures were determined, and directly sequenced proteins from Swiss-Prot (Bairoch and Apweiler 1997) (starting with June 17, 2004, Swiss-Prot implemented the 'direct protein sequencing' – DPS – keyword for proteins with sequences that were completely or partially determined by direct sequencing, as opposed to conceptual translation from mRNA). As of December 4, 2005, ~4,000 non-redundant human protein chains (they correspond to 3,764 PDB entries) were available in PDB, and as of August 3, 2004, 1,765 human protein sequences were available in Swiss-Prot.

2.2.2 Detection of TE Occurrences

For this step, the mRNA sequences that encode the above mentioned proteins were used, because both TE libraries and specific software were developed for use with DNA sequences. Because the PDB does not provide mRNA sequences that encode the proteins it hosts, NCBI's tBLASTn program was used, with protein sequences as query against the GenBank collection of human mRNAs (178,509 non-redundant sequences as of December 4, 2005). Perfect or almost perfect matches (allowing for mismatches caused by non-synonymous SNPs, ambiguous residues or His-tails in PDB protein sequences) between PDB proteins and mRNA sequences were considered for further

analysis. For proteins from Swiss-Prot, I used the links provided in the Swiss-Prot records to find appropriate mRNA sequences. More than 25,000 sequences (both protein and DNA) were linked from the records of the 1,765 DPS human proteins. Using Batch Entrez, I determined that only 6,047 represent mRNA sequences, which were used for further analyses. I then used RepeatMasker (RM) version 3.0.8 (http://repeatmasker.org) with cross_match version 0.990329 (http://www.phrap.org/phredphrapconsed.html) as searching engine, and '-s' (greater sensitivity) and '-gccalc' (GC aware) parameters to detect occurrences of putative TE cassettes in transcripts. Scores of RM matches are obtained from Smith–Waterman alignments provided by cross_match with scoring matrices tailored by RM for the GC background of every batch of sequences analyzed when the '-gccalc' option is used. Those scores can be further altered by RM based on the complexity of matches (see RM documentation at http://www.repeatmasker.org/ webrepeatmaskerhelp.html or using the '-help' option). RM cut-off scores were determined to minimize the number of false positives reported, with the lowest value of 180 for old repeats (see RM documentation). All matches reported by RM were considered for further analysis without imposing additional score or length thresholds. Overlaps between TE occurrences and protein sequences were detected using custom Perl scripts. In the case of transcripts coding for DSP proteins, the references provided in the Swiss-Prot records were checked to learn whether the TE fragment overlapped with the DPS fragment. This was necessary because many mRNA sequences are splicing variants of the same gene, and for many proteins the DPS was done only for short segments.

2.2.3 Phylogenetic Reconstruction and Analysis

For each of the proteins in Tables 2.1, 2.2 , I searched for similar proteins in NCBI and ENSEMBL protein databases using BLAST software (BLASTp, PSI-BLAST) available from NCBI. Where necessary, tBLASTn searches against genomic sequence were conducted to retrieve additional sequences. Sequence alignments, highlighting and manipulations were performed using ClustalW (Thompson*, et al.* 1994), Bioedit v7.0.1 (http://www.mbio.ncsu.edu/BioEdit), and manually, where necessary. Distances were computed with the modified Nei-Gojobori method (Nei and Gojobori 1986), and trees were built using the neighbor-joining method (Saitou and Nei 1987) implemented in MEGA3 (v3.1) software (Kumar*, et al.* 2004). MEGA3 was also used for computing dN/dS ratios.

2.2.4 Statistical Analyses

A sequence randomization test was conducted for every PDB protein found to have a putative TE-encoded fragment to assess the probability of a random match to that particular TE in the coding region of the gene. For each case, 10,000 random sequences with identical length and nucleotide composition to the CDS of the gene were generated. RM with '-s' and '-gccalc' options was then used to detect random TE matches in these sequences, which I further used to create the distribution of expected random TE matches per sequence. For this step I only used the matches to TEs of the same type as the TE identified in the real gene. P-values associated with the matches in the real genes were determined from these distributions. The hypotheses of neutral evolution were tested

using the Z-test implemented in MEGA3 (v3.1) software (Kumar, *et al.* 2004). The significance of difference between average vertebrate-invertebrate pairwise identity of PTP domain and the L3-like segment was tested using the Wilcoxon (Mann–Whitney) rank sum test implemented in Minitab 14 (http://www.minitab.com).

2.2.5 Representation of Protein Structures

PyMOL (http://pymol.sourceforge.net) was the software of choice to create cartoon representation of the protein structures.

2.3 Results and Discussion

Among the 3,764 PDB entries with non-redundant protein chains, RM reported the presence of a TE fragment in only 24 cases (Tables 2.1, 2.2). No additional cases were found among the 1765 directly sequenced proteins from the Swiss-Prot collection. A common feature of all TE cassettes identified in the 24 proteins is their low RM score, which are all close to the empirically set thresholds for false positive matches, which therefore required further validation. A TE match could be validated if the exaptation of the TE cassette could be explained in the context of the phylogeny of the host gene. Not surprisingly, phylogenies of 21 proteins (Table 2.1) do not support the presence of a TE cassette as reported by RM. In 13 examples, where the putative TE cassette is located within one exon, the encoded fragment is conserved in invertebrate orthologs, which is inconsistent with the known activity times of the reported TEs. In eight examples, the putative TE cassette spans across multiple exons, which could be reasonably explained only by intron gain during vertebrate evolution. This would be a reasonable

Table 2.1. False TEs found by RM in protein coding sequences. The phylogenies of host genes do not support these matches as real TE cassettes.

Gene[a]	Corresponding PDB Structure[b] (PDB ID:chain)	mRNA gi	CDS Length (nt)	GC content of CDS (%)	Type of Matching TE	Length of TE in mRNA (nt)	Divergence from TE consensus (%)	RM Score	p-value[c]	Notes
AASDHPPT	2C43:A	7106835	930	43.33	L4	67	28.1	186	0.0001	Located in exon 6; contains the STOP codon.
AMY1A	1Z32:X	38648818	1536	41.67	L3	62	18	181	0.0014	Spans exons 5-6.
APEX1	1DEW:A	219477	957	52.66	L1PA14	121	27.6	275	0	Spans exons 3-4.
ARFIP2	1I49:A	38569401	1026	55.17	LTR27*	59	22.8	229	0	Located in exon 4.
CASP7	1SHJ:A	1125072	912	48.46	L3	65	36.9	184	0.001	Located in exon 6.
CAT	1QQW:A	29720	1584	49.18	L3*	68	19.7	204	0.0003	Located in exon 5.
CDK7	1UA2:A	34783068	1041	41.5	L4	113	30.6	184	0.0004	Spans exons 5-6.
FKBP4	1QZ2:A	33873773	1380	53.04	L3	96	36.5	191	0.0014	Spans exons 8-9.
GALM	1SO0:A	15928900	1029	53.26	L4	47	22.2	183	0.0004	Located in exon 4.
ITGB3	1U8C:B	2443451	2355	54.56	L3*	70	29	197	0.0024	Located in exon 5.
MHCII HLA-DQA1	1JK8:A	619821	681	51.1	L4	29	13.8	182	0.0001	Partial CDS. There is no precise location of this gene in the human genome.
MME	1DMT:A	29625	2253	41.32	L2*	36	17.1	184	0.0014	Located in exon 3; alignment also includes 3 nt 5' of exon 3.
PIK3CG	1E8Y:A	1507821	3306	50.67	L3	73	23.5	194	0.0014	Spans exons 2-3.
PRDX5	1OC3:A	8745393	645	59.22	MIR*	132	28.8	200	0	Located in exon 6; contains the STOP codon.
RAP1GA1	1SRQ:A	32452013	1992	58.99	L3	158	34.7	183	0.001	Spans exons 10-12.
RNH1	1A4Y:A	33875733	1386	64.57	L2	101	25.8	190	0.0002	Spans exons 6-7.
SPTA1	1OWA:A	4507188	7290	48.63	MER5A	66	31.2	193	0.0014	Located in exon 2.
SUFU	1M1L:A	6689887	1302	58.68	L3*	89	25	209	0.0006	Located in exon 2.
TMPO	1H9F:A	31566193	1365	46.3	L3	75	33.8	185	0.0027	Located at the 3' end of exon 2.
XPO1	1W9C:A	53759152	3216	37.1	L4	43	23.3	202	0.0023	Located at the 3' end of exon 17.
XRCC5	1RW2:A	12408650	2199	42.56	L3	123	31.7	189	0.0003	Spans exons 19-21.

Legend: [a]Genes are presented alphabetically by their official NCBI gene symbol; [b]Even though several structures might be available for certain proteins, only one is shown here; [c]p-values represent the tail probability of a random match to the specific TE type found in the gene's CDS (see Materials and methods for details on the sequence randomization test). In the case of APEX1, ARFIP2, PRDX5, and SPTA1 genes, p-values were computed based on all matches to L1, LTR, MIR, and MER5 repeats, respectively; *As of December 2005, these putative TEs can be visualized in the USCS Human Genome Browser (May 2004 assembly; http://genome.cse.ucsc.edu/).

scenario (Rogozin, *et al.* 2003), but one that I could not confirm for any of the eight cassettes, because either the fragment and gene structure were conserved in invertebrate orthologs, or the fragment was not well conserved in vertebrates. In either case, the data suggest that the putative TE cassettes are probably random matches to TE consensus sequences and not real TE cassettes. In three cases (Table 2.2), however, the phylogenies of the host proteins suggest that exaptation events could have occurred in the past as detailed below.

Table 2.2: Human proteins with TE-encoded fragments

Gene	PDB structure	mRNA gi	CDS length (nt)	GC content of CDS (%)	TE type	Length of TE cassette (nt)	Divergence from TE consensus (%)	RM score	*P*-value
CAPN1	2ARY:A	49900978	2145	59.53	MIRm	34	17.6	182	0.0003
GZMA	1OP8:A	184022	789	43.73	L3	85	34.1	183	0
PTPN1	1G7F:A	17390366	1308	50.69	L3	101	27.7	198	0.0013

2.3.1 The Protein Tyrosine Phosphatase, Non-Receptor Type 1 (PTPN1)

PTPN1, also known as PTP1B, is a 435-amino-acid protein that belongs to the large family of protein tyrosine phosphatases (PTPs), which catalyze protein dephosphorylation. Its sequence was initially determined by direct sequencing (Charbonneau, *et al.* 1989), its three dimensional structure was first determined by Barford et al. (1994) and its functionality has been detailed by several consequent studies (links to all PDB structures can be accessed via the Swiss-Prot record of PTPN1, accession number P18031). RM finds remnants of an L3 LINE in the coding region of the

corresponding mRNA (NCBI gi:17390366) between coordinates 499 and 599 (Figure 2.1).

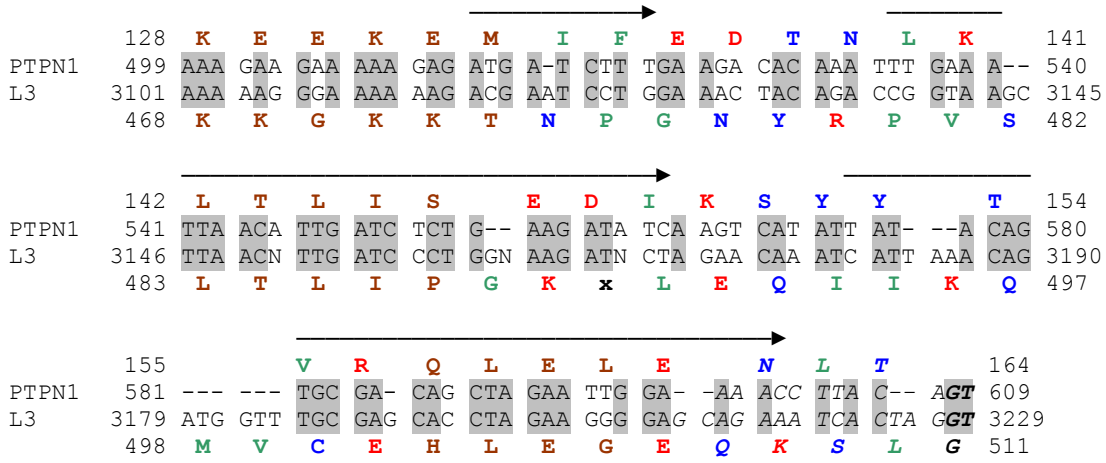

Figure 2.1: Alignment of human PTPN1 mRNA (gi:17390366) with the L3 consensus sequence, as determined by RepeatMasker. The alignment was manually extended (*italics*) until the end of the fifth exon, including the donor splice site. β-strands occurring in human PTPN1 are shown above the alignment. Conserved nucleotides are shaded, and encoded amino acids are shown above the corresponding second codon position. Residues for which the reading frame was preserved between L3 ORF2p and PTPN1 are shown in brown. Residues from regions of non-preserved reading frame are shown in green (hydrophobic), red (charged) and blue (polar).

Although the length, divergence from consensus and RM score are similar to those of the TE cassettes that can be considered false positives, several arguments support the validity of the L3 cassette. The first argument is the origin of the L3 cassette in the second open reading frame of the L3 element (ORF2p). The L3 non-LTR retrotransposon is among the most ancient TEs reconstructed *in silico* and is characterized by the presence of two ORFs (Kapitonov and Jurka 2003). The second ORF is estimated to be ~902 residues long (Repbase v10.12, http://www.girinst.org) and contains a well-

conserved reverse-transcriptase (RT) domain between coordinates 457 and 712, characteristic of retrotransposons and retroviruses. The L3 fragment found in the PTPN1 mRNA corresponds to part of this RT domain (residues 468–506, Figure 2.1). In fact, the L3 cassette donates almost unchanged the core residues of two β-strands that are part of a four-strand anti-parallel β-sheet (Figures 2.1, 2.2). It is difficult to imagine that such a complex structure could have been generated by a sequence that did not have previous coding capacity. We see a similar situation in GZMA (see Chapter 2.3.2), consistent with the exaptation hypothesis, which implies the reuse of characters (i.e. protein coding sequences) for different functions.

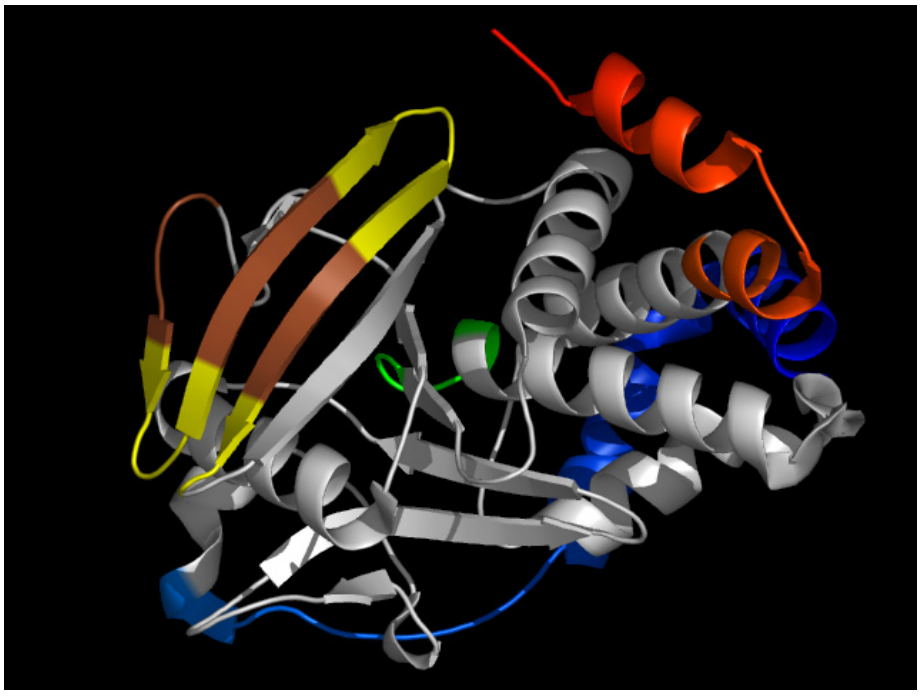

Figure 2.2: The three dimensional structure of human PTP-1B (PDB ID: 1G7F): N-terminus in blue, C-terminus in red, PTP domain in white, base of the active site cleft in green (Cys^{215} is the essential catalytic residue), L3 encoded fragment in brown (preserved reading frame) and yellow (non-preserved reading frame).

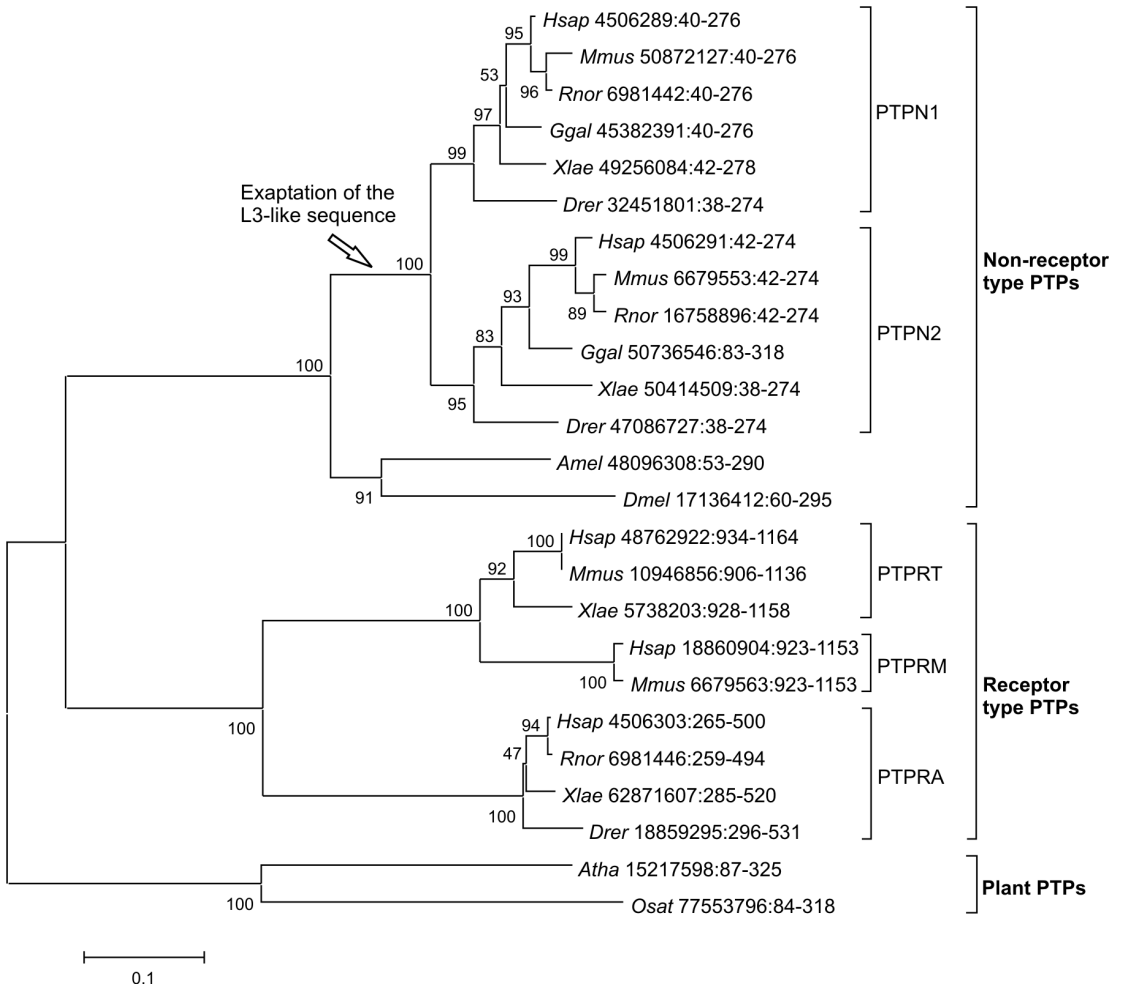

Figure 2.3: Phylogenetic history of human PTPN1. For constructing the tree, I used the neighbor-joining method [57], 188 sites after complete gap deletion, Poisson corrected distance and 1,000 bootstrap replicates. The fragment corresponding to the L3 (coordinates 89–127 in Figure 2.4) was excluded for this step. Two plant PTPs (*Atha – Arabidopsis thaliana*, *Osat – Oryza sativa*) were added to root the tree of animal PTPs. The same tree topology was obtained by using maximum parsimony and maximum likelihood methods of phylogenetic reconstruction.

A second argument supporting the validity of the L3 cassette is provided by the origin of PTPN1. It is known that PTP diversification occurred by a series of duplication

events during early vertebrate evolution (Ono*, et al.* 1999, Ono-Koyanagi*, et al.* 2000, Andersen*, et al.* 2004). This can explain why PTPN1 is located ~7.3-Mb apart from PTPRT on chromosome 20q, similar to their closest homologs, PTPN2 and PTPRM, respectively (Figure 2.3), which are located ~4.4-Mb-apart on chromosome 18p (Andersen*, et al.* 2004). The most likely scenario is that an intra-chromosomal duplication was followed by the exaptation of the L3-like sequence, followed by a larger inter-chromosomal duplication. This can explain why the L3 cassette is strongly conserved between PTPN1 and PTPN2, but seems strikingly non-conserved in the invertebrate non-receptor type PTPs (Figure 2.4). The average identity between vertebrate and invertebrate sequences for this segment (23.56±10.67%) is significantly smaller (P-value<<0.001) than that of the rest of the PTP catalytic domain (65.71±1.85%). In addition, the gene structure of invertebrate non-receptor type PTPs seems to have undergone major rearrangements (Figure 2.5).

Although plausible, the timing of the events according to this scenario implies that L3 retrotransposons were active much earlier than the current estimate, which places L3 activity before the mammalian radiation (Kapitonov and Jurka 2003). Perhaps the L3 retrotransposon is much older than the current estimate of >200 million years because of the bias towards more-conserved copies used for the reconstruction of the consensus sequence (Kapitonov and Jurka 2003). An alternative explanation could be that the fragment identified as L3 originates in an older unknown RT-carrying retrotransposon. Support for this hypothesis is provided by the strong purifying selection observed in the vertebrate lineage (Table 2.3), which maintained the similarity to the original RT domain so that it now resembles the oldest known RT-carrying retrotransposon: the L3 LINE.

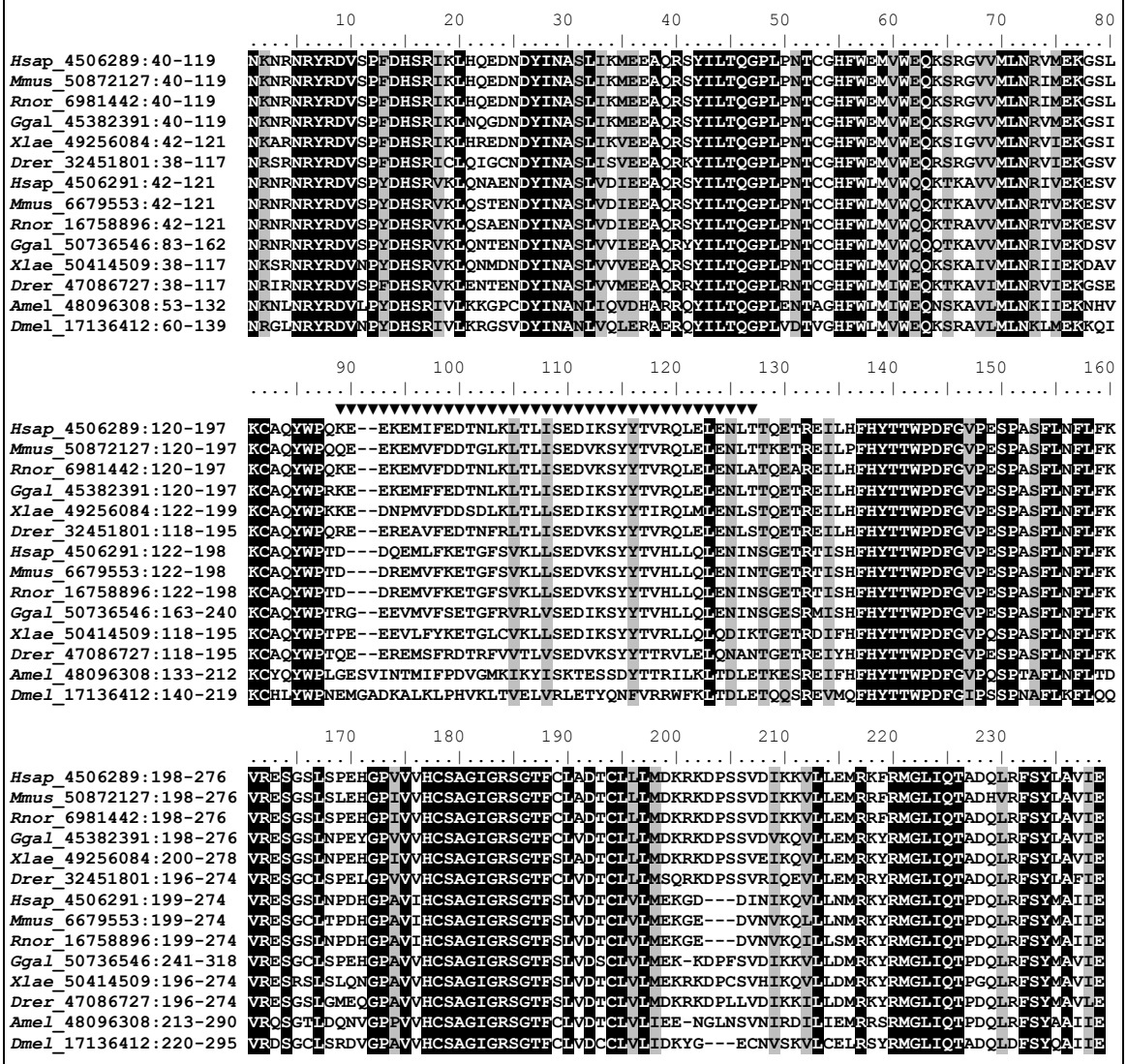

Figure 2.4: The amino-acid alignment of the PTP catalytic domain of animal non-receptor type PTPs. Species name, gi accession number and coordinates of residues included in alignment are indicated before every sequence. Identical and similar residues are highlighted in black and grey, respectively. The symbols (▼) above the alignment correspond to the L3-encoded fragment in PTPN1 (coordinates 89–127). Alignment coordinates 125–127 correspond to the manually extended alignment (shown in italics in Figure 2.1).

Table 2.3: dN/dS values calculated for the catalytic domain of vertebrate and invertebrate non-receptor type PTPs

		PTPN1						PTPN2							
	Sp.	*Hsap*[a]	*Mmus*[b]	*Rnor*[c]	*Ggal*[d]	*Xlae*[e]	*Drer*[f]	*Hsap*[a]	*Mmus*[b]	*Rnor*[c]	*Ggal*[d]	*Xlae*[e]	*Drer*[f]	*Amel*[g]	*Dmel*[h]
PTPN1	*Hsap*[a]		0.142	0.090	0.022	0.271	0.151	0.657	0.475	0.573	0.387	0.481	0.505	0.737	0.964
	Mmus[b]	0.054		0.151	0.081	0.264	0.200	0.426	0.381	0.438	0.394	0.514	0.421	0.650	0.981
	Rnor[c]	0.035	0.084		0.056	0.282	0.120	0.463	0.445	0.477	0.386	0.551	0.434	0.602	0.892
	Ggal[d]	0.043	0.078	0.062		0.301	0.162	0.571	0.579	0.571	0.446	0.395	0.366	0.789	0.764
	Xlae[e]	0.059	0.083	0.064	0.088		0.428	0.800	0.600	0.668	0.492	0.542	0.440	0.555	1.048
	Drer[f]	0.109	0.143	0.128	0.136	0.139		0.468	0.350	0.436	0.438	0.503	0.341	0.691	0.985
PTPN2	*Hsap*[a]	0.219	0.236	0.221	0.234	0.234	0.241		0.086	0.095	0.276	0.272	0.449	0.956	1.096
	Mmus[b]	0.225	0.248	0.224	0.243	0.234	0.243	0.073		0.000	0.241	0.303	0.366	1.309	1.029
	Rnor[c]	0.229	0.253	0.231	0.218	0.225	0.248	0.062	0.124		0.262	0.309	0.372	1.069	1.088
	Ggal[d]	0.152	0.175	0.158	0.217	0.224	0.211	0.119	0.131	0.144		0.256	0.437	0.969	0.983
	Xlae[e]	0.189	0.198	0.186	0.199	0.180	0.199	0.155	0.190	0.186	0.122		0.406	1.019	0.874
	Drer[f]	0.199	0.225	0.213	0.207	0.191	0.227	0.213	0.178	0.191	0.150	0.156		0.712	0.809
	Amel[g]	0.293	0.297	0.285	0.347	0.337	0.289	0.347	0.372	0.358	0.345	0.349	0.304		0.698
	Dmel[h]	0.385	0.461	0.438	0.354	0.360	0.341	0.369	0.358	0.368	0.342	0.369	0.452	0.290	

Legend: [a]*Homo sapiens*, [b]*Mus musculus*, [c]*Rattus norvegicus*, [d]*Gallus gallus*, [e]*Xenopus laevis*, [f]*Danio rerio*, [g]*Apis melifera*, [h]*Drosophila melanogaster*. Order of sequences is the same as in Figure 2.3. Modified Nei-Gojobori method using p-distance and complete deletion was used (Jukes-Cantor correction was not applied because the p-distance for many pairs is higher than 0.75). Shaded cells correspond to pairs for which assumption of neutrality cannot be rejected by a Z test. Variance was computed analytically for the L3 corresponding segments because of the low number of codons. Significance was tested with the variance computed both analytically and by bootstrap for the rest of the domain. The segment corresponding to the L3 consensus (89-127 in Figure 2.4) is not considered homologous between vertebrate and invertebrate sequences, and therefore dN/dS values are meaningless for those pairs (presented in faded black). Values in the upper-right were computed for the segment matching the L3 consensus (89-127 in Figure 2.4). Values in the lower-left side were computed for the rest of the PTP domain (195 codons).

According to Ohno, gene duplications create the raw material for evolutionary "innovations" (Ohno 1970). He argues that newly duplicated genes are free of functional constraints and can undergo significant changes until they acquire new functions. Provided that duplications are the documented source for PTP diversification as discussed earlier, it is easy to imagine that the future PTPN1 could have easily acquired a TE fragment after the activation of a cryptic splice site in a manner similar to how genes currently acquire Alu fragments (Makalowski*, et al.* 1994, Lev-Maor*, et al.* 2003). The

position of the L3-like fragment at the end of exon 5 (Figure 2.5) supports this hypothesis (because non-active TE sequences are expected to mutate beyond recognition (Makalowski 2001), it is not surprising that we cannot extend the alignment into the 3' intronic region). One could also note that the ratio of dN to dS between the L3 consensus and the fragment identified in the human PTPN1 mRNA is almost one in either L3 or human PTPN1 reading frames (0.95 and 0.71, respectively – the assumption of neutrality cannot be rejected by a Z-test in either example). This is consistent with a period of neutral evolution that might have affected the TE in intron before exaptation. However, it is also consistent with a period of positive selection that the fragment could have experienced following exaptation, but before PTPN1 acquired a new specific function. This can explain why, despite the reasonably good nucleotide conservation, the coding function of the TE fragment has changed considerably (Figures 2.1, 2.2). In addition to the expected nucleotide substitutions, deletions that were not multiple of three nucleotides determined the change of reading frame for two segments (Figure 2.1). Consequently, the number of hydrophobic residues was reduced from nine to four in those regions, facilitating the formation of hairpin loops (Figure 2.2). However, it is difficult to say whether most of these changes occurred after exaptation to increase protein functional efficiency, as can occur with any random sequence (Hayashi*, et al.* 2003), or if they occurred in a fortuitous manner that actually facilitated the exaptation event. Whatever the case, it is remarkable that following the exaptation event and the subsequent intra-chromosomal duplication, both PTPN1 and PTPN2 acquired specific functions (Table 2.3) that probably do not exist in invertebrates, which have much fewer PTPs.

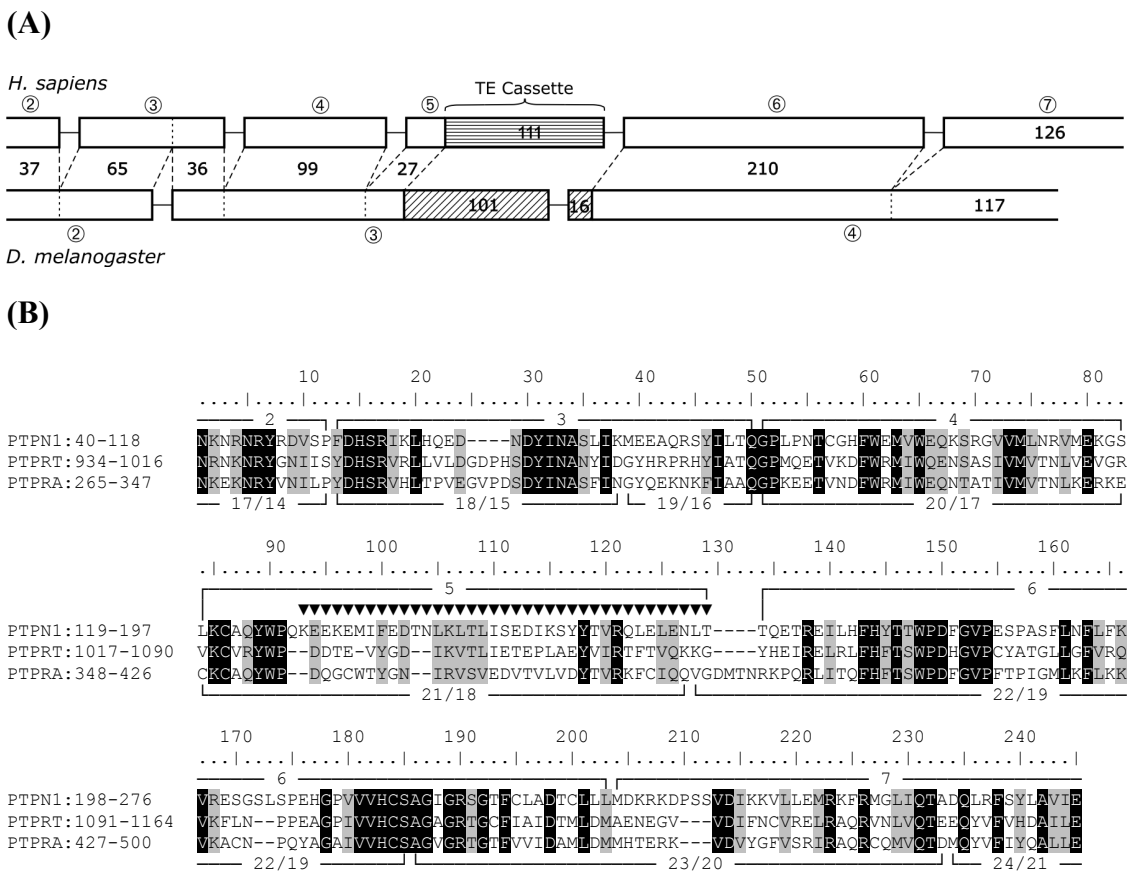

Figure 2.5: A comparison of PTPN1 gene structure with other invertebrate and human PTPs. **(A)** Gene structure and correspondence between the PTP domain of human PTPN1 and *D. melanogaster* PTP61F gene. Exon/intron boundaries were determined using Spidey (www.ncbi.nlm.nih.gov/spidey) and the following sequences (NCBI gi numbers): 17390366 (mRNA), 51511747 (genomic) for human, and 24655162 (mRNA), 56411837 (genomic) for *D. melanogaster*. Sizes, in nucleotides, of corresponding exonic blocks are indicated between the two genes if identical, or in each exon if different between the two species. Exon numbers are shown in circles. Introns are not represented to scale. **(B)** Alignment of the PTP catalytic domain of human PTPs located on chromosome 20. Exon boundaries and numbers are shown above the alignment for PTPN1, and below the alignment for PTPRT and PTPRA. For the latter, the first number corresponds to PTPRT, and the second for PTPRA. While the gene structures of PTPRT and PTPRA are identical, that of PTPN1 is completely different after exon 5 which contains the L3-encoded fragment. It is hard to determine, however, the direction of causality, i.e. exaptation determined change of gene structure or vice-versa.

2.3.2 The μ-calpain (CAPN1)

CAPN1 is one of many Ca^{2+}-dependent intracellular proteinases (calpain family) that exist in the human genome. Together with CAPN2, CAPN1 is ubiquitously expressed, whereas other calpains are expressed in specific tissues such as muscle (CAPN3) (Sorimachi, *et al.* 1989), digestive tract (CAPN9) (Lee, *et al.* 1998), or testis (CAPN11) (Dear, *et al.* 1999, Rojas, *et al.* 1999). CAPN1 is a 22-exon gene, whose second exon contains a MIRm element (Figure 2.6a). This element appears to be conserved in other calpains (RM does not report it in other species because of lower conservation), such as CAPN2 and CAPN11, as well as in other vertebrates including fish (Figure 2.6a). This is not surprising, because the MIR tRNA related region, to which the CAPN1 fragment corresponds, is known to be conserved in non-mammalian vertebrates (Gilbert and Labuda 1999).

(A)

```
                67   S  L  G    Y  K  D  L  G  P  N  S  77
CAPN1          298 GAGCCTGGG--TTACAAGGACCTGGGTCCCAATTCC 331
MIRm            34 GGGCCTGGGAGTCAGAAGGACCTGGGTTCTAATCCC 69
```

(B)

```
Hsap1          298 GAGCCTGGG--TTACAAGGACCTGGGTCCCAATTCC 331
Mmus1          256 TAGCCTGGG--CTTCAAGGAACTGGGTCCTCATTCC 289
Drer1          401 TTCCCTGGG--CTTCAATGAACTCGGCCCGCGATCC 434
Hsap2          310 GGCCCTGGG--CTTCAAGGAGTTGGGGCCCTACTCC 343
Hsap11         308 CTCACTGGG--CTTCAAGGACCTGGGCCCCAACTCC 341
MIRm            34 GGGCCTGGGAGTCAGAAGGACCTGGGTTCTAATCCC 69
```

Figure 2.6: Alignment of MIRm element with calpain sequences. **(A)** RM alignment with human CAPN1 (gi:49900978). Encoded amino acids are shown above corresponding second codon positions. **(B)** Alignment of CAPN1 MIR-like homologous sequences from other species, mouse (gi:31982478) and zebrafish (gi:47087454), and calpain types, human CAPN2 (gi:12408645)and CAPN11 (gi:45446753). The number following the species name indicates the calpain type.

In spite of this, the MIRm element appears to be missing from basal vertebrates according to the UCSC Genome Broswer (Figure 2.7). A similar situation exists for six additional false positive TE matches (Table 2.1), which shows that masking low-scoring RM matches may not always be appropriate.

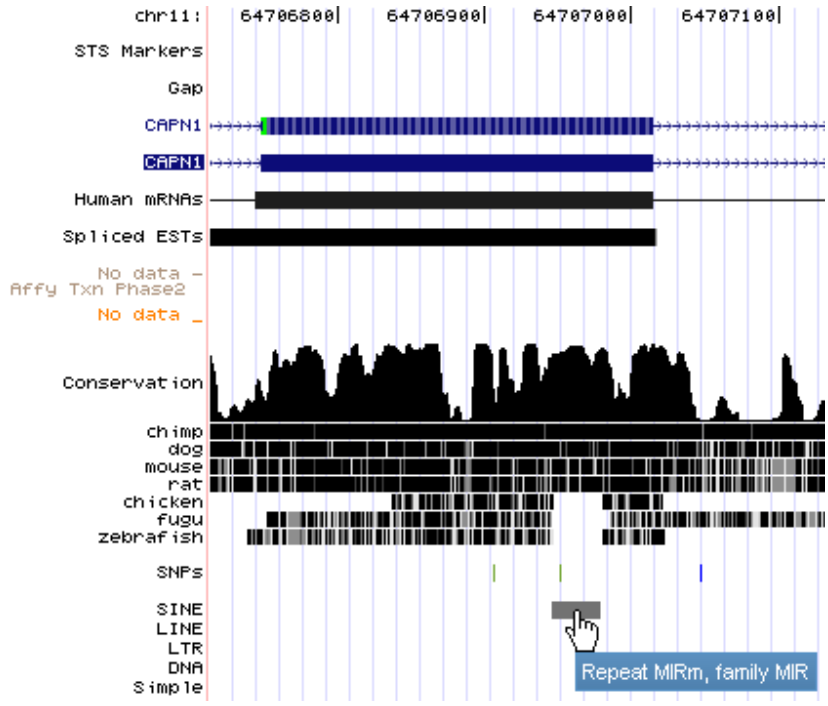

Figure 2.7: Partial view of the second exon of human CAPN1 in the USCS Genome Browser (http://genome.cse.ucsc.edu). The MIRm element erroneously appears to be missing in lower vertebrates.

The MIR-encoded fragment (residues 67-77; Figure 2.6a) is part of the calpain peptidase domain (residues 55-354 as determined by a Pfam search). A multiple sequence alignment of eukaryotic calpains (Figure 2.8) reveals that not all of them contain a fragment homologous to the MIR-encoded CAPN1 fragment. Invertebrate calpains, as well as human calpains 3, 5, 6, 9, and their vertebrate orthologs contain a shorter fragment instead. Similarly to the case of PTPN1, where the L3-encoded fragment was

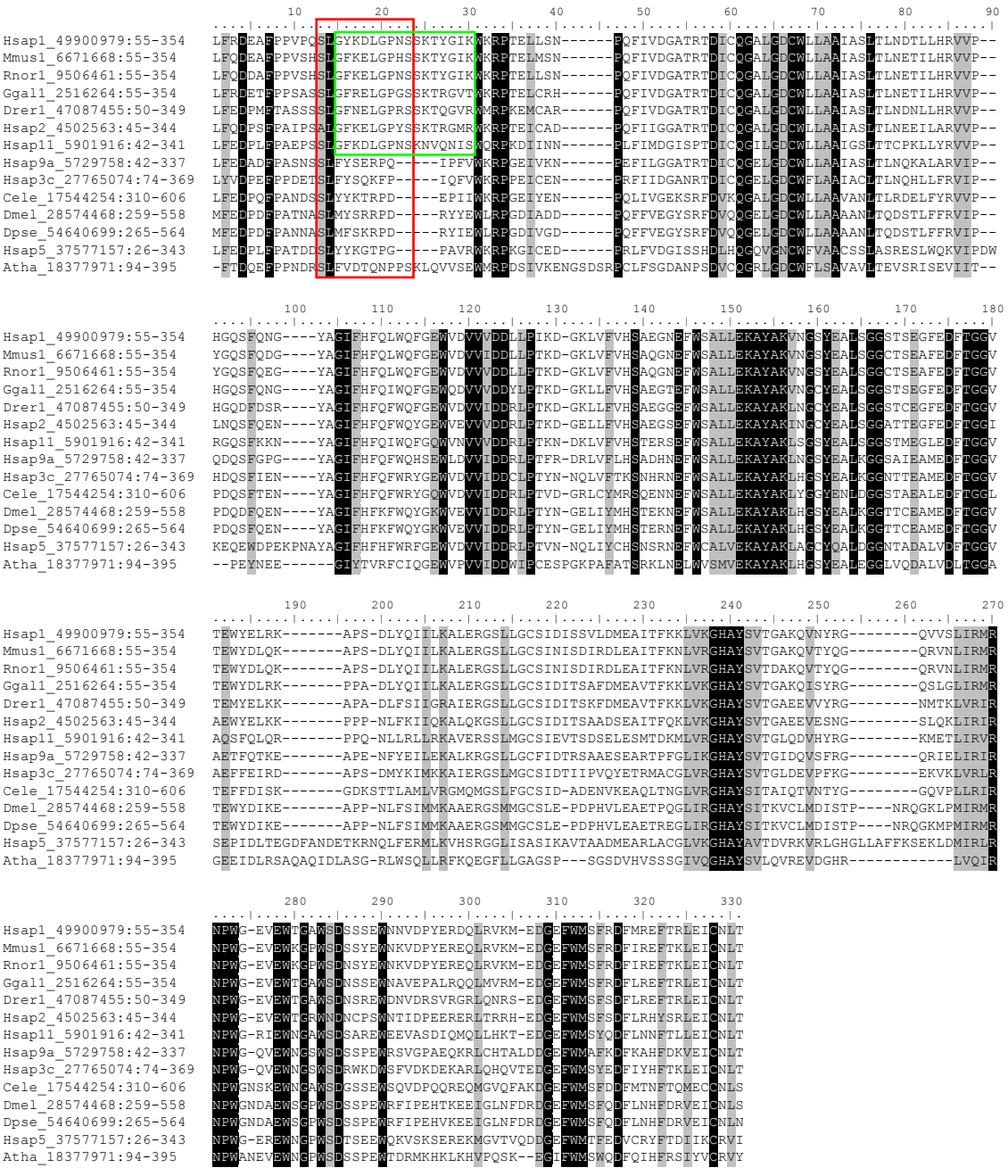

Figure 2.8: Multiple sequence alignment of eukaryotic calpains. Numbers following species names indicate the calpain type (for vertebrates), while letters (where the case) indicate the splicing variant. Sequence gi numbers and coordinates included in the alignment follow. Black and grey shadings indicate identical and similar residues, respectively. The red box delimitates the portion of the alignment (coordinates 13-23) that corresponds to the MIR-encoded human CAPN1 fragment, while the green one includes the inferred exapted fragment.

conserved only in vertebrate species, the MIR-encoded fragment is restricted to a subset of vertebrate calpains (1, 2, 8, 11 – CAPN8 not shown), generating a distinct conservation pattern (red box in Figure 2.8). Unlike the rest of the fragment, Ser^{67} and Leu^{68} appear to be highly conserved in all calpains, as they are part of a short α-helix, highly conserved as well (see the following PDB structures: 1DF0 for rat CAPN2, 1ZIV for human CAPN9). Interestingly, the α-helix is not conserved in human CAPN2 (PDB ID: 1KFX), where Ser^{57} is replaced by Ala.

The phylogeny of eukaryotic calpains reveals that all calpains containing the MIR-like fragment form a distinct cluster (Figure 2.9). This leads to the conclusion that the exaptation of the MIR fragment probably occurred in the common ancestor of eukaryotic calpains 1, 2, 8, and 11. This coincides with the first round of tandem duplications inferred to have taken place after the first chordate genome duplication (Jekely and Friedrich 1999). In fact, calpain family appears to have been generated by several rounds of chromosomal and tandem duplications, which might have facilitated the exaptation event. This is because function of duplicated genes is not a prerequisite and they may freely change after the duplication event. In this process they could lose functionality completely (pseudogenes), become functionally redundant with the original gene, or acquire new functions, case in which dramatic changes will not be further tolerated.

A more careful analysis of the alignment reveals that an extended fragment, up to position 30 in Figure 2.8, shows a conservation pattern similar to that of the MIR-encoded cassette. Because it immediately follows it, this fragment covers the five-residue gap observed in the alignment, and is conserved only among the same limited number of

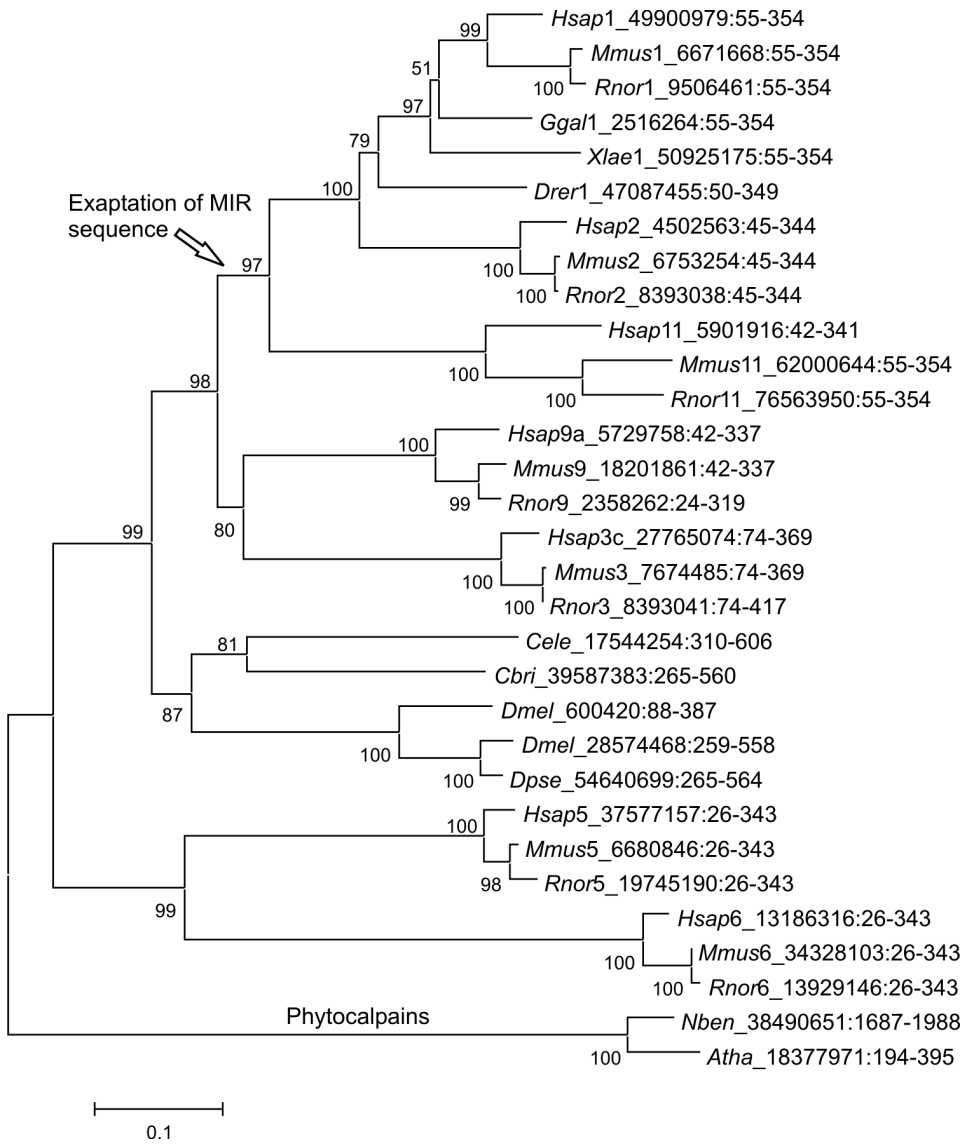

Figure 2.9: The phylogenetic tree of eukaryotic calpains. Neighbor-joining method, complete gap deletion, 282 sites, Poisson corrected distance, and 1000 bootstrap replicates were used for constructing the tree. The arrow indicates the inferred exaptation event. *Cele* - *Caenorhabditis elegans*; *Cbri* - *Caenorhabditis briggsae*; *Dpse* - *Drosophila pseudoobscura*; *Nben* - *Nicotiana benthamiana*.

calpains, this entire fragment could have been subject to exaptation as well (green box in Figure 2.8). It is certainly conceivable that sequence adjacent to the MIR element was included in the exaptation event. Residues 69-84 correspond to a loop region (green in

Figure 2.8) that connects a short α-helix at its N-terminus to a short β-strand at its C-terminus. This region of reduced structural complexity might have also facilitated the exaptation event. This is particularly relevant for the MIR element, which had no previous coding capacity, and consequently it is unlikely that the same fragment could have been tolerated in structurally more complex regions. In the cases of PTPN1 and GZMA, where, interestingly, both exapted fragments form β-strands, the TE-cassettes originated from L3 ORF2p, thus having had "innate" coding capacity.

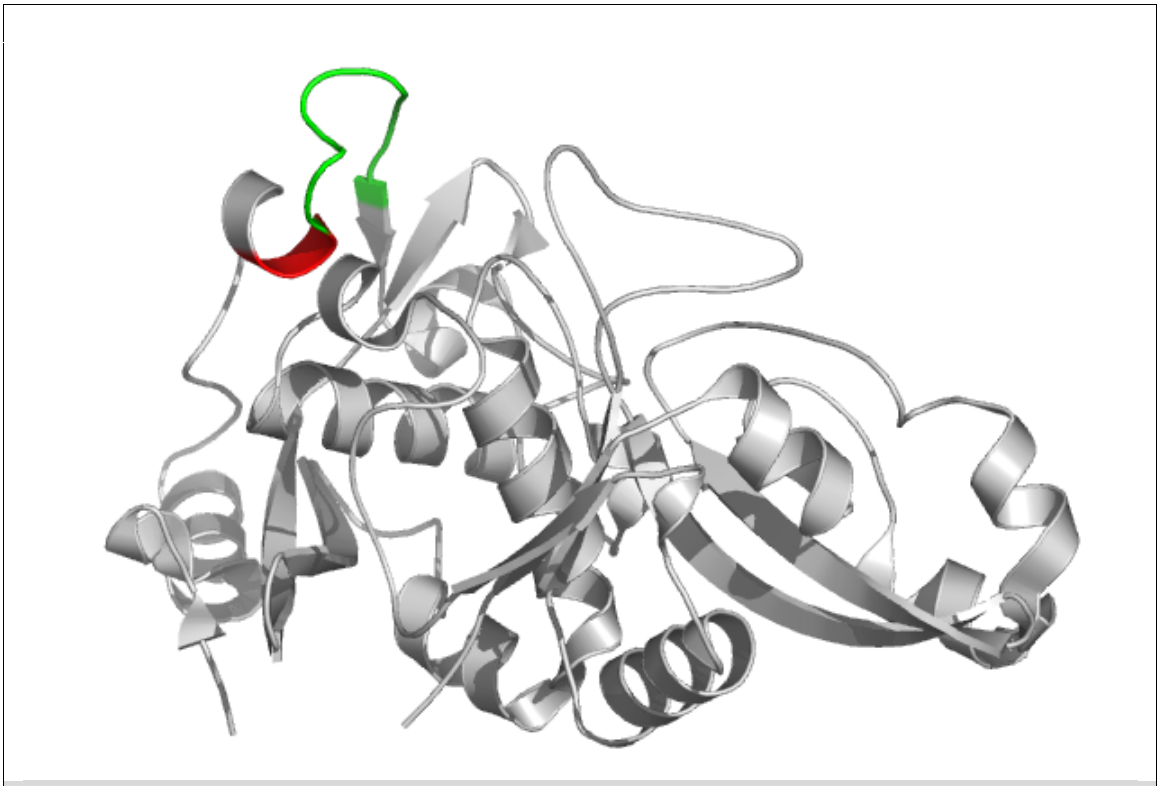

Figure 2.10: Three dimensional structure of human CAPN1 (PDB ID: 2ARY). Conserved Ser^{67} and Leu^{68} are colored in red. The inferred exapted fragment, that corresponds to non-conserved residues Gly^{69}-Lys^{84} (positions 15-30 in Figure 2.8), is colored in green.

On the other hand, it is certainly arguable whether the inclusion of the conserved Ser[67] and Leu[68] residues in the MIR alignment is due to chance, and if so, whether the entire match is due to chance. In the sequence randomization test I only found three random matches to MIR elements (p-value: 0.0003; Table 2.2), from which only one was MIRm, but from a different region than the fragment in CAPN1. Even if virtually null, the estimated chance of a random match is, however, probably irrelevant, because the TE fragments that are very likely to be false positives have similar p-values for (Table 2.1). Because of this, the phylogeny should remain the ultimate argument for the validity of a real match. In the case of calpains, the phylogeny and the scenario of diversification by gene duplication proposed by Jékely and Friedrich (1999) remains to be revised for at least two reasons: a) new calpains have been characterized (e.g. CAPN11); b) it was suggested that the ancestor of CAPNs 1 and 2 was the source of the tandem duplications that occurred after the first chordate genome duplication, while our data indicate CAPN9 as a better candidate, due to its closer resemblance (i.e. does not have the exapted fragment) to the ancestor of animal calpains (Figure 2.8).

2.3.3 The Granzyme A (GZMA)

GZMA is another protein that contains a putatively exapted TE fragment. Unlike in the cases of PTPN1 and CAPN1, where the TE cassette is contained in one exon, RM identifies an L3-like fragment that spans across two exons (Figure 2.11).

Figure 2.11: Alignment of GzmA transcript (gi:184022) with L3 consensus sequence as reported by RM. Identical nucleotides and corresponding amino acids are shaded. Exon limits are indicated above the alignment. GzmA residues 77-80 form a β-strand which is represented by an arrow.

A situation like this could be very easily interpreted as a false positive match, as one would expect TEs to be contiguous segments or interrupted by younger TEs, but not by long introns (2175 nt in this cases). However, gain of an intron after the exaptation event could be an alternative explanation, as the gain of introns was documented in vertebrates (Rogozin*, et al.* 2003), but in none of the eight additional cases in which the TE match spans across multiple exons (Table 2.1), gene's phylogeny does support such events. A look at the multiple sequence alignment of granzymes (Figure 2.12) reveals yet another possible scenario for the presence of the L3-like fragment in the GZMA transcript. While the L3-like sequence in exon 2 appears highly conserved across all vertebrate granzymes and invertebrate trypsins, the sequence from exon 3 appears to have limited conservation even among vertebrate granzymes. This would suggest that only the fragment in exon 3 was actually exapted (green box in Figure 2.12) and the match in exon 2 represent a random one. This explanation alone can raise doubts over the validity

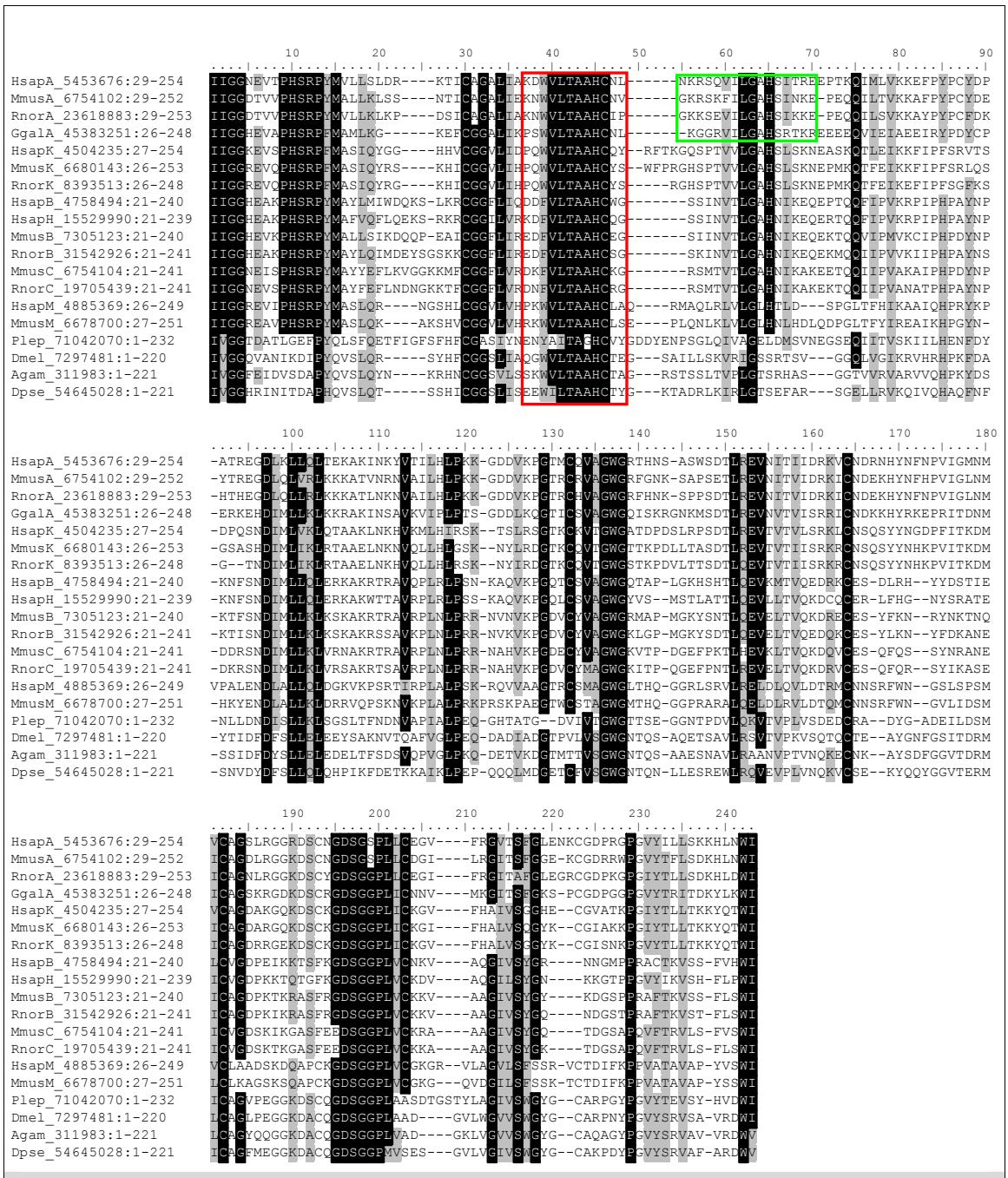

Figure 2.12: Multiple sequence alignment of vertebrate granzymes and invertebrate trypsins. Letters following species name indicate the granzyme type. The red box indicates the portion of the alignment that corresponds to the L3-like fragment in exon 2 of human GZMA. GZMM and *Dpse* trypsin have one more amino acid encoded by exon 2, not included in the box. *Dmel* and *Agam* trypsins are intronless, and gene structure is unknown for *Plep* trypsin. The green box corresponds to the L3-like fragment in exon 3, and indicates the putatively exapted TE-cassette encoded fragment. *Plep* - *Pontastacus leptodactylus*.

of the inferred exaptation event because it is harder to demonstrate that a shorter fragment was derived from a particular TE (there is a greater chance for a random match). Additionally, the L3-like fragment in exon 3 contains a core sequence (alignment coordinates 60-67 in Figure 2.12) conserved across all vertebrate granzymes, which could suggest that the segment might not be derived from a TE. There are, however, a number of facts that support the exaptation hypothesis. First of all, the L3 fragment had previous coding capacity, as it originates in ORF2p. In fact, it actually donates unchanged the core residues of what is now β4 (Hink-Schauer*, et al.* 2003) (Figures 2.11, 2.13), similarly to the case of PTPN1. Formation of a β-strand part of a more complex anti-parallel β-sheet would have probably been impossible with sequence originating in a non-coding TE region, such as the MIR sequence in CAPN1.

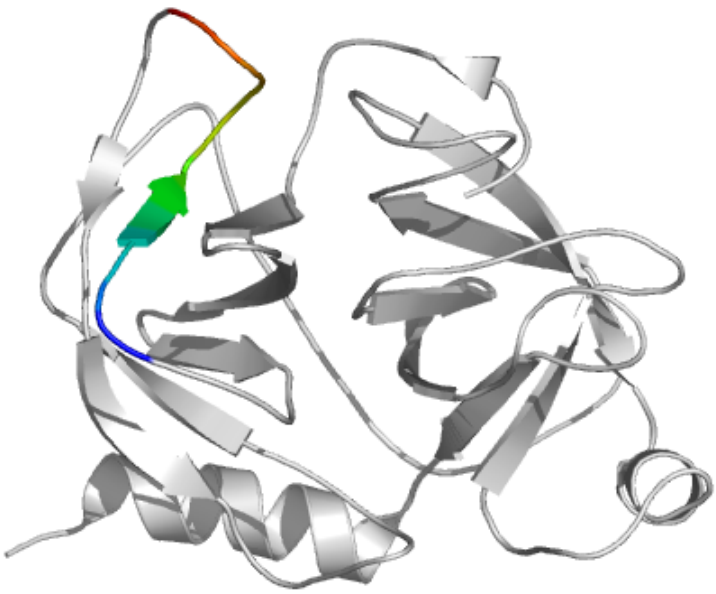

Figure 2.13: Three dimensional structure of human GZMA monomer (PDB ID: 1OP8). Residues 73-88 (coordinates 55-70 in Figure 2.12) that correspond to the L3-like fragment in exon 3 are colored in standard rainbow colors.

In addition, the L3-like fragment is located at the 5' end of exon 3, which could be explained by the activation of a cryptic acceptor site which led to exaptation of partial L3 sequence into GZMA. Once could extend the alignment into the 5' intronic region, but the alignment is very week and is not reported by RM, just as one would expect with old TE sequences from introns (Makalowski 2001).

Granzymes, similarly to PTPs and calpains, form a diverse family of proteins generated by multiple duplication events. Even though not completely documented, tandems of granzymes can be found on different chromosomes, such as GZMA and GZMK on chromosome 5, and GZMB and GZMH on chromosome 14 (Grossman*, et al.* 2003). The case of GZMA follows thus the pattern established by PTPs and calpains, where duplication events seem to have facilitated exaptation of TE sequences. It is interesting to observe that the 5' end of exon 3 is variable in length and is not conserved among different granzymes (Figure 2.12). This might be the result of modifications following gene duplication events, which might allow certain changes until a new function is gained by the newly arisen gene. The conservation pattern is not clear enough for one to precisely place the exaptation event on the phylogenetic tree of granzymes, but the common ancestor of vertebrate GZMA (Figure 2.14) appears as a good candidate. This is because GZMA forms a tandem with GZMK on chromosome 5, and we already have the example of the L3 and MIR exaptations in PTPN1 and CAPN1, respectively, which followed an intra-chromosomal duplication event. Fish seem to lack granzyme homologs, which places the exaptation event in the common ancestor of tetrapods, closer to the estimate of L3 activity time than the exaptation of the L3 fragment in PTPN1.

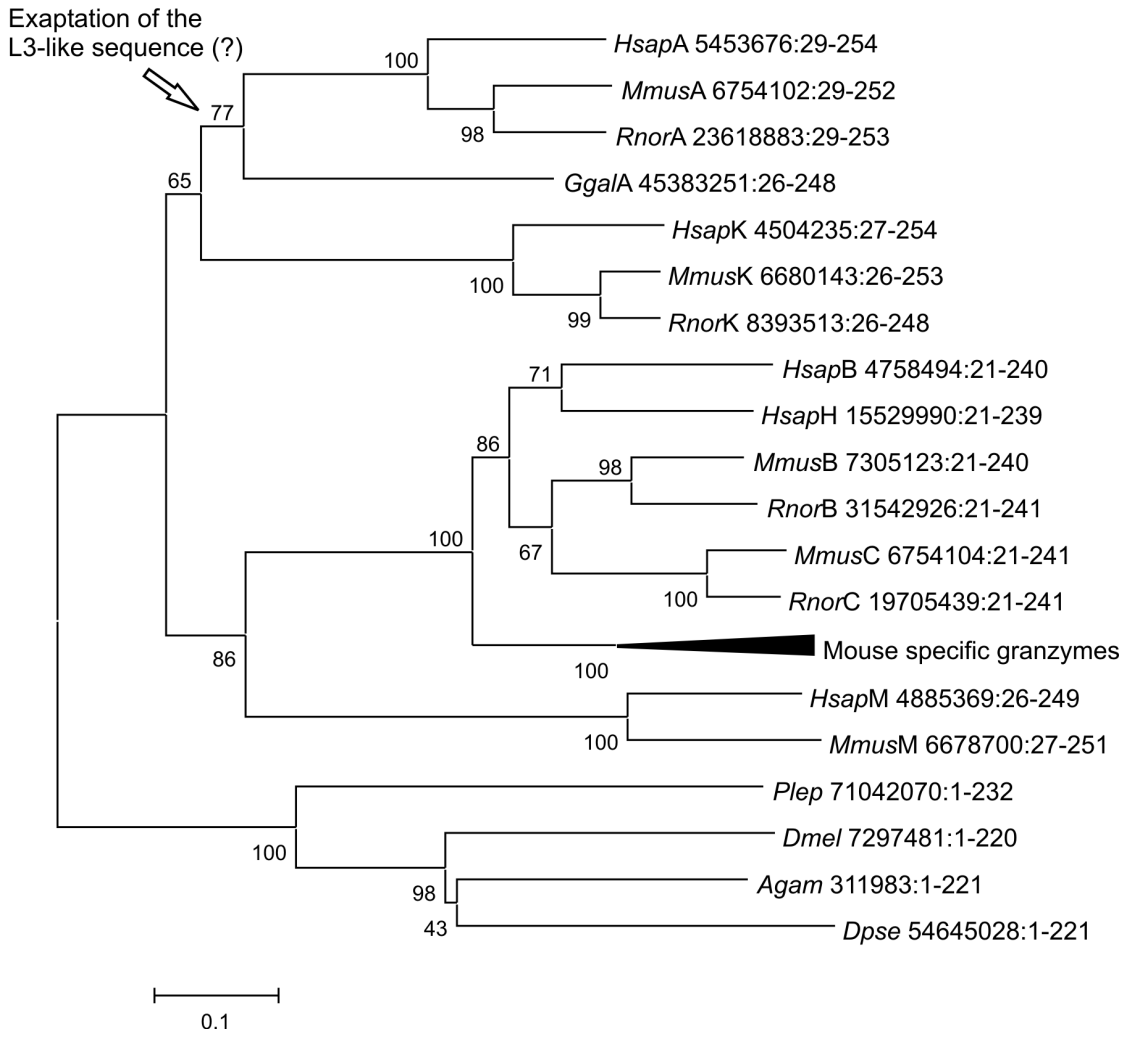

Figure 2.14: The Phylogeny of granzymes as reconstructed from the protein alignment of the trypsin-like domain. This represents a more accurate view of granzyme relationships than the phylogeny based on nucleotide alignments presented by Grossman *et al.* (2003). As no invertebrate granzymes have been found, closest invertebrate trypsins form an outgroup. Neighbor-joining method, complete deletion, 198 sites, Poisson corrected distance, and 1000 bootstrap replicates were used for construction of the phylogenetic tree. Mouse specific granzymes (D, E, F, G, N) were compressed to save space.

Regardless of whether we consider the GZMA a true case of exaptation or not, the importance of this case is also given by the fact that it illustrates a possible approach that

can be employed for detection of short TE fragments. In the case of GZMA, the L3-like sequence in exon 3 was fortuitously complemented by coding sequence in exon 2 and generated a significant match to the L3 consensus sequence. This complementation approach can be applied systematically to extend alignments into fragments that otherwise would produce scores below threshold because of their limited length. A similar approach has been successfully utilized in detecting "invisible" protein domains, with the main difference that complementation was done with protein sequences (van Rossum*, et al.* 2005). In our case, complementation needs to be done with nucleotide sequences for two main reasons. Many TEs do not have protein coding regions, and even if they do, frame shifts could disrupt any similarity between the ancestral protein and current sequence. Secondly, TE detection software has been developed to work with DNA, not protein sequences. Additionally, this approach can be applied to any genomic regions, without restriction to protein coding genes, for which multiple species alignments can be obtained. We are aware that aligning short and diverged nucleotide sequences is more problematic then aligning protein sequences, but the identity of the fragment discovered can always be confronted with the phylogenetic history of the gene/genomic region investigated as shown in this study.

2.4 Conclusions

The confirmation that TEs are present at the protein level is by no means a surprise, and they are certainly not the only category of DNA sequence to be exapted successfully into functional proteins. Hayashi et al. showed that any random sequence

could acquire biological functions if it had sufficient time to evolve (Hayashi*, et al.* 2003). It is, however, their prevalence and mobility within genomes that make TEs important players in molecular and genomic evolution.

2.4.1 Gene Duplications – Key Events that Favor Exaptation

One common feature of the PTP, calpain and granzyme protein families is that they were all diversified by multiple gene duplication events. A newly duplicated gene is likely to be free of functional constraints, and therefore can more easily accommodate major changes (Ohno 1970), such as the exaptation of TE sequences. If those genes are preserved and acquire new specific functions, the influence of TEs is then directly reflected through the function of the host protein. This aspect can be further investigated in other protein families known to have been diversified through extensive gene duplications.

2.4.2 Phylogenies – The Key for Validating Low Scoring TE Cassettes

Another common feature of the TE cassettes uncovered by this analysis is that they all have lengths, divergence from TE consensus and RM scores similar to those of cases considered to be false positives (Tables 2.1, 2.2). Therefore, an accurate distinction between random matches and real TE cassettes cannot be made based on any of those criteria. Moreover, not even the sequence randomization test can distinguish between the two because the P-values in all examples are small. Even so, they are overestimated,

because matches from any region of the same repeat type were considered, and even correcting for multiple tests would yield significant P-values. In these conditions, it is only the phylogenetic history of a gene that can confirm the validity of an RM match, as shown for the TE cassettes in PTPN1, CAPN and GZMA. Interestingly, all three cassettes are derived from old repeats, which is consistent with the idea that a nonaptation period is usually required for the fortuitous shaping of such elements before successful exaptation into the ORF of a gene. In contrast to the two sequences derived from L3 ORF2p, which are both part of anti-parallel β-sheets in PTPN1 and GZMA, the sequence derived from the noncoding tRNA-like MIR region forms a simple loop region in CAPN1. The fate and importance of the exapted TE fragment therefore appears to be determined by its original role in the parent TE.

However, we cannot completely exclude the hypothesis of sequence convergence for any of the TE cassettes. This is because all real TE cassettes are likely to be old, and are therefore short and highly diverged from their original sequence, which means that random matches that would resemble such TE fragments are likely to occur (21/24 putative TE cassettes seem to be random matches). However, the exaptation scenario should be favored when support from phylogenies exists, because the probability of having both a random match and phylogenetic support for the same protein fragment is lower than having a random match alone. Therefore, I urge scientists to treat low scoring RepeatMasker matches with special attention because some might prove to be real "treasures among the junk" (Nowak 1994).

2.4.3 TE Cassettes – Discrepancy Between the Frequency of Occurrence in Transcripts and Functional Proteins

One striking result of this analysis is that much fewer TE cassettes (~0.1%) can be found in functional proteins than one would expect (~4%) from the translation of TE-containing transcripts (Nekrutenko and Li 2001, Lorenc and Makalowski 2003). In contrast to this finding, most TE cassettes at the transcript level are derived from young TEs, and appear in a minor, alternatively spliced form of cognate mRNAs (Nekrutenko and Li 2001, Sorek*, et al.* 2002, Lorenc and Makalowski 2003). They can even persist as such over long evolutionary periods (Krull*, et al.* 2005), indicating that they might represent neither successful exaptations for protein coding purposes nor the intermediate stages of such events. They must have a different important role or otherwise they would be lost. At least one account provides a clue as to what that role might be. Oh *et al.* (2001) showed that coexpression of the a, b, and g subunits of wild-type human epithelial sodium channel (hENaC) with an Alu-containing splicing variant of the a subunit (haENaCCAlu) enhanced the expression of the amiloride-sensitive current in oocytes. The expression of TE-containing transcript variants, or even of pseudogenes, can thus regulate the expression or enhance the function of the functional protein coding form. The significant number of TE-containing transcripts might indicate that the role of TEs in regulation of gene expression and function is more important that it is currently acknowledged, and requires further insight.

2.4.4 The Number of Functional Proteins with TE Cassettes Is Currently Underestimated

A number of recent studies have uncovered isolated examples of functional proteins with TE fragments exapted into their coding region (Bejerano*, et al.* 2006, Cordaux*, et al.* 2006). In spite of this new evidence and the few real TE cassettes that can be found in functional proteins with the approach detailed herein, it is likely that the real number is underestimated. One reason for this is that transmembrane-, signal-, disordered- and low-complexity protein regions are significantly under-represented in the PDB collection (Peng*, et al.* 2004), because of the way targets are selected for structural genomics (Brenner 2000). We can only hope that further studies will characterize more proteins from the under-represented classes. A second reason is that all TE-cassettes that I found are derived from old TEs (Table 2.2), which might cause the exaptation events to be obscured by long evolutionary periods. In addition, old TEs are usually difficult to identify because of their highly diverged and fragmented sequence. Similarity searching techniques currently employed for finding TEs are not optimized for these types of sequences; therefore, I would encourage the scientific community to implement better techniques for detecting fragments of older TEs. For example, the use of position weight matrices instead of consensus sequences for finding diverged MIR copies seemed to be a promising approach (Chalei and Korotkov 2001) and could be applied to other TEs. Despite these reasons, the real proportion of TE-containing proteins is probably closer to our estimate of ~0.1% than to previous estimates of ~4% (Nekrutenko and Li 2001, Lorenc and Makalowski 2003) because we do not expect to find TE cassettes of young elements in functional proteins.

2.4.5 Young TEs: Subject to Future Exaptation Events

An important conclusion of this study is that functional proteins are unlikely to contain TE cassettes derived from young TEs, such as Alu and L1s. This is in contrast to previous reports (Nekrutenko and Li 2001, Lorenc and Makalowski 2003), which estimated that they represent up to 60% of the human TE cassettes in ORFs. Even if that might be true at the transcript level, it seems unlikely that young TEs could be found in functional proteins because long evolutionary periods are needed for successful exaptation events. For example, Alu elements, which are found only in primate genomes, did not have enough time to evolve and adapt to new coding functions, but they seem to be currently undergoing that process (Krull, *et al.* 2005). As a result, they often cause problems when inserted into protein coding regions (Deininger and Batzer 1999). By contrast, there are examples of young repeats that contributed to the human proteome, but did not undergo exaptation. Elements such as the human endogenous retroviruses HERV-FRD were co-opted by the human genome, and their protein product has a function similar to that of the original retrovirus gene (Renard, *et al.* 2005). Nonetheless, we should not be surprised if exaptation of currently young TEs will eventually yield functional proteins – we just need to give nature enough time.

This chapter was published in Trends in Genetics, Vol. 22, Gotea, V., Makalowski, W., Do transposable elements really contribute to proteomes?, 260-267,

Chapter 3

Alu Retrotransposons in Protein Coding Sequences

3.1 Introduction

Alu elements represent the most successful class of SINE retrotransposons in the primate lineage, representing the most numerous class of TEs in the human genome, in which more than one million copies occupy ~10% of its sequence. Their name is given by the fact that they contain a number of sites recognized by the *Alu*I restriction enzyme (Houck, *et al.* 1979). Alu elements have a dimeric composition, where both the left and right monomers originated in the 7SL RNA gene (Sinnett, *et al.* 1991, Kriegs, *et al.* 2007), and therefore they do not have protein coding capacity. The two monomers are connected by an A rich linker, so that the Alu consensus is 282 nt long (Price, *et al.* 2004), to which a polyA tail is usually added. Despite their lack of protein coding capacity, Alu elements have multiplied tremendously in the primate lineage, by a mechanism that is believed to involve the retrotransposition machinery of L1 LINE retrotransposons. In a similar manner, the MIR class of SINE retrotransposons is believed to have used the retrotransposition machinery of L2 LINE elements.

Alu elements have inserted throughout the human genome, having no apparent insertional bias (Cordaux, *et al.* 2006). However, they appear to be enriched in gene rich regions as compared to other classes of TEs, such as LTR or LINE elements, which is thought to be due to their lower content in putative regulatory signals leading to different

retention rates in those regions (Thornburg, *et al.* 2006). When inserted in introns, fragments of Alu elements are often included in alternatively spliced transcripts (Sorek, *et al.* 2002) due to their sequence predisposition to provide splicing signals, especially when inserted in reverse orientation relative to the gene orientation (Makalowski, *et al.* 1994). These transcripts are considered by several authors to automatically produce functional protein products (Deragon and Capy 2000, Li, *et al.* 2001, Nekrutenko and Li 2001, Kriegs, *et al.* 2007), despite the lack of experimental evidence for any of those. Even though a few studies have suggested that Alu-cassettes can be part of functional proteins (Gerber, *et al.* 1997, Hoenicka, *et al.* 2002), solid evidence for this claim is still elusive. In fact, whether Alu elements contributed to the evolution of primate lineage is unknown, as no single aspect of the Alu impact on primate genomes has been shown to be a major factor in the evolution of primates or to contribute to the phenomenal success of Alu amplification.

In an effort to add to the growing body of evidence supporting different aspects of Alu impact on primate genomes, I used several computational methods to try to clarify the contribution of Alu elements to proteomes. I used protein homology modeling to infer if transcripts containing Alu cassettes can produce functional proteins using the structure of the protein translated from the transcript variant that misses the Alu cassette as a model. I also tested whether Alu cassette exons are subject to specific selection forces that would indicate a significant role at the protein level. At the same time, I investigated new ways in which Alu elements can contribute to enriching the proteome repertoire of the cell and enhancing its functionality.

3.2 Materials and Methods

3.2.1 Inferring Functionality of Alu-Cassette Containing Transcripts from Protein Homology Modeling

Alu elements are present in the introns of many genes, and often they are included in mature mRNA molecules as alternatively spliced exons. Even though in many such cases they are part of the CDS, no strong evidence supports the claim that the corresponding protein can be functional. One way to address this problem is to infer the functionality of the *Alu*ternative (Kreahling and Graveley 2004) variants using the structural information of the non-Alu variant, when this information exists. For this purpose, I used the ScrapYard database (see Chapter 5), and later an updated dataset of human transcripts (see Chapter 3.2.2) to find Alu-containing transcripts. A tBLASTn search was then carried with the non-redundant set of human proteins from PDB against the Alu-containing transcripts. With the help of custom Perl scripts cases where the Alu cassette was included in the alignment were further subjected to homology modeling, because homology modeling cannot be carried for segments having no homolog with structural information. For the homology modeling step I used Genemine 3.5.2 (http://www.bioinformatics.ucla.edu/genemine), which uses a protein alignment to infer the structure of the protein of choice based on the structure of the aligned homolog (in our case this can be just an alternatively spliced variant). In most cases, the alignment provided by tBLASTn was used directly, but manual adjustments were made where necessary. Cartoon representations of the structural models were made with PyMOL (http://pymol.sourceforge.net). From the model produced by Genemine, inferences

regarding the impact of the Alu cassette on the stability of the protein structure or known active site with consequences for the protein functionality were made.

3.2.2 Inferring Selection Acting on *Alu*ternative Exons from Human-Macaque Comparisons

The recent sequencing of the macaque genome offered new possibilities for studying selection acting on primate specific sequences, such as Alu elements. Human and macaque diverged from their most recent common ancestor about 25 million years ago (Mya) (Kumar and Hedges 1998), more than four fold the time since the human-chimp split, which is estimated at 6 Mya (Chen and Li 2001). The more distant divergence between human and macaque (in an average gene, 12 non-synonymous and 22 synonymous changes separate the two species) helps in clarifying the signature of natural selection acting on protein coding genes, which would have been almost impossible to distinguish at the level of divergence between human and chimp, with less than 3 non-synonymous and 5 synonymous changes for an average gene (Gibbs*, et al.* 2007). As of September 17, 2005, I downloaded 52,140 human transcripts from GenBank that were annotated as having a "complete CDS", as well as 29,459 RefSeq human transcripts. Using patdb to eliminate redundancy from these two datasets resulted in a set of 66,812 non-redundant transcripts that were further scanned with RepeatMasker (http://www.repeatmasker.org) for occurrence of Alu elements. Among those, 7,262 transcripts contained at least one Alu fragment, 297 of which contained at least one Alu fragment fully contained within the CDS. Using the UCSC Genome Browser (Kent*, et al.* 2002) I mapped these transcripts on the human genome and determined the transcripts

where the exons containing the Alu fragments contain no other sequence. This is to eliminate unknown influences that other categories of DNA might have on the selection acting on those exons. A conservative approach is to consider only cases where the Alu element provided both the acceptor and the donor splice sites, condition satisfied by 88 exons. For these cases, I further used the Galaxy framework (Giardine*, et al.* 2005) to extract human-macaque alignments corresponding to the coordinates of the 88 exons found to contain only Alu sequence. Furthermore, I eliminated from further analyses exons for which macaque is missing a part or the entire exon (16 cases), for which frame-shift causing indels exist between the two species (11 cases), for which the splice sites are not conserved in macaque (5 cases), or if a premature termination codon exists in macaque (9 cases). For the remaining 47 exons, I computed the number of synonymous and non-synonymous changes between the two species using the modified Nei-Gojobori method (Nei and Gojobori 1986). The significance of selection suggested by the dN/dS measure was tested with Fisher's exact test (Fisher 1922).

3.2.3 Investigating the Impact of Alu Elements on the Signaling Molecules

Alternatively spliced exons containing Alu fragments can be found fully contained within the CDS, as well as at the beginning or at the end of it. In the case of the former, a successful translation of the transcript would produce a protein with a different N-terminus. This offers the possibility that a signal peptide is introduced at or removed from the N-terminus of the protein. The implications can be major, because new signaling molecules can be created, as well as proteins can be arrested in the cytoplasm

instead of being exported. Using the dataset described in Chapter 3.2.2, I searched for cases where the START codon is annotated in an Alu element, by intersecting the CDS and RM annotations with the help of custom Perl scripts. I then compared these cases with the major alternative form of the gene in terms of the presence or absence of a signal peptide at the N-terminus of the protein product using SignalP 3.0 (Bendtsen*, et al.* 2004).

3.2.4 Investigating the Contribution of Alu Elements to the Human Selenoproteome

Another interesting scenario of Alu contribution to genomic novelty is the possibility that in special situations they can promote the incorporation of selenocysteine (Sec) residues in the polypetide chain. In this way, they could contribute to enlarging the repertoire of the human selenoproteins, with direct positive effects in the cell defense against oxidative stress, for example. They could do so simply by providing an in-frame UGA termination codon, which could be translated into a Sec residue instead of a termination signal. The recoding of the UGA termination signal into a Sec residue is a feature of the metazoan, archae- and eubacterial translational machineries which is facilitated by the presence of a few special features: a unique Sec-tRNA^{Sec}, a unique elongation factor to deliver the tRNA, and a cis-acting element in the mRNA that forms a stem-loop structure, refer to as the **SE**leno**C**ysteine **I**nsertion **S**equence (SECIS) (Berry*, et al.* 1993, Copeland 2003). It is interesting to note that plants and fungi do not have the mechanism required for Sec incorporation (Copeland 2005). The relative position and structure of the SECIS element to the UGA codon differ considerably among metazoa,

archaea, and eubacteria. Another difference is that while bacteria require only an elongation factor (SelB) which binds to both the SECIS element and the tRNA. In the case of metazoa, Sec incorporation requires both an elongation factor (eEFSec) and a specific SECIS binding protein (SBP2) (Copeland 2003).

The number of selenoproteins in human is relatively small, only 25 being known until present (Kryukov*, et al.* 2003). The restrictive factor seems to be the number of mRNA with the required UGA and SECIS element, since all the other features (eEFsec, SBP2, Sec-tRNASec) exist and function within humans. The number of mRNA could be potentially increased during evolution if an mRNA featuring a SECIS-like structure would fortuitously acquire an in-frame UGA codon. The expansion of Alu elements during primate evolution prompts to think of this possibility. They are known to be included as alternatively spliced exons in many genes, and often times they provide an in-frame premature termination codon that determines a disease phenotype (Deininger and Batzer 1999, Sorek*, et al.* 2002). If an *Alu*ternative exon happens to be included in a SECIS-containing mRNA, it can very well be recoded into Sec, thus potentially contributing to the enrichment of the selenoproteome. To investigate this interesting evolutionary scenario, I used the same dataset described in Chapter 3.2.2 to search for transcripts where UGA termination codons are annotated within Alu fragments. I then scanned the downstream sequence for the presence of another in-frame termination codon which would provide the end of the CDS, and then used the SECISearch 2.19 (Kryukov*, et al.* 2003) to scan for the presence of a SECIS element downstream of the Sec residue. I also used the same protocol to search for potential Sec containing transcripts in the mouse and rat transcriptomes from the ScrapYard database (see Chapter 5).

3.3 Results and Discussion

3.3.1 Functionality of Alu-Cassette Containing Proteins Inferred from Homology Modeling

The search for alternatively spliced variants in PDB of Alu containing transcripts yielded no surprising results, namely no Alu fragment had a direct hit to a protein from PDB. Alu cassettes were included in alignments rather due to the perfectly matching flanking regions, and they appeared in many cases as in-frame additions to the polypeptide chain, as shown in Figure 3.1 for the human survivin. This is in agreement with previous studies (Pavlicek*, et al.* 2002), and attests to the lack of support of the functionality of Alu-containing proteins.

```
pdb|1E31:A   1  MGAPTLPPAWQPFLKDHRISTFKNWPFLEGCACTPERMAEAGFIHCPTENEPDLAQCFFC 60
                MGAPTLPPAWQPFLKDHRISTFKNWPFLEGCACTPERMAEAGFIHCPTENEPDLAQCFFC
gi|7416052  27 MGAPTLPPAWQPFLKDHRISTFKNWPFLEGCACTPERMAEAGFIHCPTENEPDLAQCFFC 206

pdb|1E31:A  61 FKELEGWEPDDDPI----------------------EEHKKHSSGCAFLSVKKQFEELT 97
                FKELEGWEPDDDPI                      EEHKKHSSGCAFLSVKKQFEELT
gi|7416052 207 FKELEGWEPDDDPIGPGTVAYACNTSTLGGRGGRITREEHKKHSSGCAFLSVKKQFEELT 386

pdb|1E31:A  98 LGEFLKLDRERAKNKIAKETNNKKKEFEETAKKVRRAIEQLAAMD  142
                LGEFLKLDRERAKNKIAKETNNKKKEFEETAKKVRRAIEQLAAMD
gi|7416052 387 LGEFLKLDRERAKNKIAKETNNKKKEFEETAKKVRRAIEQLAAMD  521
```

Figure 3.1: Protein alignment of the human survivin (the chain A of the 1E31 PDB structure) with the translated sequence of its Alu containing splicing variant, survivin-2B (gi:7416052). The Alu cassette appears as a 23 amino acid addition to the survivin sequence, indicated by a gap in the normal variant.

For a small number of Alu-containing transcripts, including survivin-beta, the structure of the non-Alu variant was found in PDB (Table **3.1**). For these cases, inferences were made regarding the functionality of the protein product containing the Alu-encoded cassette.

Table 3.1: Genes with alternatively spliced transcripts that contain an Alu cassettes in their CDS. The 3D structure of the normal variant was determined for the protein products of these genes, and it was used for inferring the functionality of the Alu-containing variant.

Gene symbol	Transcript with Alu cassette (gi number)	Chromosomal location	Length of the Alu-encoded cassette (aa)	PDB ID of the protein without the Alu cassette
BIRC5	7416052	17q25	23	1E31
CHEK2	54112406	22q12.1	43	1GXC
PPIL3	19557632	2q33.1	36	1XYH
SULT1C2	8117858	2q11.1-q11.2	32	1ZHE
ADARB1	75709170	21q22.3	40	1ZY7

Survivin, the short name of the baculoviral IAP repeat-containing 5 protein (BIRC5), is a member of the inhibitor of apoptosis (IAP) gene family which contains proteins that inhibit apoptotic cell death. IAP proteins usually contain multiple baculovirus IAP repeat (BIR) domains, but survivin contains only one such domain (Ambrosini*, et al.* 1997, Chantalat*, et al.* 2000). Survivin has three splicing variants, one of which misses the third exon, called survivin-ΔEx3, and one which has an additional exon, called survivin-2B or survivin-beta (Mahotka*, et al.* 1999). The additional 69-nt exon in survivin-2B is provided by an AluY element, inserted in sense orientation relative to the gene and which provides both the acceptor and donor splice sites. At the protein level, this translates into a 23 amino acid fragment which is located within the BIR domain of survivin. Using the structure of the major survivin form determined by Chantalat *et al.* (2000) and the alignment of its sequence with survivin-2B, it can be determined what structural unit of the protein would be interrupted by the inclusion of the Alu-encoded cassette (Figure 3.2). It becomes clear that the 23 extra amino acids will disrupt the α-helix that hosts His^{77}, which is one of the four Zn-binding residues, and thus

will affect the stability of the entire BIR domain. Cys^{84}, which also binds the Zn atom is also likely to be displaced because the folding of the entire domain fragment downstream of the Alu-encoded cassette is likely to be affected. If the fragment preceding the cassette folds in the same way as the major form, or if the folding of survivin-2B will yield a different functional structure is unclear, and needs to be determined experimentally.

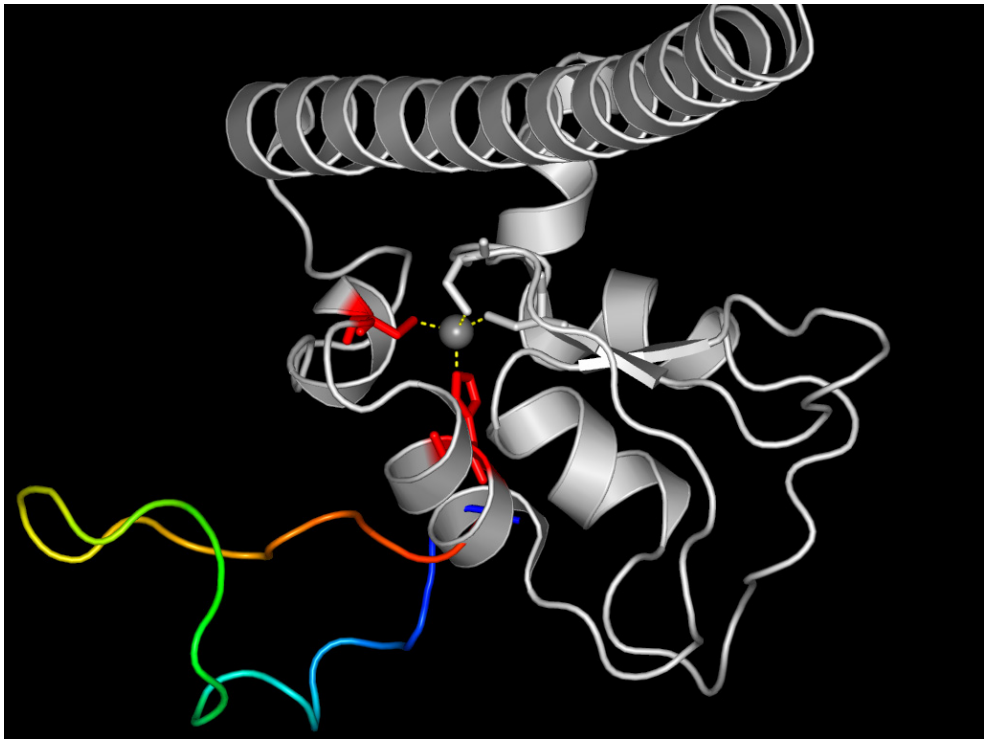

Figure 3.2: Three dimensional structure of human survivin (PDB ID: 1E31) in white, with the position of the alternatively spliced Alu-encoded cassette predicted by Genemine (shown in standard rainbow colors). Side chains of His^{77} and Cys^{84} are shown in red.

The consequence of this insertion seems likely to be the destabilization of the survivin BIR domain, and thus a probable loss of function. In fact, this observation agrees well with previous studies which indicate that survivin-2B is not only differentially

expressed in different tissues (Mahotka*, et al.* 2002, Li*, et al.* 2007), but also has antagonistic effects as compared to survivin and survivin-ΔEx3 variants (Mahotka*, et al.* 1999). The antagonistic effects could be very well explained by the simple loss of function of survivin when is spliced with the *Alu*ternative exon, or by destabilizing interactions with the normal variants at the level of transcripts, but at least it is clear that survivin-2B is not able to perform the function of the other two known survivin variants.

CHK2 checkpoint homolog (*S. pombe*) (CHEK2) is a cell cycle checkpoint regulator and putative tumor suppressor. It contains a forkhead-associated (FHA) protein interaction domain essential for activation in response to DNA damage, and in its activated form phosphorilates BRCA1, CDC25 family of phosphatases, and the p53 tumor suppressor, as reviewed by Li*, et al.* (2002). CHEK2 shows extensive alternative splicing in tumors, where about 90 variants were discovered (Staalesen*, et al.* 2004). One of these variants contains an 129-nt Alu cassette provided by an anti-sense FLAM_C element which provides the acceptor splice site (15 nt of the exon and the donor splice site are provided by adjacent anonymous sequence). As in the case of survivin, the structural information of the major form (Li*, et al.* 2002) helped in understanding what impact this cassette can have. The Alu-cassette would disrupt (Figure 3.2) the second β-strand in the protein (Li*, et al.* 2002), but which is not part of the FHA domain as indicated by a Pfam search. It can be speculated that such an insert would allow the functional FHA domain to fold and function properly. However, none of the known FHA domain structures present such a long insertion between β2 and β3, thus the possibility of improper folding still exists. It can be argued that the most likely possibility is that the

Alu-containing variant of CHEK2 is non functional or functions aberrantly, because the splicing variant was found in tumors (Li, *et al.* 2002), but this requires further experimental evidence.

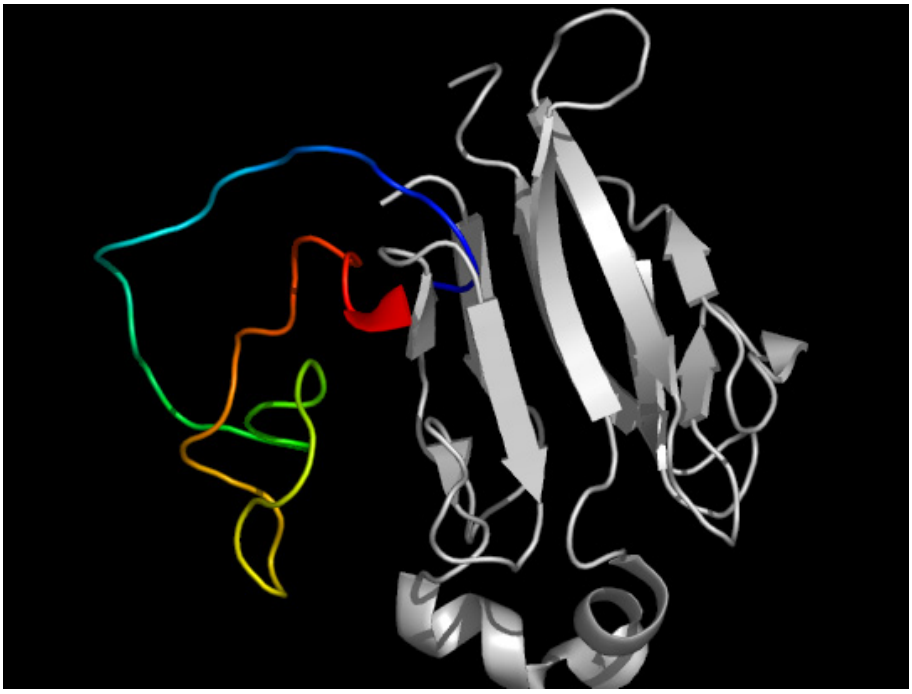

Figure 3.3: Three dimensional structure of the human CHEK2 (in white, PDB ID: 1GXC) with the alternatively spliced Alu cassette shown in standard rainbow colors.

Peptidylprolyl isomerase (cyclophilin)-like 3 (PPIL3) is a member of the cyclophilin family of proteins, which catalyze the cis-trans isomerization of peptidylprolyl imide bonds in oligopeptides. They have been proposed to act either as catalysts or as molecular chaperones in protein-folding events (Gothel and Marahiel 1999). As in the previous cases, PPIL3 has several splicing isoforms, and one of them, PPIL3a (Zhou, *et al.* 2001), contains an Alu cassette derived from a reverse oriented (relative to gene orientation) AluSx element which provides both the donor and acceptor

splice sites. The 106-nt long cassette replaces a slightly shorter (94 nt) exon, thus compensating for the frame shift that would be otherwise introduced. Consequently, the model produced does not have an extra loop, but contains the Alu-encoded cassette nicely within the structure. However, the major form normally interacts with several water molecules in the catalytic process, but the Phe^{56} in PPIL3a is likely to interfere with the water molecules, thus with the catalysis itself. In fact, it appears that PPIL3a is insoluble (Huang, pers. comm.), raising the question of whether it can actually perform its function.

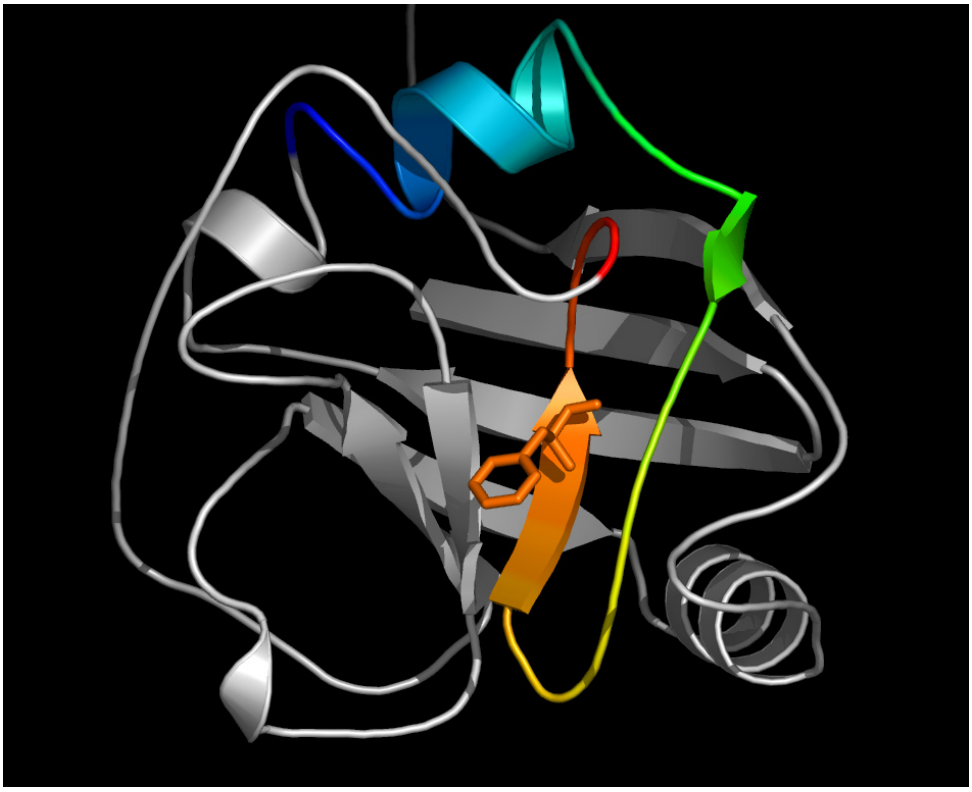

Figure 3.4: A structural model of PPIL3a variant of PPIL3 based on PDB ID 1XYH. The Alu-encoded cassette is shown in standard rainbow colors, and Phe^{56} is represented with its side chains.

Sulfotransferase family, cytosolic, 1C, member 2 (SULT1C2) is a member of Phase II enzymes of chemical defense, responsible for the biotransformation of many drugs, neurotransmitters, and hormones (Freimuth, *et al.* 2000). There are three SULT families (SULT1, SULT2, SULT4), SULT1 having four subfamilies (A, B, C, E) and eight members, including SULT1C2 (Blanchard, *et al.* 2004). Just as the previously described proteins, SULT1C2 has several splicing variants, one of them containing an Alu cassette. What is interesting about this variant is that the splicing of the Alu cassette is coupled with the alternative splicing of the downstream exon which compensates for a frame shift, as shown in Figure 3.5:

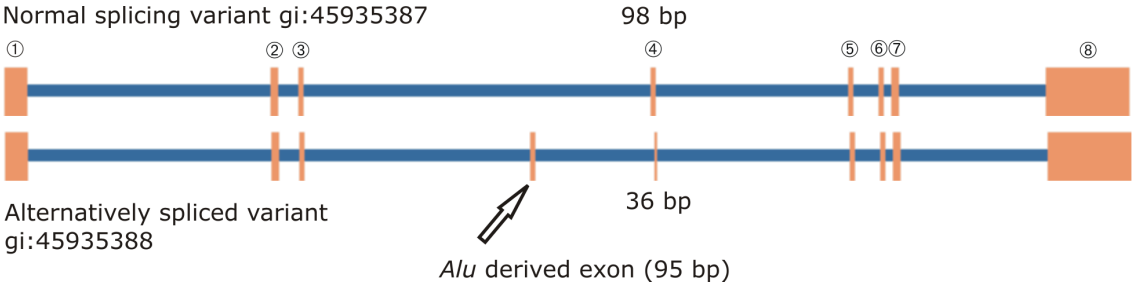

Figure 3.5: Two splicing variants of SULT1C2. The Alu-containing variant also features alternative splicing of the downstream exon, and compensates for a possible frame shift.

This results in a situation similar to PPIL3a, where the Alu cassette replaces a shorter fragment, which leads to a model (based on the structure of SULT1C1) where the Alu-encoded fragment is nicely integrated into the structure of the domain (Figure 3.6). Unfortunately, it is not clear what the active sites are despite having a 3D structure (Dombrovski, *et al.* 2006), and therefore inferences about the functionality of this variant are hard to be made.

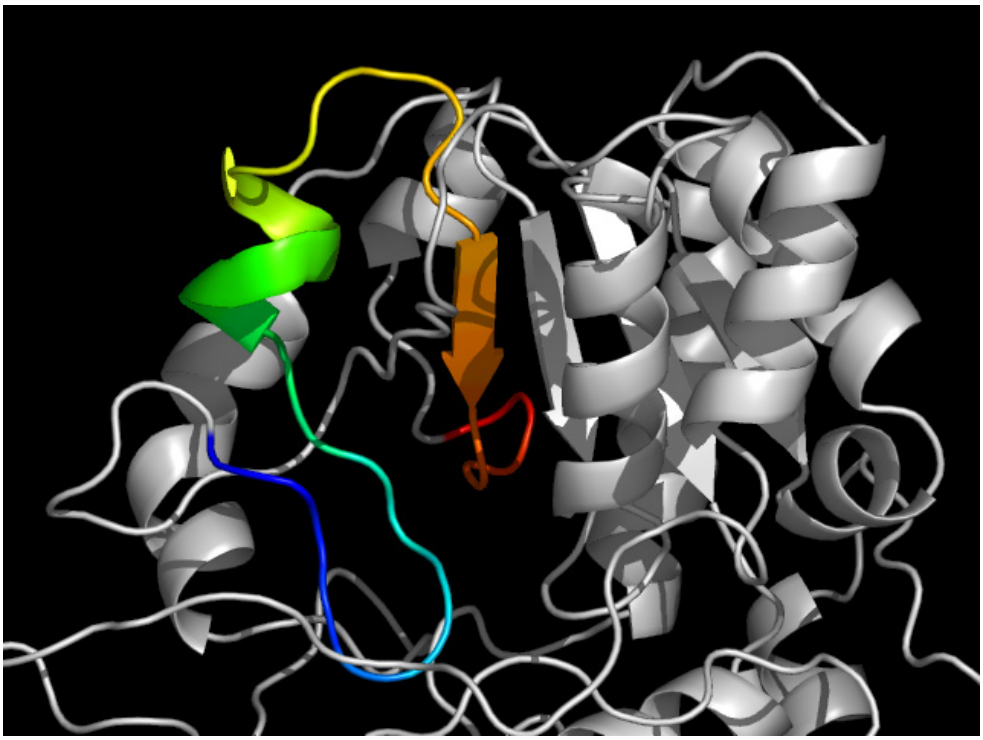

Figure 3.6: Structural model of the SULT1C2 Alu-containing variant based on the structure of SULT1C1 (PDB ID: 1ZHE).

Adenosine deaminase, RNA-specific, B1 (RED1 homolog rat) (ADARB1) is an enzyme responsible for pre-mRNA editing of the glutamate receptor subunit B by site-specific deamination of adenosines. It is also called DRADA2, and its four splicing isoforms were characterized for their activity *in vitro* by Lai *et al.* (1997). DRADA2b is the longest isoform, which also contains an Alu cassette. The Alu cassette is shared by the DRADA2c isoform, but which has a shorter C-terminus as compared to DRADA2b. From the model based on PDB ID: 1ZY7 (Figure 3.7), one could infer that the 40 amino acids introduced by the Alu cassette do not interfere with the catalytic Zn center which is located in a deep pocket in the enzyme surface surrounded by positive electrostatic potential that likely serves as the dsRNA binding site as shown by Figure 2b in Macbeth

et al. (2005). In fact, the evidence from *in vitro* experiments suggests that DRADA2b has reactivity comparable with DRADA2a isoform (Lai, *et al.* 1997), which lacks the Alu cassette, so it seems that the Alu cassette might be tolerated in the protein as long as it does not interfere with its function. To what degree the Alu cassette interferes with the folding of the protein can only be addressed experimentally, so the complete assessment of the influence of the Alu cassette cannot be done computationally.

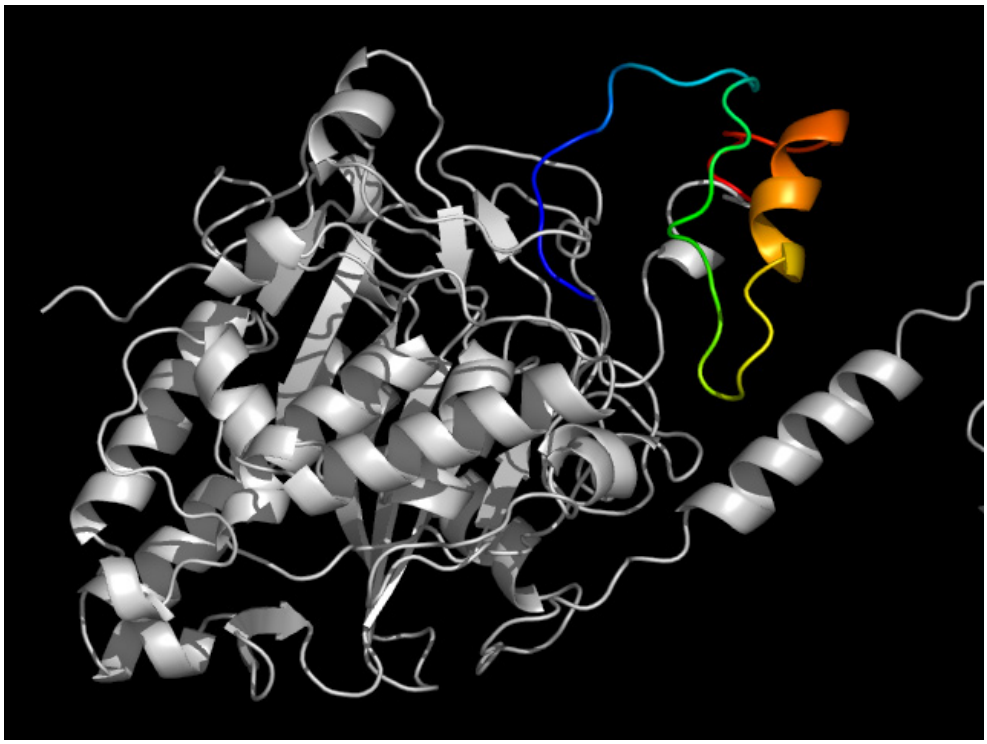

Figure 3.7: Model of the DRADA2b isoform based on PDB ID 1ZY7. The Alu-encoded cassette is shown in standard rainbow colors.

Aside from the five cases presented above, there are a few others where the structure of a more distant homolog exists (Table 3.2). In such cases it is much harder to infer the functionality of the variants with the Alu cassettes, because the locations of the active sites, and sometimes the active sites themselves, are unknown.

Table 3.2: Genes with Alu-containing splicing variants, but for which only structures of more distant homologs exist. Note that for computing the alignment lengths and identities, the Alu cassettes and other alternatively spliced exons (in the case of 1MX1) were excluded from the alignment.

Gene symbol	Transcript with Alu cassette (gi number)	Chromosomal location	Length of the Alu-encoded cassette (aa)	PDB ID of the homolog structure	Alignment length/identity (aa / %)
FLJ37464	40255185	16q22.1	40	1MX1	395 / 45.6
MKNK1	34147650	1p33	41	1NXK	303 / 32.8
PKP2	52630430	12p11	44	1XM9	459 / 42.3
UBE2J2	37577125	1p36.33	16	1JAS	119 / 32.8

In the case of PKP2, for example, the Alu-encoded fragment is located in a loop rich region, which might suggest that if no serious interference with the folding of the protein exists, then the cassette might be tolerated in the structure (Figure 3.8).

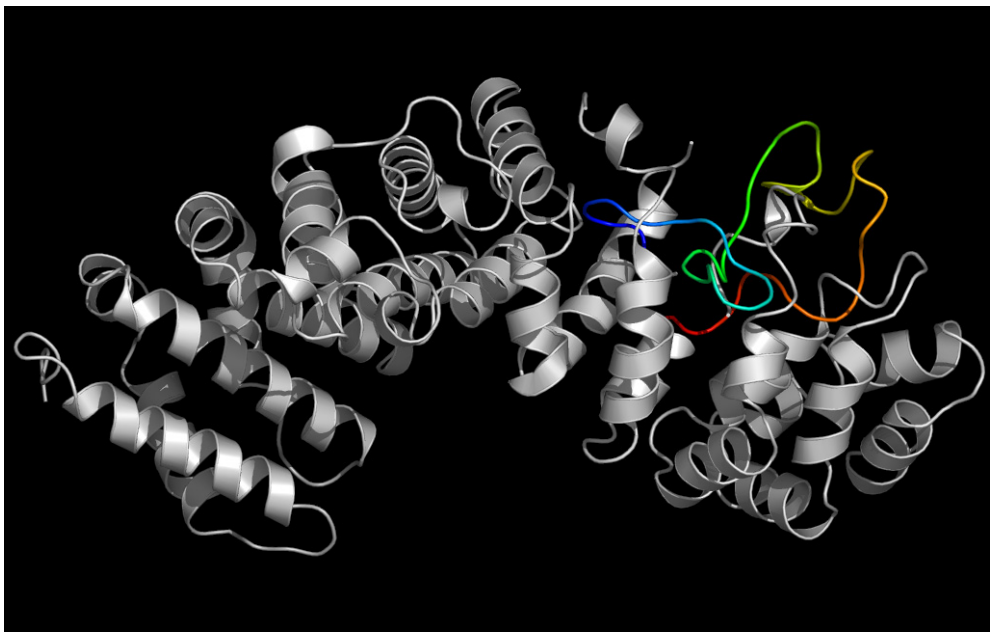

Figure 3.8: Structural model of the Alu-containing isoform of PKP2 based on PDB ID 1XM9.

In conclusion, there is no general pattern that applies to the functionality of the Alu cassettes at the protein level. The impact of the insertion depends on its position relative to functionally important sites. In some cases it can be inferred that an Alu cassette might disrupt the function of the host protein, which might indicate that those cassettes might have a regulatory function at the transcript level, such as regulating the protein level post-transcriptionally by being targets for microRNAs (Smalheiser and Torvik 2006). Even though in other cases the Alu-encoded cassettes might be tolerated, such as in the case of DRADA2b, strong evidence suggesting that Alu elements can contribute to proteomes is still missing. This agrees well with the fact that Alu elements do not have coding capacity, and thus for exaptation to create phenotypes with increased fitness a longer evolutionary time might be required. Alu elements have appeared recently during the evolution of primates, and they might just be "en route to protein-coding function" (Krull*, et al.* 2005).

3.3.2 Selection Acting on *Alu*ternative Exons

For determining the signature of natural selection acting on Alu cassettes, a dataset of 88 *Alu*ternative exons was compiled in which the Alu element provided both the donor and acceptor splice sites, so that the possible influence of other types of sequences would be eliminated. Most of these exons originated in Alu elements inserted in reverse orientation relative to the gene, which agrees with the fact the Alu sequence has a predisposition of providing splicing sites on its reverse complement (Makalowski*, et al.* 1994). In Figure 3.9, the usage of the nucleotides mapped to the Alu consensus

sequence (Price, *et al.* 2004) is shown, and it is obvious that not only most of the exons were generated by reverse oriented Alu elements, but most of them also originate from the right Alu monomer.

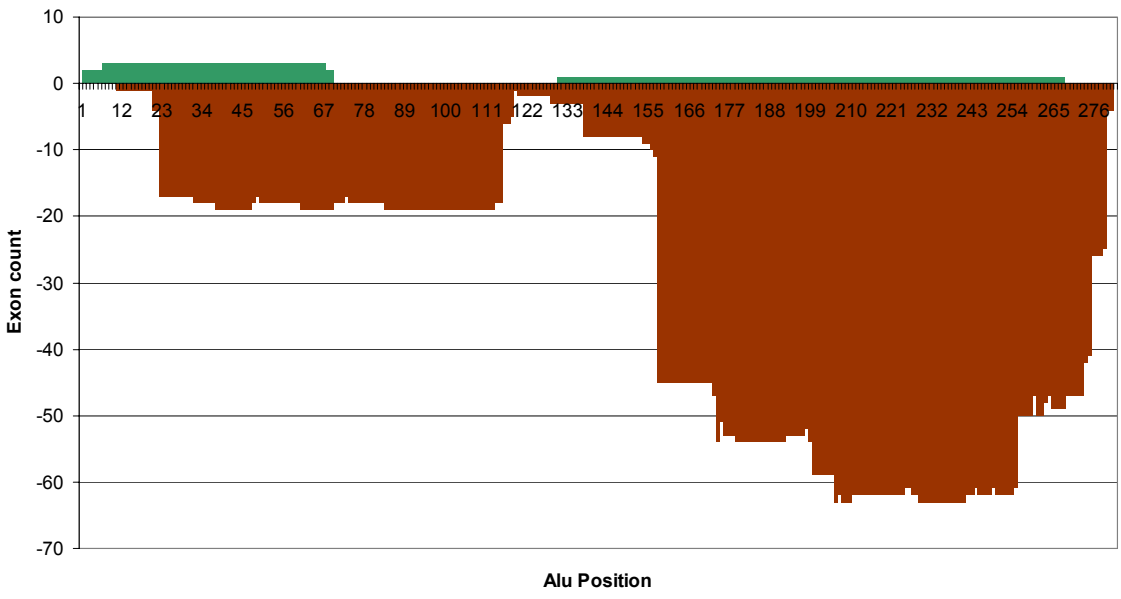

Figure 3.9: Frequency of usage for each nucleotide in the Alu consensus sequence in 88 human exons, where both the donor and acceptor splice sites are provided by an Alu element.

Out of these 88 exons, only 47 were kept for further analysis, due to lack of conservation for the rest of them in macaque: in 16 cases a part or the entire exon is missing in macaque, in 11 cases frame-shift causing indels exist between the two species, in 5 cases the splice sites are not conserved in macaque, and in 9 cases a premature termination codon exists in macaque. In seven cases, the comparison revealed no synonymous mutations (dS=0, dN/dS could not be thus computed), which could be an indication of positive selection. The remaining 40 exons showed great variation in the dN/dS, as shown in Figure 3.10.

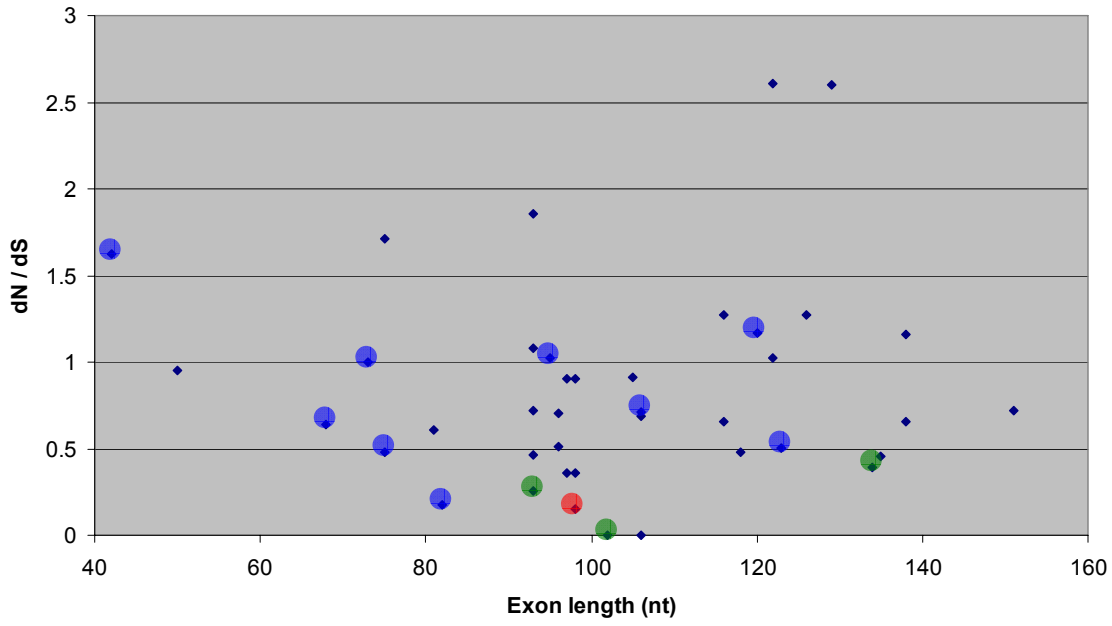

Figure 3.10: Selection acting on 40 *Alu*ternative exons, as indicated by dN/dS. Blue circles denote exons inserted in the middle of a known functional domain, green and red circles indicate near significant (*P*-value < 0.1), and significant (*P*-value < 0.05), respectively, purifying selection as indicated by Fisher's exact test.

It is interesting to note that, with one exception, none of the *Alu*ternative exons shows an excess of synonymous or non-synonymous substitutions. However, the Alu cassettes in the splicing variants of RGR, GYG2, and MLLT10 genes appear to be subject to near significant (*P*-value = 0.074, 0.068, 0.064, respectively) purifying selection as indicated by Fisher's exact test (green circles in Figure 3.10). Also, the Alu cassette in the variant of JARID1B gene appears to be subject to positive selection, as 8 non-synonymous and none synonymous mutations could be detected between human and macaque over the 108-nt long exon (*P*-value = 0.084; not represented in Figure 3.10 because dS=0).

Interestingly, Fisher's exact test indicates that the 98-nt AluSx cassette in BAXε (red circle in Figure 3.10) splice variant of BAX (Shi*, et al.* 1999) appears to be subject to significant purifying selection (2 non-synonymous, 4 synonymous substitutions, *P*-value = 0.033). The 98-nt exon introduces a frame shift which causes the appearance of a premature termination codon in the following exons, and obviously determines the translation of a different C-terminus. One might argue that the premature termination codon might be used to regulate the expression post-transcriptionally as reported for other genes (Lareau*, et al.* 2007, Ni*, et al.* 2007), or that the different C-terminus itself might play an important role in the apoptotic activating process which the BAX gene is responsible for. However, the original discovery of the BAXε variant present evidence for its expression in mouse tissues (Shi*, et al.* 1999), raising serious doubts about the validity of this variant. This is because Alu elements are not present in the mouse genome, and the transcription of the Alu cassette is not supported by any additional cDNA or EST sequence.

In conclusion, looking for signature of natural selection acting on *Alu*ternative exons fails to provide new strong evidence for the protein coding potential of Alu cassettes. One might argue that dN/dS is not an appropriate measure for investigating selection on such short sequences, and even that the Fisher's exact test is too conservative to discover significant trends. Unfortunately, SNP data are not available for these cassettes, or at least not in a quantity (less than five SNPs per exons) that would allow carrying a McDonald-Kreitman test (McDonald and Kreitman 1991) as applied for other human genes (Bustamante*, et al.* 2005).

3.3.3 *Alu*ternative Variants and Signal Peptides

The contribution of Alu derived sequences to signaling capabilities of the cell is an interesting evolutionary scenario given that Alu elements are primate specific and they contribute to the splicing variants of many genes. Searching for transcripts with the START codon of the CDS annotated in an Alu cassette resulted in 71 transcripts, in which the contribution of the Alu cassette to the CDS is shown in Figure 3.11:

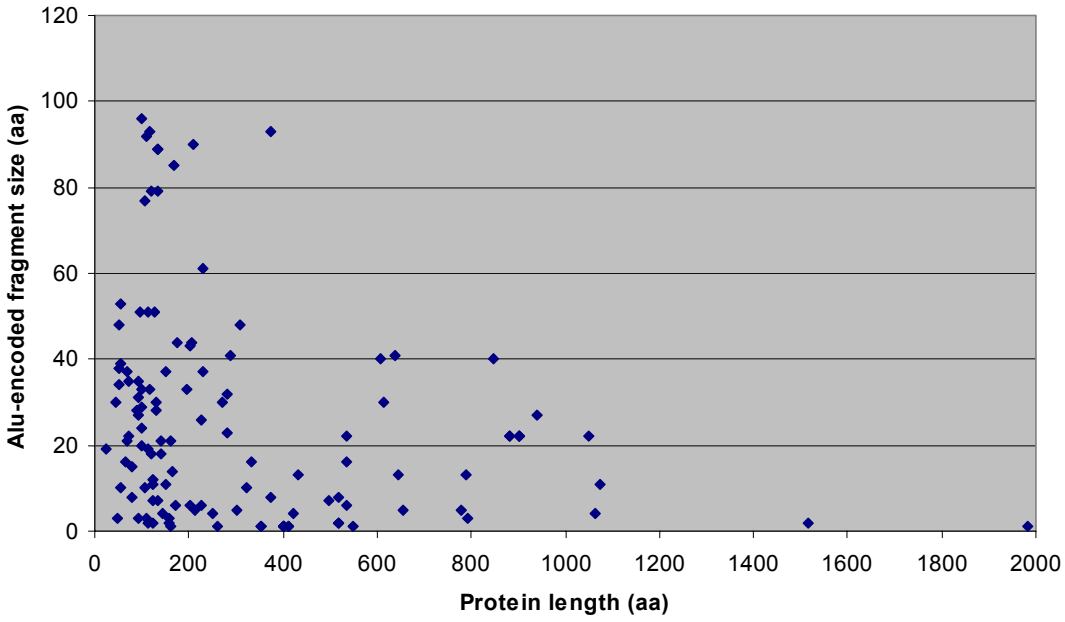

Figure 3.11: Transcripts with the START codon annotated in an Alu cassette. The size of the protein product is indicated on the X axis, while the contribution of the Alu-encoded fragment is indicated on the Y axis.

It is interesting to note that the contribution of the Alu cassette to the CDS can be as little as one amino acid, as in the case of NM_001013671 (gi:61966796) where an AluSx element provides the donor splice site of an internal exons (the acceptor splice site is donated by anonymous sequence), which is likely to be an alternative one but no

evidence for alternative transcripts exists (only 27 among the 71 transcripts have an alternatively spliced variant). However, in this case, the important contribution is providing the start codon, which in the case of the alternatively spliced variants, implies that the protein product has a different N-terminus. This alternative N-terminus could contain a signal peptide, which could determine the fate of the product in terms of cellular localization. A total of 21 among the 71 transcripts found are mono-exonic, and only two of them have alternatively spliced variants. Many of these are annotated to encode a short protein (only two of them are longer than 200 amino acids), and in most cases the protein product is made almost entirely of Alu-encoded sequence, which would suggest that they are aberrant proteins (Kriegs, *et al.* 2005), or false annotations of transcripts that do not encode any protein product.

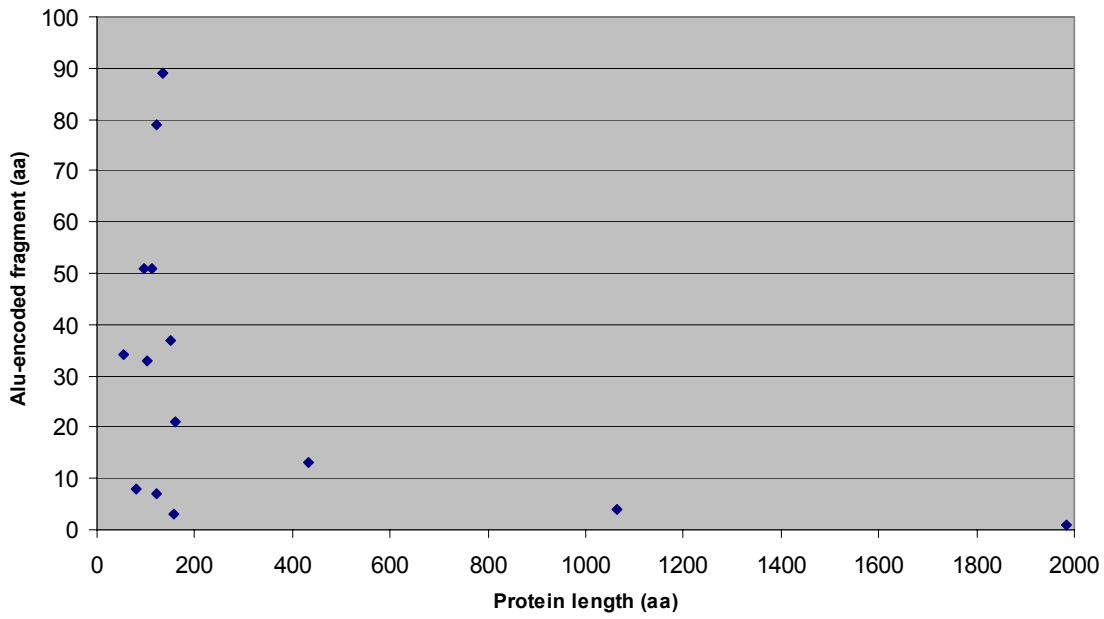

Figure 3.12: Contribution of Alu cassettes to transcripts with a predicted signal peptide.

In the case of eight transcripts, SignalP 3.0 predicts the existence of a signal peptide at the N-terminus of the protein. These are presented in Figure 3.12. Four of these are mono-exonic, three of them having no alternative variant. Four of the eight transcripts have an alternative splicing variant, which suggests that Alu elements can contribute to changing cellular localization of the protein products. One should note that the contribution of Alu cassettes to the CDS can be minimal, as in the case of MYH7B, which is a 1983 amino acid long protein, but to which Alu contributes with only one amino acid. This eliminates the argument that Alu elements do not have protein coding potential, as they contribute very little to the protein itself. Another interesting observation is that most of these cases are short proteins, which Frith *et al.* (2006) documented that are usually discarded in the annotation process. Because of this bias, there may be in fact more such proteins, but we simply do not have knowledge of their existence.

A reversed scenario is one in which an alternative splicing variant with an Alu cassette containing the START codon might not provide the signal peptide, while the normal variant would have such a feature. This would result in the arrest of the protein in cytoplasm, which might have different phenotypic effects, depending on the protein being arrested. In the dataset of 71 transcripts, there are three such examples (COL23A1, TMTC1, MRPL30), where the *Alu*ternative form misses a signal peptide, but the normal variant has one.

In conclusion, it can be said that Alu elements can contribute to differential cellular function of different proteins through the mechanism of alternative splicing, thus adding to the complexity of cell signaling and functionality.

3.3.4 The Contribution of Alu Elements to Selenoproteins

In previous chapters, cases where the Alu cassette is located in the middle or at the beginning of the CDS were shown. However, most transcripts which contain an Alu cassette in CDS are cases where the Alu cassette contains a termination codon. In the dataset of 7,262 transcripts with an Alu cassette, 821 contained an Alu cassette in CDS, among which 437 (53.2%) were transcript where the termination codon was annotated in the Alu cassette. The usual consequence is that a truncated protein is produced, and possibly a disease phenotype if the *Alu*ternative variant becomes constitutively spliced (Mitchell*, et al.* 1991). For this study, I explored, as described in Chapter 3.2.4, a rather unexpected consequence of introducing a termination codon in frame, namely the possibility that UGA termination codons could be translated into a Sec residue.

Among the 437 cases of termination codons introduced by Alu cassettes, 215 were cases of UGA codons. Among these, only in three cases a putative SECIS element with a COVE score higher than 15 was located downstream of the UGA codon. The COVE score is determined using a covariance model of the SECIS element (Kryukov*, et al.* 1999, Kryukov*, et al.* 2003), and an empirical cutoff of 15 is recommended for an element to be considered for further analyses (http://genome.unl.edu/SECISearch.html).

The highest COVE score (32.06) was obtained by the SECIS element located in the transcript 1 variant (gi:47578098) of selenoprotein N (SEPN1). This variant has two in-frame UGA codons, one of them being located in exon 3, which is an alternatively spliced AluJb cassette (Moghadaszadeh*, et al.* 2001). Later experiments have showed that this variant is unlikely to produce a functional protein, as the Alu-borne UGA is unlikely

to be recoded into a Sec residue and functions as a premature termination signal (Petit*, et al.* 2003).

The second case is that of a mitochondrial ribosomal protein MRPL48 variant, where an AluSx cassette provides an in frame UGA codon, as well as another termination codon one residue downstream from the potential Sec residue. Figure 3.13 shows the alignment between the two MRPL48 variants, and their translation. A SECIS element with a COVE score of 15.46 (Figure 3.14A) is located 192 nt downstream of the UGA codon, which is a similar distance to that found in K, O, S, TR1, TR2, and TR3 selenoproteins. Another character in common with the above mentioned selenoproteins is that between the UGA and the next termination codon there is only one other amino acid (a serine in this case). These similarities provide support for the incorporation of a Sec residue in this case, even though experimental evidence would need to confirm this prediction. However, the possibility that the Alu-containing MRPL48 variant can encode

```
                          M  S  G  T  L  E  K  V  L  C  L  R  N  N  T  I  F  K  Q  A  F  S
> NM_016055.4|657-770    ATGAGCGGAACCTTGGAAAAGGTGCTGTGCCTGAGGAACAATACCATTTTTAAGCAAGCCTTTTCT
> gi|71852588|657-895    ATGAGCGGAACCTTGGAAAAGGTGCTGTGCCTGAGGAACAATACCATTTTTAAGCAAGCCTTTTCT
                          M  S  G  T  L  E  K  V  L  C  L  R  N  N  T  I  F  K  Q  A  F  S


 L  L  R  F  R  T  S  G  E  K  P  I  Y  S  V
CTCTTAAGGTTTAGAACTTCAGGAGAGAAGCCCATCTATTCTGTAG-----------------------------------------
CTCTTAAGGTTTAGAACTTCAGGAGAGAAGCCCATCTATTCTGTAGAGAGAGACGGTCTTGCTTGTGGCCCAGGCTTGAGTACAGTG
 L  L  R  F  R  T  S  G  E  K  P  I  Y  S  V  E  R  D  G  L  A  C  G  P  G  L  S  T  V
                                             <-------------------- AluJo -------------


                                                                                    G ...
------------------------------------------------------------------------------------GT...
GCATGATCATAGCTCCCTGCAGCCTCGAACTCCTGGGTTCAAGCAATCCTCCTGCCTCAGCCTCTGGAGTAGCTGGGATTACAGGT...
 A (*) S  *
------------------------ AluJo ---------------------------------------------------->
```

Figure 3.13: The translation of the *Alu*ternative variant of MRPL48 (gi:71852588) provides a UGA codon in frame, followed by a serine, and a UAG termination codon which could serve as the end of the CDS.

a selenoprotein is a fascinating evolutionary event that is specific to primates. The SECIS sequence, which is present in other mammalian species as well, such as mouse, has evolved towards a potentially functional element, yielding a COVE score below 3 in mouse, 14.95 in macaque, and 15.46 in human. One should note that in the normal variant, the SECIS element is included in the CDS, which in theory would offer protection from neutral mutations, but it seems that the strong purifying selection acting on this segment (dN/dS = 0.208, *P*-value = 0.003) allowed for mutations that resulted in forming a compatible SECIS structure. One cannot argue that mutations to create this

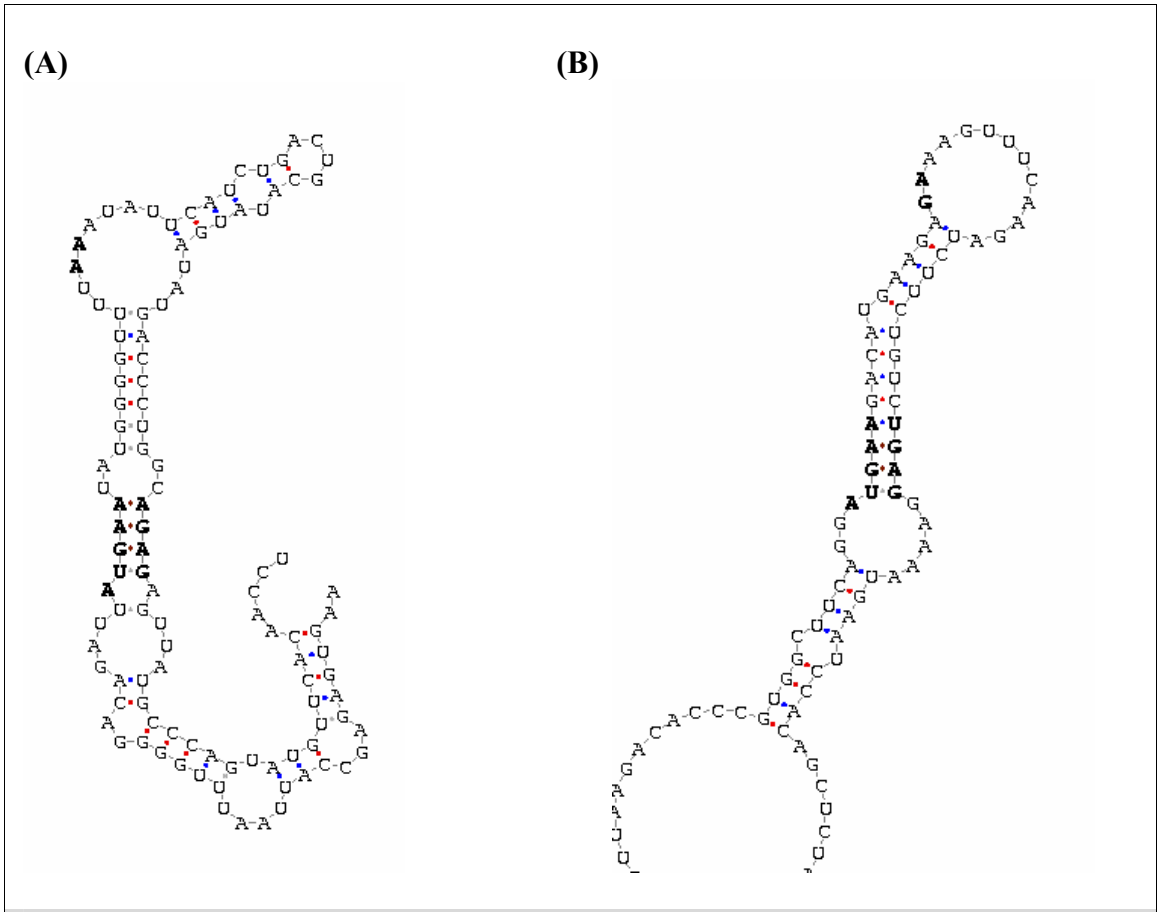

Figure 3.14: Secondary structures of SECIS elements in Alu-containing transcripts of MRPL48 (**A**) and CHEK2 (**B**) genes. Their COVE scores determined by SECISearch are 15.46, and 17.07, respectively.

structure were favored, since they could have had no function without the presence of an UGA codon in frame. Therefore the presence of a high-scoring SECIS element is very likely the result of chance mutations in a purifying selection environment. Primate species could have benefited from this evolutionary chance after they acquired the Alu cassette that now provides an in-frame UGA codon, and thus could add one more selenoprotein to their selenoproteome.

A third example is that of the Alu-containing variant of CHEK2 (Figures 3.14B, 3.15). This is slightly different that the case of MRPL48 because the extension of the CDS after the putative Sec residue extends almost to the end of the transcript (position 1744-1746) which would include the SECIS element (positions 1555-1643) in the CDS. This is rather unusual for selenoproteins, because all 25 known selenoproteins have their SECIS in the 3' UTR. However, a recent study (Mix*, et al.* 2007) indicates that SECIS elements can function even if they are included in the CDS. Moreover, recent evidence indicates that CHEK2 is activated by oxidative stress, which might be an indication that, indeed, CHEK2 can function as a selenoprotein. Selenoproteins are known to be actively involved in protection against oxidative damage, and due to this property they are used in several cancer treatments (Loflin*, et al.* 2006, Lu*, et al.* 2006, Smalheiser and Torvik 2006, Papp*, et al.* 2007).

In the light of recent evidence (Mix*, et al.* 2007), cases where the SECIS element is located upstream of the UGA codon might also be considered for further experimental analyses. I found one such case, in an mRNA sequence (gi:51466863) predicted based on two ESTs (gi:3593846 and gi:27847386). The putative SECIS element in located in the 5' UTR (between coordinates 36-128), and has a COVE score of 10.57, thus below the

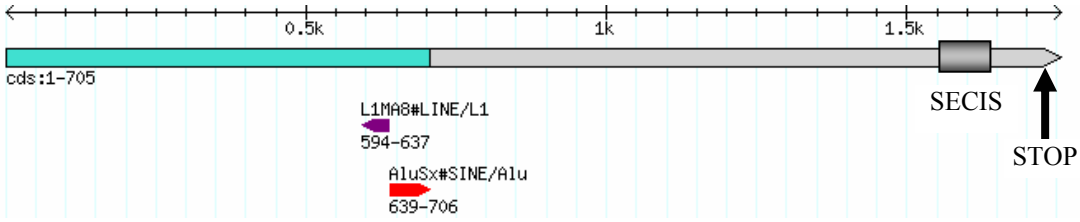

Figure 3.15: Location of the Alu cassette, SECIS element (gray box), and termination codon (indicated by an arrow) on the annotated CHEK2 Alu-containing variant (gi:45356853). Original CDS annotation, which includes a UGA codon at position 703-705, and location of TE fragments are displayed as shown in the ScrapYard database.

empirical, thus permissive, threshold of 15. Because the SECIS element is located in the 5' UTR, this transcript does not resemble classic selenoproteins, but given that SECIS elements upstream of UGA codons could be functional, this case could be considered for experimental validation. The search for other putative selenoproteins generated by the presence of UGA-containing TE cassettes in other transcriptomes (mouse, rat), yielded no other significant results. This might be due to the poorer pool of annotated transcripts, thus such cases might be revealed in the future. Nonetheless, the two cases of human transcripts that can putatively promote incorporation of Sec residues in proteins represents an interesting evolutionary, with possible beneficial effects for fitness, namely a better resistance to oxidative stress.

3.4 Conclusions

As revealed by this study, a large number of human genes present alternatively spliced transcripts with Alu cassettes included in the CDS, as one would expect from the most numerous TE in the human genome. Whether Alu sequences can provide functional protein fragments still remains an elusive conclusion, as no strong evidence for this was found by the several computational methods used. However, in a few cases, such as DRADA2b, the possibility that Alu is tolerated in proteins exists. This creates the possibility that Alu exaptation events will lead to increased variability at the protein level, with potential beneficial effects. By computational methods, a few novel predictions were made about the possible contribution of Alu elements to the enrichment of the human proteome, such as providing signal peptides, and creating novel selenoproteins. While these predictions are very exciting from the evolutionary point of view, they need further experimental support. Nonetheless, it is only through computational methods, which provide a perfect complement for experimental biology, that so diverse scenarios can be explored, given the vast amount of genomic data generated in recent years.

Chapter 4

Transposable Elements Are a Significant Source of Transcription Regulating Signals

4.1 Introduction

Exaptation of TE fragments into protein coding sequences can have direct impact on the phenotype by altering the proteome, as shown in the previous chapters. Equally important is the influence that TEs can have on gene expression through their capacity of providing transcriptional regulatory elements novel to specific genes. After discovering that long terminal repeats (integral parts of some retrotransposons) carry promoter and enhancer motifs, it became clear that integration of such elements in proximity of a host gene must have an influence on this gene expression (Sverdlov 1998). Many TEs have been described in the last decade that can add a variety of functions to their targeted genes. These include polyadenylation sites, promoters, enhancers, and silencers (Makalowski 1995). It seems that a sizable fraction of eukaryotic, gene-associated regulatory elements arose in this modular fashion by insertion of TEs, and not only by point mutations of static neighboring sequences. When a TE is inserted upstream from a gene, a few short motifs can be conserved if they were subjected to selective pressure as promoters or enhancers of transcription. Even though the rest of the TE sequence might evolve beyond recognition due to absence of functional constraints, TEs are hence exapted into a novel function. A recent survey that analyzed 846 functionally characterized *cis*-regulatory elements from 288 genes showed that 21 of those elements

(~2.5%) from 13 genes (~4.5%) reside in TE-derived sequences (Jordan, *et al.* 2003). The same study showed that TE-derived sequences are present in many more (~24%) promoter regions, defined as ~500 bp located 5' of functionally characterized transcription initiation site. Similarly, van de Lagemaat *et al.* showed that the 5' UTRs of a large proportion of mammalian mRNAs contain TE fragments, suggesting that they play a role in regulation of gene expression (van de Lagemaat, *et al.* 2003). One should note that the TE influence on gene regulation upon insertion in promoter regions is only due to chance similarity of TE sequence to various *cis*-regulatory elements, or to the presence of regulatory elements that were active in regulating the transcription of the TE itself. In order to estimate what potential TEs have in regulating the expression of neighboring genes, the content in putative transcription regulating signals was evaluated for TEs located in promoter regions of all annotated human genes.

4.2 Materials and Methods

4.2.1 Finding TEs in Promoter Sequences

For the purpose of this study, the July 2003 assembly of the human genome available from the Golden Path at the University of California Santa Cruz (http://genome.ucsc.edu/goldenPath/hg16/), and corresponding gene annotation (we used the refFlat files which contain annotation for RefSeq and predicted genes) were used. For every gene, we extracted 2,000 nucleotides upstream from the annotated transcription start coordinate, which were assimilated with gene promoter regions. The 20,193 excised

promoter sequences were then scanned for occurrence of TEs using the May 15, 2002 version of RepeatMasker (http://www.repeatmasker.org) with default options, but ignoring simple repeats and low complexity regions ("-nolow" parameter).

4.2.2 Identification of Transcription Signals

TRANSFAC database of transcription factor binding sites, maintained by Biobase (http://www.biobase.de), was used as a source of verified transcription signals. We relied upon the MATCH program (Kel*, et al.* 2003) from the same software suite for finding such putative signals in human promoter regions. MATCH uses predefined positional weight matrices (PWM), which we chose based on the TRANSFAC classification of transcription factor binding sites (http://www.gene-regulation.com/pub/databases/transfac/cl.html). Representative high-quality matrices were chosen for each class of transcription factors (Table 4.1). If high-quality matrices were not available, low-quality matrices were chosen only if they were based on more than ten experimentally characterized binding sites, in order to reduce the false positive identification rate (Qiu*, et al.* 2002). The MATCH profile was created using matrix similarity cutoff values corresponding to a false negative rate of 50% (FN50 values). While this setting potentially excludes half of the biologically significant transcription factor binding sites (TFBS), it drastically reduces the number of false positives matches. Only binding sites completely overlapping with TE sequences were kept for further analysis.

Table 4.1: Representative position weight matrices (PWM) from TRANSFAC database used for identifying transcription factor binding sites in human promoter regions.

Class	Factor name	Matrix ID	Quality	Matrix similarity cutoff[a]
Superclass: Basic Domains				
Leucine Zippers	XBP-1	V$AP1_C	High	0.98
	CRE-BP1	V$CREBP1_Q2	High	0.96
	C/EBPα	V$CEBP_C	High	0.93
Helix-Loop-Helix	E12	V$E12_Q6	High	0.97
	MyoD	V$MYOD_01	High	0.94
Helix-Loop-Helix / Leucine Zipper	USF	V$USF_Q6	High	0.95
	c-Myc	V$MYCMAX_01	High	0.97
RF-X	RF-X2	V$RF-X1_01	High	0.94
Helix-Span-Helix	AP-2γ	V$AP2_Q6_01	Low	0.92
Superclass: Zinc-coordinating Domains				
Zinc Finger – Nuclear Receptor	GR	V$GRE_C	High	0.92
	ER	V$ER_Q6	High	0.94
	HNF-4α1	V$HNF4_01	High	0.86
Cys4 Zinc Fingers	GATA-1	V$GATA1_02	High	0.97
	GATA-3	V$GATA_C	High	0.96
Cys2His2 Zinc Fingers	YY1	V$YY1_02	High	0.92
	Egr-1	V$EGR1_01	High	0.96
Superclass: Helix-Turn-Helix				
Homeo Domain	HNF-1A	V$HNF1_01	High	0.90
	Oct-2B	V$OCT_C	High	0.93
Paired box	Pax-6	V$PAX6_01	High	0.88
	Pax-5	V$PAX_Q6	High	0.86
Fork head / Winged Helix	HNF3-α	V$HNF3B_01	High	0.94
	E2F-1	V$E2F_Q6	High	0.91
Tryptophan Clusters	c-ETS-1 p54	V$ETS1_B	High	0.94
	IRF-1	V$IRF1_01	High	0.97
Superclass: beta-Scaffold Factors				
Rel Homology Region	p50	V$NFKAPPAB_01	High	0.96
	p65	V$NFKB_Q6_01	High	0.91
STAT	p91	V$STAT_01	High	0.97
MADS Box	MEF-2A	V$MEF2_02	High	0.93
	SRF	V$SRF_C	High	0.93
TATA Binding Proteins	TBP	V$TATA_C	High	0.95
HMG	Sox-9	V$SOX9_B1	High	0.95
	SSRP1	V$TCF4_Q5	High	0.98
Heteromeric CCAAT Factors	CP1B	V$NFY_Q6	High	0.96
Grainyhead	CP2	V$CP2_02	High	0.93
Runt	AML-3	V$AML_Q6	High	0.97

Legend: [a] The matrix similarity cutoff corresponds to a false negative rate of 50% (FN50).

4.2.3 Testing the Significance of TE Content in TFBS

Because TFBS are short, they can be found by chance on any DNA sequence, including TEs. Over- or under- representation of certain binding sites in TE sequences as compared to random DNA would provide evidence that TEs have a significant potential in altering gene regulation when inserted in their close proximity, thus this aspect needed to be tested. For this purpose, sets of randomly generated sequences mimicking the number, length, and GC content of individual elements found by RepeatMasker for every TE class (Table 4.2) were created. Even though controversial, the choice of randomly generated sequences offers a non-biased dataset to compare the sequence content of different TE classes in regulatory elements. Non-repetitive intergenic sequences have the big disadvantage of unknown origin. It is widely accepted that much more than 50% of the human genome originated in TEs, thus a data set of randomly selected non-repetitive intergenic sequences can be in fact an uncontrollable mix of TE-derived sequences. Moreover, the annotation of regulatory elements is incomplete, making impossibly to avoid selecting already functional regulatory sequences. Putative binding sites were identified using MATCH with the same settings used for the set of real sequences. Normal distributions of binding site occurrences per element were generated using 1000 samples of size 100 for every combination of binding site - TE class. The sample unit was the number of binding sites found by MATCH on real TEs or randomly generated sequences, and therefore the size of datasets was different for every TE class (see Table 4.2 for number of TEs found for every TE class). The significance of difference was assessed with Student's t-test, assuming equal variances. A stringent significance

threshold of 0.0001 was set in order to reduce the number of false positive findings.

4.3 Results and Discussion

4.3.1 TE Content of Promoter Regions

Among the 20,193 gene promoter regions analyzed, 16,665 (~83%) were found to contain TE-derived sequences. This represents a much higher percentage than the ~24% previously reported (Jordan*, et al.* 2003), and it is probably due to the longer 5' upstream region analyzed (2000 bp instead of 500 bp). It should be noted, however, that about half of these TE-derived sequences do not carry putative regulatory elements (Figure 4.1), and consequently, transcriptional regulation of an additional 3,377 genes (~17%) would remain unaffected after the insertion of TE fragments in their regulatory regions (Figure 4.2). The remaining 13,288 (~64%) are still a significant number of genes whose transcriptional regulation could be fortuitously influenced by TEs. In reality, the number is probably much smaller because many of these putative signals would likely be non-functional. A summary of RepeatMasker findings in the 20,193 promoter sequences is presented in Table 4.2. Miscellaneous repetitive elements, such as SVA and SVA2 SINE-like elements, were not included because they are unlikely to influence gene regulation at genomic scale due to their very low frequency within promoter regions (only 35 occurrences were detected).

Table 4.2: Summary of RepeatMasker findings on human promoter sequences.

TE Class	Number of TEs Detected	Number of Promoter Regions containing each TE class	% of Total Sequence	Minimum TE Length (bp)	Average TE Length (bp)	Maximum TE Length (bp)	GC Content (%)
SINE	30271	13759	15.72	11	210.7	427	51.29
LINE	11356	7165	6.18	11	220.8	2000	39.59
LTR	5534	3633	3.79	11	277.8	2000	45.42
DNA	4311	3137	1.79	11	168.5	1880	39.62

It is interesting to note that the number, as well as the fraction of total sequence, of SINE elements found in promoter regions, is almost three fold larger than that of LINE elements. This is in agreement with previous reports based on smaller datasets (Jordan, *et al.* 2003), but in contrast to the proportion of SINE/LINE elements within the human genome. While LINE elements account for the largest fraction of the human genome among TEs (20.42% vs. only 13.14% for SINE elements), they are only twice less frequent (~0.8 million vs. ~1.5 million copies) as compared to SINE elements (Lander, *et al.* 2001). Our observation is not surprising because a genome wide survey already showed that SINE elements are more frequent in GC-rich regions, while LINE elements are more frequent in AT-rich regions (Korenberg and Rykowski 1988). SINE density in AT-rich regions tends to be higher near genes (Smit 1999). The reason for this appears not to be due to insertion site preference, because both SINE and LINE elements seem to insert randomly in the genome (Arcot, *et al.* 1998, Ovchinnikov, *et al.* 2001). One hypothesis that may explain this pattern proposes that SINE elements are subjected to differential retention rates influenced by their ability to regulate protein translation if readily transcribed from open chromatin such as is found near genes (Liu, *et al.* 1995, Chu, *et al.* 1998, Schmid 1998). The findings from this study suggest that differential

retention of different TE classes might be also determined by their content in transcription factor binding sites.

One can also note that the number and proportion of LTR elements and DNA transposons in promoter regions is the lowest among the four classes (Table 4.2, Figure 4.1). This is probably due to their higher divergence and fragmentation, which makes their detection harder, or even impossible, with current similarity searching techniques. It is known that LTR elements are remnants of more ancient retroviruses, and many of the DNA elements, such as MER75 or Charlie8, are labeled as "fossils". Similarly to the case of LINEs, the lower number of LTR elements found in promoter regions might be due to lower retention rates caused by the intrinsic nature of their sequence, which is known to carry promoter and enhancer motifs (Sverdlov 1998). Additionally, the insertion of still active, younger, and more abundant SINE and LINE elements such as Alu and LINE1, respectively, may have gradually replaced older elements from the 2,000 nucleotide promoter regions analyzed. Consequently, fewer cases in which LTR and DNA elements occupy the entire or most of the promoter region are found as compared to similar instances of LINE elements (Figure 4.1).

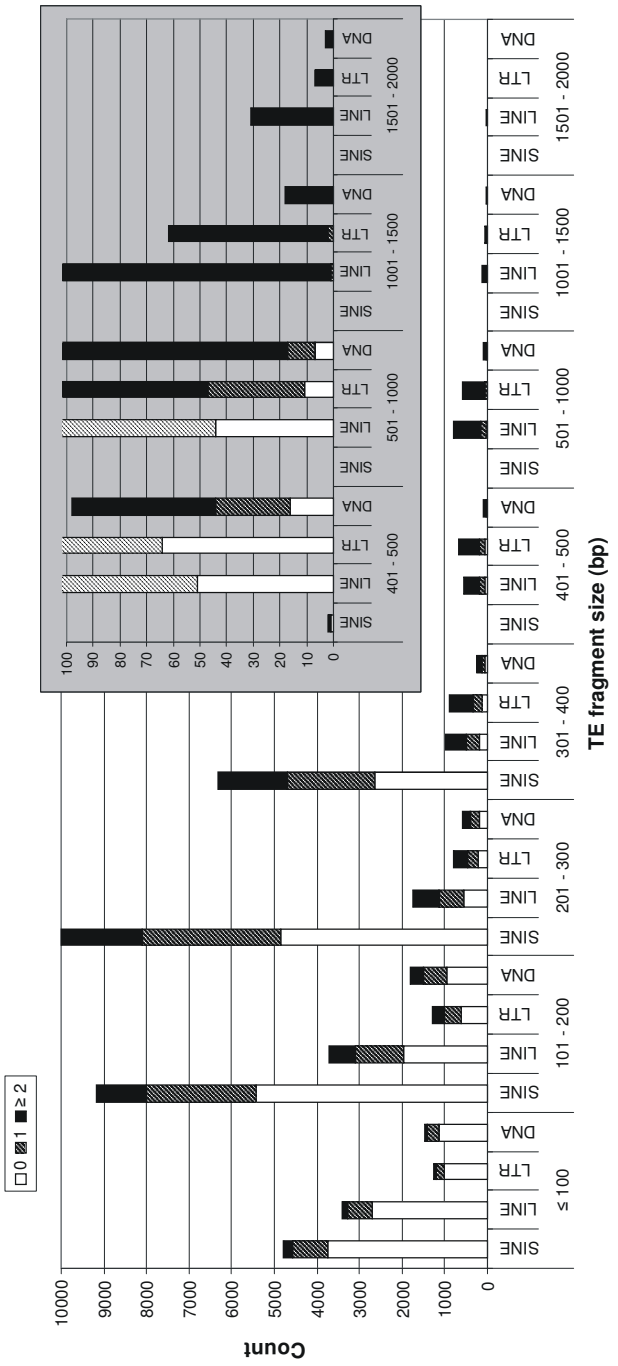

Figure 4.1: Size distribution of TE fragments found in gene promoter regions and distribution of transcription factor binding site occurrence on each TE size subclass. The last four bins are 500-bp bins, which, for clarity, are presented in higher-resolution scale as well. Open boxes indicate no occurrence of putative regulatory signals in TE fragments, hashed boxes indicate one occurrence, and solid boxes indicate two or more such occurrences.

4.3.2 TE Content in Transcription Regulating Signals

The number of potential transcription factor binding sites found in promoter-residing TEs using MATCH is shown in Table 4.3. Using the sampling technique described above, significance of whether TEs contain more binding sites than random sequences was inferred. While statistical significance does not imply biological function, it shows that TEs, when inserted into promoter regions, have an increased potential to alter gene expression in a manner specific to the signals they contain as compared to random sequences.

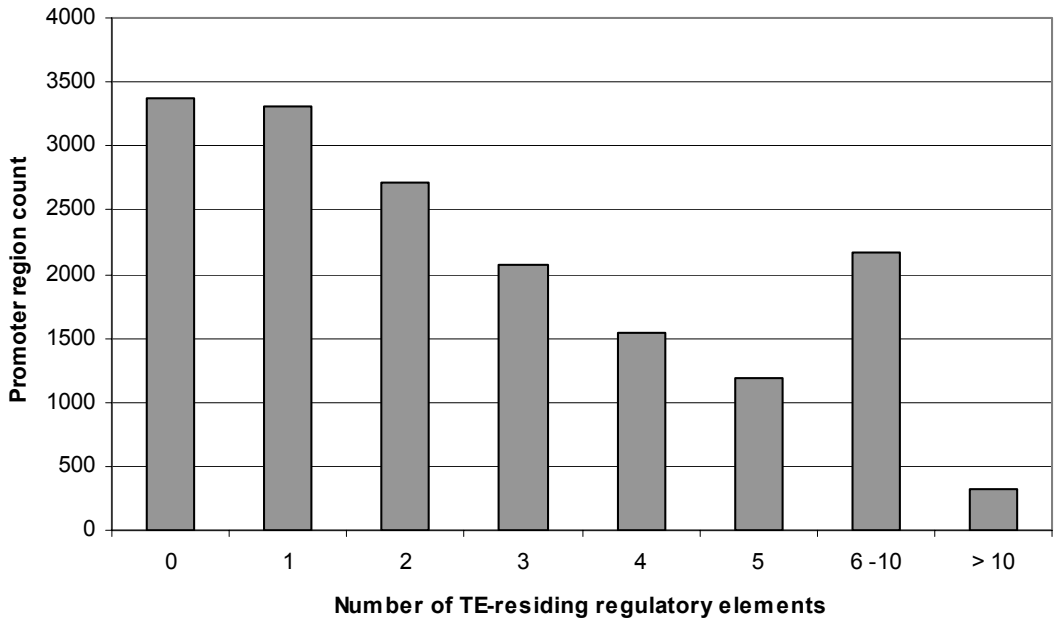

Figure 4.2: Distribution of promoter regions based on their content in transcription regulating signals contributed by TE-derived sequences. Promoter regions that do not contain TE-derived sequences are not included in this distribution. The maximum number of TE-contributed putative regulating signals found in a single promoter region is 32.

Table 4.3: Comparison of numbers of putative transcription factor binding sites identified by MATCH on real TE and randomly generated sequences.

Binding Site Class	TE Class											
	SINE			LINE			LTR			DNA		
	Obs.[a]	Rnd.[b]	p-value[c]	Obs.[a]	Rnd.[b]	p-value[c]	Obs.[a]	Rnd.[b]	p-value[c]	Obs.[a]	Rnd.[b]	p-value[c]
Superclass: Basic Domains												
Leucine Zipper	1335	1363	0.493	472	530	$4.87 \cdot 10^{-06}$	565	364	$1.20 \cdot 10^{-97}$	206	153	$1.62 \cdot 10^{-29}$
Helix-Loop-Helix	2368	1385	$2.43 \cdot 10^{-122}$	703	258	$1.46 \cdot 10^{-210}$	469	265	$2.15 \cdot 10^{-119}$	234	73	$5.29 \cdot 10^{-247}$
Helix-Loop-Helix / Leucine Zipper	873	1780	$2.12 \cdot 10^{-120}$	158	277	$4.92 \cdot 10^{-37}$	320	268	$1.04 \cdot 10^{-16}$	109	77	$3.58 \cdot 10^{-19}$
RF-X	188	836	$1.25 \cdot 10^{-219}$	183	200	0.00012	159	169	0.0028	75	54	$7.05 \cdot 10^{-16}$
Helix-Span-Helix	2515	3572	$3.67 \cdot 10^{-117}$	622	278	$7.19 \cdot 10^{-171}$	733	393	$5.07 \cdot 10^{-221}$	245	82	$7.40 \cdot 10^{-264}$
Superclass: Zinc Domains												
Zinc Finger / Nuclear Receptor	2764	2628	0.0004	1259	871	$7.99 \cdot 10^{-102}$	961	649	$8.36 \cdot 10^{-172}$	233	260	$1.06 \cdot 10^{-06}$
Zinc Finger / GATA	2402	6440	*0*	2776	3336	$3.17 \cdot 10^{-70}$	1892	1826	$8.50 \cdot 10^{-07}$	641	969	$3.67 \cdot 10^{-180}$
Zinc Finger / Cis2His2	134	64	$2.43 \cdot 10^{-15}$	33	16	$1.10 \cdot 10^{-14}$	45	17	$3.91 \cdot 10^{-51}$	5	6	$4.92 \cdot 10^{-06}$
Superclass: Helix Turn Helix												
Homeo Domain	430	244	$3.89 \cdot 10^{-36}$	385	303	$7.15 \cdot 10^{-18}$	136	109	$2.46 \cdot 10^{-10}$	86	80	0.0003
Homeo / Paired Box	575	771	$3.07 \cdot 10^{-12}$	416	339	$1.47 \cdot 10^{-20}$	325	201	$4.09 \cdot 10^{-102}$	123	86	$5.92 \cdot 10^{-34}$
Fork head / Winged Helix	4155	3558	$1.12 \cdot 10^{-17}$	1579	1470	0.00014	571	852	$2.73 \cdot 10^{-160}$	468	392	$8.08 \cdot 10^{-27}$
Tryptophan Clusters	338	351	0.0117	293	107	$2.48 \cdot 10^{-121}$	276	81	$2.79 \cdot 10^{-275}$	42	34	$1.68 \cdot 10^{-09}$
Superclass: beta-Scaffold Factors												
Rel Homology Regions	349	771	$2.11 \cdot 10^{-59}$	301	165	$1.64 \cdot 10^{-44}$	406	163	$2.44 \cdot 10^{-173}$	60	54	0.842
STAT	44	305	$8.69 \cdot 10^{-124}$	141	103	$6.91 \cdot 10^{-16}$	180	78	$3.54 \cdot 10^{-154}$	53	34	$1.07 \cdot 10^{-23}$
MADS Box	144	74	$1.20 \cdot 10^{-13}$	94	69	$2.29 \cdot 10^{-08}$	56	38	$1.80 \cdot 10^{-08}$	20	25	$5.02 \cdot 10^{-06}$
TATA Binding Proteins	783	440	$2.11 \cdot 10^{-59}$	895	488	$2.57 \cdot 10^{-175}$	391	184	$1.22 \cdot 10^{-220}$	230	133	$6.36 \cdot 10^{-117}$
HMG	791	2238	*0*	1487	1338	$5.15 \cdot 10^{-19}$	742	709	$1.69 \cdot 10^{-07}$	391	369	$2.52 \cdot 10^{-05}$
Heteromeric CCAAT / Histone Fold	134	935	$9.67 \cdot 10^{-299}$	356	488	$3.64 \cdot 10^{-31}$	869	285	0	83	131	$4.88 \cdot 10^{-66}$
Grainyhead	259	577	$1.16 \cdot 10^{-91}$	147	95	$7.12 \cdot 10^{-22}$	317	84	0	57	22	$3.49 \cdot 10^{-64}$
Runt	367	278	$6.34 \cdot 10^{-10}$	195	120	$3.78 \cdot 10^{-38}$	129	74	$1.78 \cdot 10^{-63}$	49	32	$1.18 \cdot 10^{-15}$

Legend: [a] Number of putative transcription factor binding sites identified on promoter residing TEs. [b] Number of putative transcription factor binding sites identified on randomly generated sequences. [c] P-values indicating the significance of difference between observations in the two sets of sequences are highlighted or *italicized* if the number of putative TFBS in real TEs as compared to randomly generated sequences is over- or under-represented, respectively. Significance level was set to 0.0001.

Unlike the other three TE classes, LTR elements are likely to carry almost all of the binding site classes (Table 4.3). This might be a consequence of their original function of providing regulatory elements for retrovirus protein coding genes (Sverdlov 1998). The fork-head / winged helix binding site is the only one being under-represented in LTR elements, the RF-X binding site is significantly over-represented only at 0.01 error level, but the remaining 18 classes are all over-represented in LTR elements. LINE elements, and particularly LINE1 elements, are known to contain YY1 P*ol*II (Athanikar*, et al.* 2004) and antisense promoters that have been shown to influence the transcription of adjacent genes (Speek 2001). This study finds that LINE elements are likely to carry 14 over-represented classes of binding sites, double the number of transcriptional signals over-represented in SINE elements, which do not contain *Pol*II promoters. Active SINE elements carry, however, *Pol*III A and B boxes (Schmid and Rubin 1995), but which do not influence the transcription of protein coding genes. The difference in over-represented signals might offer an alternative hypothesis for different retention rates near genes observed for different TE classes. Carrying more transcription regulating signals can cause in more cases alteration of gene expression, thus with more possible deleterious effects. Consequently, elements with more regulatory signals are subjected to negative selection, and elements with fewer gene regulatory signals are more likely to be tolerated and fixed as they are less likely to disrupt the regulation of genes upstream of which they are inserted. The comparison is particularly interesting between SINE and LINE elements, the former being the most abundant TE class in promoter regions, as previously discussed. The fact that Alu elements are primate specific, and not present in rodents, for example, might indicate that they are indeed, at least for the most part, tolerated rather

than positively selected. While it seems reasonable to have fewer promoter-residing LTR elements, other factors, such as the age and genomic abundance, might explain why DNA transposons are the least present in promoter regions.

Another interesting observation is that all 20 binding site classes are likely to be over-represented in at least one TE class, but only three (helix-loop-helix, TATA binding proteins, runt) are over-represented in all four TE classes. These are all transcription factor binding sites that control the expression of many genes (Beltran*, et al.* 2005, Kitayner*, et al.* 2005, Zukunft*, et al.* 2005), while binding sites over-represented in only one of the TE classes (Table 4.3), RF-X (Hasegawa*, et al.* 1991), zinc finger / GATA (Pikkarainen*, et al.* 2004, Liew*, et al.* 2005), and heteromeric CCAAT factors / histone folds (Linhoff*, et al.* 1997), appear to have more specific functions. I should reemphasize that these findings do not necessarily imply biological significance, in spite of statistical significance. Two reasons might be invoked here. One is the fact that 2000-nucleotide long 5' flanking regions are admittedly not perfect substitutes for verified promoter sequences. Our approach is supported by the fact that 5' flanking regions were shown to be enriched in promoter sequences (Suzuki*, et al.* 2002), and several studies have successfully used the same 5' flanking region to study influence of promoter sequences (Wang*, et al.* 2001, Zukunft*, et al.* 2005). Secondly, further studies are necessary to show what proportion of TE-residing transcription factor binding sites is in fact functional. MATCH findings should be taken as "potential" until experimental evidence can be provided, in spite of stringent criteria being used for defining over-representation (FN50 values in finding potential binding sites, 0.0001 statistical significance cutoff). Specific examples of such regulatory elements are known to reside in Alu elements (Hamdi*, et al.*

2000, Clarimon*, et al.* 2004, Oei*, et al.* 2004), for example, but the influence that TEs can have at genomic level is of particular interest for this study. It emphasizes the *potential* of these TE-borne signals to act as currently unknown regulatory elements or to gain function when carried into new genomic locations by their host TEs, making them contributors to changes in the genomic expression landscape, thus to genomic novelty.

4.4 Conclusions

Following these findings, a few conclusions could be drawn. In humans, SINE retrotransposons are found to be the most abundant class of TEs in promoter regions, in agreement with previous studies based on smaller datasets. Interestingly, they are three times more numerous and voluminous than LINE retrotransposons, in spite of a inverse SINE/LINE proportion at genomic level.

The abundance of different TE classes in promoter regions may be a reflection of different retention rates due to their content in potential regulatory signals. As expected, LTR elements carry most type of binding sites, followed by LINE, DNA, and SINE elements in that order, and consequently, SINE elements are the most abundant in promoter regions, followed by LINE, LTR, and DNA elements.

It is clear that TEs have a big potential to influence gene regulation at genomic scale by carrying potential transcription regulating signals. When TEs are inserted in promoter regions, their sequence can fortuitously provide regulatory elements that can influence the expression of the genes in their vicinity. In this way, TEs can contribute to

creating elements of novelty in the process of genomic evolution, and favorable influences could by eventually fixed in populations.

This chapter was published in Gene, Vol. 365, Thornburg, B.G., Gotea, V., Makalowski, W., Transposable elements as a significant source of transcription regulating signals, 104-110, Copyright Elsevier Ltd (2006).

Chapter 5

The ScrapYard Database

5.1 Motivation

The advances in sequencing technologies have led to an exponential increase in the amount of sequence data available for analysis. This is probably best exemplified by the increase of GenBank size (http://www.ncbi.nlm.nih.gov/Genbank/index.html) and by the sequencing to near completion of more and more genomes, such as macaque (Gibbs, *et al.* 2007), dog (Lindblad-Toh, *et al.* 2005), and even twelve species from the genus Drosophila (http://rana.lbl.gov/drosophila). During recent years, the computational power increased tremendously as well, but despite this, analysis of transposable elements are still computationally expensive, especially when dealing with large datasets, such as complete transcriptomes or genomes. Large scale analyses are necessary for a better understanding of the impact of transposable elements on genome organization and evolution, thus having pre-computed TE annotation is a desiderate for many researchers.

Following Makalowski's concept of genomic scrap yard (Makalowski 1995, 2000), the creation of the "**S**crap **Y**ard **D**ata**B**ase" (SYDB) was proposed. With an initial focus on TE exaptation into protein coding sequences, the SYDB is meant to be a database of TE-containing transcripts from different vertebrate species. Among these, the human receives special attention because of the high bio-medical interest and the quality of data available.

5.2 Implementation and Current Content of the SYDB

The SYDB was set up with data from three species: human, mouse, and rat. Transcripts from each species were retrieved from Genbank using two ENTREZ queriez combined with the Boolean OR operator:

```
1: human[ORGN] AND "bimol mrna"[Properties] AND "srcdb refseq"[Properties]
2: human[ORGN] AND "bimol mrna"[Properties] AND "complete"[Title] AND
"cds"[Title]
```

We also included sequences annotated by the H-Invitational project (Imanishi*, et al.* 2004) which provided additional annotations where not available in the Genbank records. We excluded from our analysis all sequences for which the coding sequence (CDS) was not annotated, because we are interested in the position of the TE fragments relative to the CDS. Records with identical sequence were eliminated using the **`patdb`** program (http://blast.wustl.edu/blast/README.html), but redundancy information was kept so that searches using accession, gi numbers, and descriptions of the corresponding redundant records could still be performed. Identification of TE occurrences on these transcripts was accomplished using RepeatMasker version of May 22, 2003 (http://www.repeatmasker.org). Only records on which the presence of a TE was detected were further included in the SYDB. The content of the SYDB was organized into a MySQL database (version 4.0.27 for SUN Solaris 2.8), whose schema is shown in Figure 5.1.

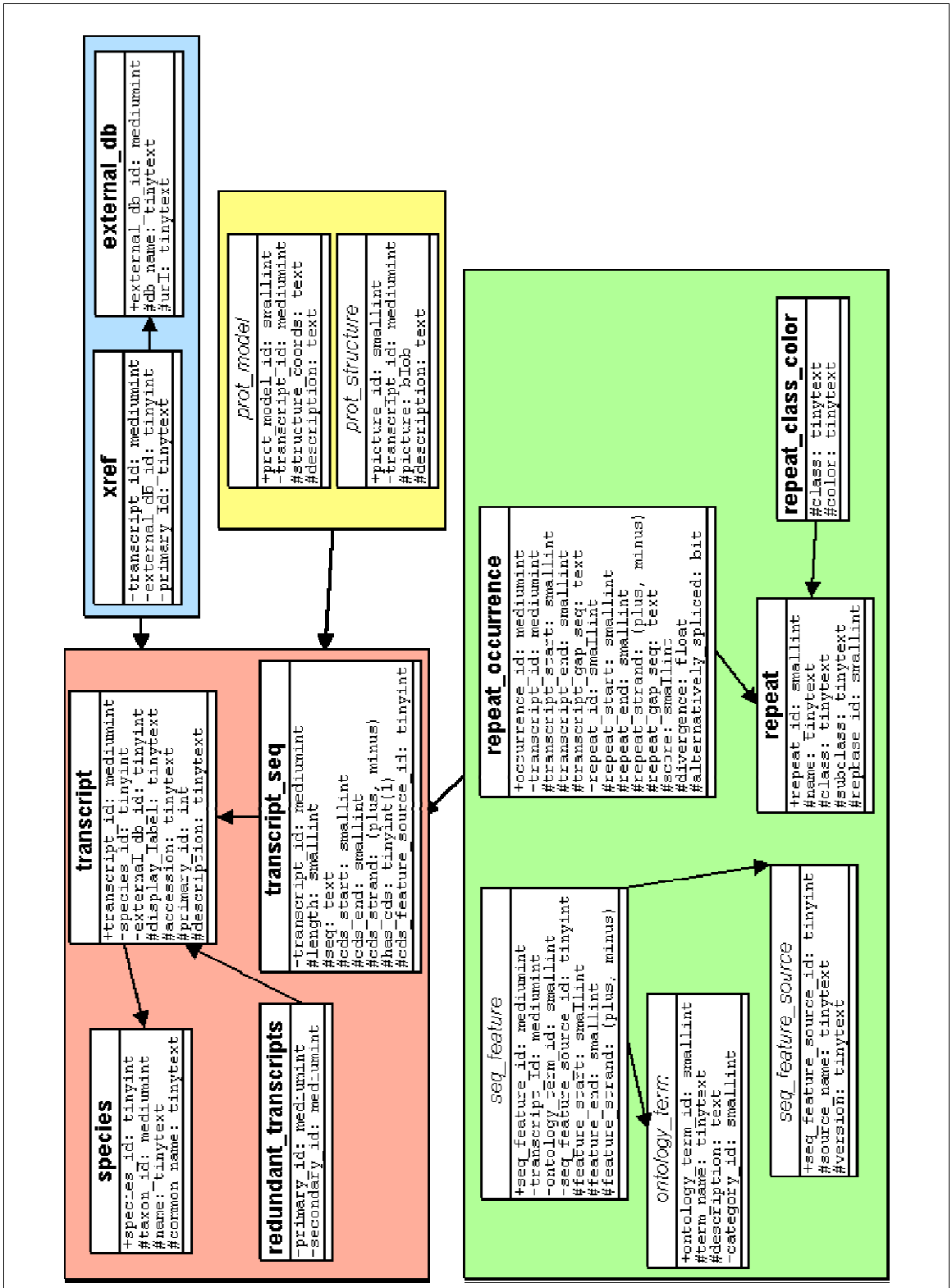

Figure 5.1: The schema of the SYDB, representing tables and keys in each table. Tables with title in italics are not yet implemented. "+" represents a primary key in a table, "-" foreign key, that is keys primary in other tables, and "#" are local fields.

Currently, the SYDB contains TE annotations for 18,295 human, 8895 mouse, and 2150 rat transcripts as detailed in Table 5.1.

Table 5.1: Basic statistics of the three mammalian transcriptomes analyzed.

	Human	Mouse	Rat
Initial number of sequences	86,544	82,226	28,938
Number of non-redundant sequences	66,262	69,112	24,548
Total length (bp)	144,410,976	102,037,991	45,664,657
Average mRNA length (bp)	2,179	1,476	1,860
GC level (%)	49.98	50.01	51.33
Number of unique sequences containing a TE-cassette in any region	15,783	7,320	1,716
Number of unique sequences containing a TE-cassette in CDS	1,397	700	368
Number of TE-cassettes lying exclusively in 5' UTR	3,823	1,179	250
Number of TE-cassettes overlapping 5' UTR and CDS	737	211	63
Number of TE-cassettes lying exclusively in CDS	1,544	1,147	548
Number of TE-cassettes lying exclusively in 3' UTR	20,987	14,618	1,554
Number of TE-cassettes overlapping 3' UTR and CDS	1,477	387	119
Number of TE-cassettes overlapping 5' UTR, CDS, and 3' UTR	189	59	74

One could note that the number of rat transcripts is much lower than the number of human and mouse transcripts, which is due to the poorer annotation of the rat genome at the time of this analysis (2003). Also, the number of sequences containing TE fragments is the highest in human, which can be attributed to richer collection of human TEs as well as to the presence of alternatively spliced Alu exons which are present in

many human genes (Sorek, *et al.* 2002). A summary of TE cassettes found in the three transcriptomes analyzed is presented in Table 5.2.

Table 5.2: Dimension of TE-cassettes found in three mammalian transcriptomes.

	Human			**Mouse**			**Rat**		
	5' UTR	CDS	3' UTR	5' UTR	CDS	3' UTR	5' UTR	CDS	3' UTR
Minimum	18	16	11	11	15	11	12	16	11
Maximum	2520	2131	2844	2580	4812	2228	1715	5471	1191
Mean	165	151	202	128	319	168	125	296	132
Median	124	102	163	109	150	134	96	123	114

The content of the database can be queried through regular SQL queries or through a simple web interface implemented using Perl CGI scripts, which is available at http://warta.bio.psu.edu/ScrapYard/database.html. The current web interface allows for searching specific TE-containing transcripts using accession, gi numbers, or general description terms (e.g. “survivin”). It also allows browsing a collection of TE-containing transcripts that satisfy a few simple criteria, such as species, TE class, location of the TE relative to the CDS as shown in Figure 5.2. In either case, a list of matching transcripts is displayed, after which the user could get more details by clicking on the desired transcripts. Along with details from the Genbank record, alignments between transcript and consensus TE sequences are displayed, as well as a visual representation of the TE location on the transcript sequence, as shown in Figure 5.3 for the human survivin-beta (accession: AB028869.1, gi: 7416052).

Figure 5.2: TE-containing transcripts could be found using a few simple criteria using the web interface of the SYDB at http://warta.bio.psu.edu/cgi-bin/ScrapYard/browse.pl.

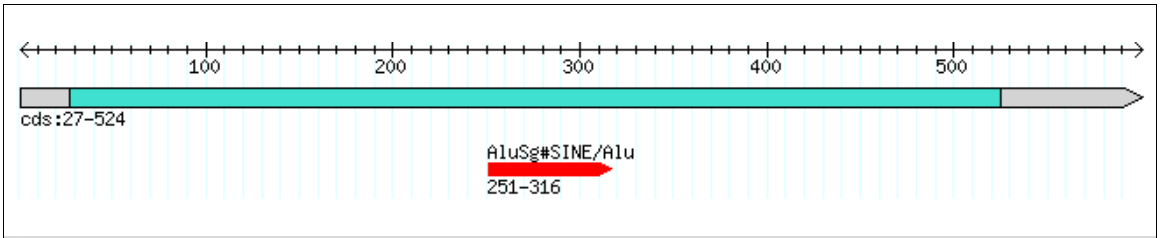

Figure 5.3: Visual representation of the human survivin-beta transcript as displayed by the web interface of the SYDB. The transcript (accession: AB028869.1, gi: 7416052) is 600 bp long, has the CDS annotated between 27-524, and contains an AluSg fragment in sense orientation between coordinates 251-316 as determined by RepeatMasker.

5.3 Discussion

The SYDB is a highly specific database, containing annotations of TE cassettes located in vertebrate transcripts (currently data from human, mouse, and rat are available). It offers direct access to TE-containing transcripts, simplifying specific searches for such sequences by eliminating the masking step which can be computationally expensive. Its use is targeted toward specific projects, mostly involving exaptation events in any region of the transcript. Originally, the SYDB was conceived to provide preliminary data for investigating exaptation events into protein coding sequences (CDS), such as for finding candidates for inferring the impact of Alu exaptation events of the host protein by the means of homology modeling (see Chapter 3.2). The SYDB was also successfully used for finding candidates for successful MIR exaptation into protein coding sequences, which were further investigated for EST support, and conservation in several mammalian species, which lead to the identification of a constitutively spliced MIR exon in the ZNF639 gene (Krull*, et al.* 2007).

In recent years, a number of other resources that can offer similar information have appeared, such as the general UCSC Genome Browser (Karolchik*, et al.* 2003), and the more specific *Alu*Gene (Dagan*, et al.* 2004) databases. Unlike the SYDB, these are all based on TE identification at the genome level. In the case of SYDB, however, the identification of TEs was done at the level of transcripts where intronic sequences are missing. This allows for identification of TEs that span one or more exons (a few examples are shown in Table 2.1), which could provide the first clues for gain of intron events, and thus offering a conceptual advantage to SYDB over other similar resources.

Using exon-spanning TEs to find gain of intron events is impossible to use with TEs identified at genomic level, and until present gain of intron events have not been documented in mammals. At the same time, finding TEs on transcripts could present some disadvantages as well. If a short fragment, normally up to 20 nt long, from a complete TE copy is exapted into an exon, would be missed by searching for TEs at transcript level, but would be otherwise found by searching for TEs at genome level. While this is a potential problem, from the exaptation point of view, which is the focus of this database, short sequences originating from TEs could be assimilated to anonymous DNA sequences. What could be potentially interesting and missed by this approach is the occurrence of splice sites in TE sequences, which would need to be documented at genome level.

In its current incarnation, the SYDB represents a database with a big potential for expansion. Adding more species would offer a wider evolutionary space where important evolutionary events, such as gain of introns or exaptation into protein coding regions, could be traced. Equally importantly, it offers a tool where manually curated cases, such as those presented in Chapter 2, could replace pure computational predictions. Individual analyses of specific cases could allow for adding additional information to the database, such as 3D models obtained thorough homology modeling for suitable candidates (see Figure 5.1). Gene ontology and interconnectivity with other resources, such as the UCSC Genome Browser, will make the SYDB more appealing to researchers, but regardless, the simplicity of use and conceptual advantages should offer the SYDB an important place in a TE researcher's tool bag.

Bibliography

1. Ambrosini, G., Adida, C., and Altieri, D.C. (1997) A novel anti-apoptosis gene, survivin, expressed in cancer and lymphoma. *Nat. Med.* 3(8): 917-921

2. Andersen, J.N., Jansen, P.G., Echwald, S.M., Mortensen, O.H., Fukada, T.*, et al.* (2004) A genomic perspective on protein tyrosine phosphatases: gene structure, pseudogenes, and genetic disease linkage. *FASEB J.* 18(1): 8-30

3. Arcot, S.S., Adamson, A.W., Risch, G.W., LaFleur, J., Robichaux, M.B.*, et al.* (1998) High-resolution cartography of recently integrated human chromosome 19-specific Alu fossils. *J. Mol. Biol.* 281(5): 843-856

4. Athanikar, J.N., Badge, R.M., and Moran, J.V. (2004) A YY1-binding site is required for accurate human LINE-1 transcription initiation. *Nucleic Acids Res.* 32(13): 3846-3855

5. Avery, O.T., MacLeod, C.M., and McCarty, M. (1944) Studies on the chemical nature of the substance inducing transformation of Pneumococcal types. *J. Exp. Med.* 79: 137-159

6. Bailey, J.A., Liu, G., and Eichler, E.E. (2003) An Alu transposition model for the origin and expansion of human segmental duplications. *Am. J. Hum. Genet.* 73(4): 823-834

7. Bairoch, A., and Apweiler, R. (1997) The SWISS-PROT protein sequence database: its relevance to human molecular medical research. *J. Mol. Med.* 75(5): 312-316

8. Bao, Z., and Eddy, S.R. (2002) Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res.* 12(8): 1269-1276

9. Barford, D., Flint, A.J., and Tonks, N.K. (1994) Crystal structure of human protein tyrosine phosphatase 1B. *Science* 263(5152): 1397-1404

10. Bejerano, G., Lowe, C.B., Ahituv, N., King, B., Siepel, A.*, et al.* (2006) A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* 441(7089): 87-90

11. Beltran, A.C., Dawson, P.E., and Gottesfeld, J.M. (2005) Role of DNA sequence in the binding specificity of synthetic basic-helix-loop-helix domains. *Chembiochem* 6(1): 104-113

12. Bendtsen, J.D., Nielsen, H., von Heijne, G., and Brunak, S. (2004) Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* 340(4): 783-795

13. Blanchard, R.L., Freimuth, R.R., Buck, J., Weinshilboum, R.M., and Coughtrie, M.W. (2004) A proposed nomenclature system for the cytosolic sulfotransferase (SULT) superfamily. *Pharmacogenetics* 14(3): 199-211

14. Brenner, S.E. (2000) Target selection for structural genomics. *Nat. Struct. Biol.* 7 Suppl.: 967-969

15. Brosius, J. (1991) Retroposons--seeds of evolution. *Science* 251(4995): 753

16. Brownell, E., Mittereder, N., and Rice, N.R. (1989) A human rel proto-oncogene cDNA containing an Alu fragment as a potential coding exon. *Oncogene* 4(7): 935-942

17. Burch, J.B., Davis, D.L., and Haas, N.B. (1993) Chicken repeat 1 elements contain a pol-like open reading frame and belong to the non-long terminal repeat class of retrotransposons. *Proc. Natl. Acad. Sci. U. S. A.* 90(17): 8199-8203

18. Bureau, T.E., and Wessler, S.R. (1992) Tourist: a large family of small inverted repeat elements frequently associated with maize genes. *Plant Cell* 4(10): 1283-1294

19. Bustamante, C.D., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M.T.*, et al.* (2005) Natural selection on protein-coding genes in the human genome. *Nature* 437(7062): 1153-1157

20. Cameron, R.A., Mahairas, G., Rast, J.P., Martinez, P., Biondi, T.R.*, et al.* (2000) A sea urchin genome project: sequence scan, virtual map, and additional resources. *Proc. Natl. Acad. Sci. U. S. A.* 97(17): 9514-9518

21. Capy, P. (1998) Classification of transposable elements. In *Molecular biology intelligence unit: dynamics and evolution of transposable elements* (Capy, P.*, et al.*, eds), 37-52, Landes Biosciences, Georgetown

22. Capy, P. (2005) Classification and nomenclature of retrotransposable elements. *Cytogenet. Genome Res.* 110(1-4): 457-461

23. Caras, I.W., Davitz, M.A., Rhee, L., Weddell, G., Martin, D.W., Jr.*, et al.* (1987) Cloning of decay-accelerating factor suggests novel use of splicing to generate two proteins. *Nature* 325(6104): 545-549

24. Chalei, M.B., and Korotkov, E.V. (2001) [Evolution of MIR-repeats in coding regions of the human genome]. *Mol. Biol. (Mosk.)* 35(6): 1023-1031

25. Chantalat, L., Skoufias, D.A., Kleman, J.P., Jung, B., Dideberg, O., *et al.* (2000) Crystal structure of human survivin reveals a bow tie-shaped dimer with two unusual alpha-helical extensions. *Mol. Cell* 6(1): 183-189

26. Charbonneau, H., Tonks, N.K., Kumar, S., Diltz, C.D., Harrylock, M., *et al.* (1989) Human placenta protein-tyrosine-phosphatase: amino acid sequence and relationship to a family of receptor-like proteins. *Proc. Natl. Acad. Sci. U.S.A.* 86(14): 5252-5256

27. Chen, F.C., and Li, W.H. (2001) Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am. J. Hum. Genet.* 68(2): 444-456

28. Chu, W.M., Ballard, R., Carpick, B.W., Williams, B.R., and Schmid, C.W. (1998) Potential Alu function: regulation of the activity of double-stranded RNA-activated kinase PKR. *Mol. Cell. Biol.* 18(1): 58-68

29. Clarimon, J., Andres, A.M., Bertranpetit, J., and Comas, D. (2004) Comparative analysis of Alu insertion sequences in the APP 5' flanking region in humans and other primates. *J. Mol. Evol.* 58(6): 722-731

30. Comfort, N.C. (1995) Two genes, no enzyme: a second look at Barbara McClintock and the 1951 Cold Spring Harbor Symposium. *Genetics* 140(4): 1161-1166

31. Copeland, P.R. (2003) Regulation of gene expression by stop codon recoding: selenocysteine. *Gene* 312: 17-25

32. Copeland, P.R. (2005) Making sense of nonsense: the evolution of selenocysteine usage in proteins. *Genome Biol.* 6(6): 221

33. Cordaux, R., Lee, J., Dinoso, L., and Batzer, M.A. (2006) Recently integrated Alu retrotransposons are essentially neutral residents of the human genome. *Gene* 373: 138-144

34. Cordaux, R., Udit, S., Batzer, M.A., and Feschotte, C. (2006) Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. *Proc. Natl. Acad. Sci. U. S. A.* 103(21): 8101-8106

35. Costas, J. (2001) Evolutionary dynamics of the human endogenous retrovirus family HERV-K inferred from full-length proviral genomes. *J. Mol. Evol.* 53(3): 237-243

36. Craig, N.L. (2002) Mobile DNA: an introduction. In *Mobile DNA II* (Craig, N.L., *et al.*, eds), 3-11, ASM Press, Washington, D.C.

37. Craig, N.L., Craigie, R., Gellert, M., and Lambowitz, A.M. (2002) *Mobile DNA II*. ASM Press, Washington, D.C.

38. Dagan, T., Sorek, R., Sharon, E., Ast, G., and Graur, D. (2004) AluGene: a database of Alu elements incorporated within protein-coding genes. *Nucleic Acids Res.* 32(Database issue): D489-492

39. Davidson, E.H., Hough, B.R., Amenson, C.S., and Britten, R.J. (1973) General interspersion of repetitive with non-repetitive sequence elements in the DNA of Xenopus. *J. Mol. Biol.* 77(1): 1-23

40. Dawson, A., Hartswood, E., Paterson, T., and Finnegan, D.J. (1997) A LINE-like transposable element in Drosophila, the I factor, encodes a protein with properties similar to those of retroviral nucleocapsids. *EMBO J.* 16(14): 4448-4455

41. Dear, T.N., Moller, A., and Boehm, T. (1999) CAPN11: A calpain with high mRNA levels in testis and located on chromosome 6. *Genomics* 59(2): 243-247

42. Deininger, P.L., Batzer, M.A., Hutchison, C.A., 3rd, and Edgell, M.H. (1992) Master genes in mammalian repetitive DNA amplification. *Trends Genet.* 8(9): 307-311

43. Deininger, P.L., and Batzer, M.A. (1999) Alu repeats and human disease. *Mol. Genet. Metab.* 67(3): 183-193

44. Deininger, P.L., and Batzer, M.A. (2002) Mammalian retroelements. *Genome Res.* 12(10): 1455-1465

45. Deininger, P.L., Moran, J.V., Batzer, M.A., and Kazazian, H.H., Jr. (2003) Mobile elements and mammalian genome evolution. *Curr. Opin. Genet. Dev.* 13(6): 651-658

46. Deragon, J.M., and Capy, P. (2000) Impact of transposable elements on the human genome. *Ann. Med.* 32(4): 264-273

47. Dewannieux, M., Esnault, C., and Heidmann, T. (2003) LINE-mediated retrotransposition of marked Alu sequences. *Nat. Genet.* 35(1): 41-48

48. Dombrovski, L., Dong, A., Bochkarev, A., and Plotnikov, A.N. (2006) Crystal structures of human sulfotransferases SULT1B1 and SULT1C1 complexed with the cofactor product adenosine-3'- 5'-diphosphate (PAP). *Proteins* 64(4): 1091-1094

49. Doolittle, W.F., and Sapienza, C. (1980) Selfish genes, the phenotype paradigm and genome evolution. *Nature* 284(5757): 601-603

50. Dunn, L.C. (1965) *A short history of genetics; the development of some of the main lines of thought, 1864-1939*. McGraw-Hill, New York,

51. Eichler, E.E. (2001) Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends Genet.* 17(11): 661-669

52. Ferrigno, O., Virolle, T., Djabari, Z., Ortonne, J.P., White, R.J.*, et al.* (2001) Transposable B2 SINE elements can provide mobile RNA polymerase II promoters. *Nat. Genet.* 28(1): 77-81

53. Feschotte, C., Zhang, X., and Wessler, S.R. (2002) Miniature inverted-repeat transposable elements and their relationship to established DNA transposons. In *Mobile DNA II* (Craig, N.L.*, et al.*, eds), 1147-1158, AMS Press, Washington, D.C.

54. Feschotte, C., and Pritham, E.J. (2005) Non-mammalian c-integrases are encoded by giant transposable elements. *Trends Genet.* 21(10): 551-552

55. Finnegan, D.J. (1989) Eukaryotic transposable elements and genome evolution. *Trends Genet.* 5(4): 103-107

56. Fisher, R.A. (1922) On the interpretation of $\chi 2$ from contingency tables, and the calculation of P. *J. R. Stat. Soc.* 85: 87-94

57. Freimuth, R.R., Raftogianis, R.B., Wood, T.C., Moon, E., Kim, U.J.*, et al.* (2000) Human sulfotransferases SULT1C1 and SULT1C2: cDNA characterization, gene cloning, and chromosomal localization. *Genomics* 65(2): 157-165

58. Frith, M.C., Forrest, A.R., Nourbakhsh, E., Pang, K.C., Kai, C.*, et al.* (2006) The abundance of short proteins in the mammalian proteome. *PLoS Genet.* 2(4): e52

59. Gerber, A., O'Connell, M.A., and Keller, W. (1997) Two forms of human double-stranded RNA-specific editase 1 (hRED1) generated by the insertion of an Alu cassette. *RNA* 3(5): 453-463

60. Giardine, B., Riemer, C., Hardison, R.C., Burhans, R., Elnitski, L.*, et al.* (2005) Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.* 15(10): 1451-1455

61. Gibbs, R.A., Rogers, J., Katze, M.G., Bumgarner, R., Weinstock, G.M.*, et al.* (2007) Evolutionary and biomedical insights from the rhesus macaque genome. *Science* 316(5822): 222-234

62. Gilbert, N., and Labuda, D. (1999) CORE-SINEs: eukaryotic short interspersed retroposing elements with common sequence motifs. *Proc. Natl. Acad. Sci. U. S. A.* 96(6): 2869-2874

63. Goldschmidt, R.B. (1938) *Physiological genetics*. McGraw-Hill Book Company inc., New York, London,

64. Gotea, V., and Makalowski, W. (2006) Do transposable elements really contribute to proteomes? *Trends Genet.* 22(5): 260-267

65. Gothel, S.F., and Marahiel, M.A. (1999) Peptidyl-prolyl cis-trans isomerases, a superfamily of ubiquitous folding catalysts. *Cell. Mol. Life Sci.* 55(3): 423-436

66. Graham, D., Neufeld, B., Davidson, E., and Britten, R. (1974) Interspersion of repetitive and non-repetitive DNA sequence in the sea urchin genome. *Cell* 1(3): 127-137

67. Grossman, W.J., Revell, P.A., Lu, Z.H., Johnson, H., Bredemeyer, A.J.*, et al.* (2003) The orphan granzymes of humans and mice. *Curr. Opin. Immunol.* 15(5): 544-552

68. Haas, N.B., Grabowski, J.M., Sivitz, A.B., and Burch, J.B. (1997) Chicken repeat 1 (CR1) elements, which define an ancient family of vertebrate non-LTR retrotransposons, contain two closely spaced open reading frames. *Gene* 197(1-2): 305-309

69. Hamdi, H.K., Nishio, H., Tavis, J., Zielinski, R., and Dugaiczyk, A. (2000) Alu-mediated phylogenetic novelties in gene regulation and development. *J. Mol. Biol.* 299(4): 931-939

70. Han, K., Konkel, M.K., Xing, J., Wang, H., Lee, J.*, et al.* (2007) Mobile DNA in Old World monkeys: a glimpse through the rhesus macaque genome. *Science* 316(5822): 238-240

71. Hasegawa, S.L., Sloan, J.H., Reith, W., Mach, B., and Boss, J.M. (1991) Regulatory factor-X binding to mutant HLA-DRA promoter sequences. *Nucleic Acids Res.* 19(6): 1243-1249

72. Hassoun, H., Coetzer, T.L., Vassiliadis, J.N., Sahr, K.E., Maalouf, G.J.*, et al.* (1994) A novel mobile element inserted in the alpha spectrin gene: spectrin dayton. A truncated alpha spectrin associated with hereditary elliptocytosis. *J. Clin. Invest.* 94(2): 643-648

73. Hayashi, Y., Sakata, H., Makino, Y., Urabe, I., and Yomo, T. (2003) Can an arbitrary sequence evolve towards acquiring a biological function? *J. Mol. Evol.* 56(2): 162-168

74. Hickey, D.A. (1982) Selfish DNA: a sexually-transmitted nuclear parasite. *Genetics* 101(3-4): 519-531

75. Hilgard, P., Huang, T., Wolkoff, A.W., and Stockert, R.J. (2002) Translated Alu sequence determines nuclear localization of a novel catalytic subunit of casein kinase 2. *Am. J. Physiol. Cell Physiol.* 283(2): C472-483

76. Hilleren, P., and Parker, R. (1999) Mechanisms of mRNA surveillance in eukaryotes. *Annu. Rev. Genet.* 33: 229-260

77. Hink-Schauer, C., Estebanez-Perpina, E., Kurschus, F.C., Bode, W., and Jenne, D.E. (2003) Crystal structure of the apoptosis-inducing human granzyme A dimer. *Nat. Struct. Biol.* 10(7): 535-540

78. Hoenicka, J., Arrasate, M., de Yebenes, J.G., and Avila, J. (2002) A two-hybrid screening of human Tau protein: interactions with Alu-derived domain. *Neuroreport* 13(3): 343-349

79. Holmes, S.E., Singer, M.F., and Swergold, G.D. (1992) Studies on p40, the leucine zipper motif-containing protein encoded by the first open reading frame of an active human LINE-1 transposable element. *J. Biol. Chem.* 267(28): 19765-19768

80. Houck, C.M., Rinehart, F.P., and Schmid, C.W. (1979) A ubiquitous family of repeated DNA sequences in the human genome. *J. Mol. Biol.* 132(3): 289-306

81. Imanishi, T., Itoh, T., Suzuki, Y., O'Donovan, C., Fukuchi, S.*, et al.* (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol.* 2(6): 856-875

82. Jacobson, A., and Peltz, S.W. (1996) Interrelationships of the pathways of mRNA decay and translation in eukaryotic cells. *Annu. Rev. Biochem.* 65: 693-739

83. Jekely, G., and Friedrich, P. (1999) The evolution of the calpain family as reflected in paralogous chromosome regions. *J. Mol. Evol.* 49(2): 272-281

84. Jordan, I.K., Rogozin, I.B., Glazko, G.V., and Koonin, E.V. (2003) Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet.* 19(2): 68-72

85. Jurka, J. (1994) Approaches to identification and analysis of interspersed repetitive DNA sequences. In *Automated DNA sequencing and analysis* (Adams, M.D.*, et al.*, eds), 294-298, Academic Press Incorporated, San Diego

86. Jurka, J., Zietkiewicz, E., and Labuda, D. (1995) Ubiquitous mammalian-wide interspersed repeats (MIRs) are molecular fossils from the mesozoic era. *Nucleic Acids Res.* 23(1): 170-175

87. Jurka, J., Klonowski, P., Dagman, V., and Pelton, P. (1996) CENSOR--a program for identification and elimination of repetitive elements from DNA sequences. *Comput. Chem.* 20(1): 119-121

88. Jurka, J. (1998) Repeats in genomic DNA: mining and meaning. *Curr. Opin. Struct. Biol.* 8(3): 333-337

89. Jurka, J. (2000) Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet.* 16(9): 418-420

90. Kajikawa, M., Ohshima, K., and Okada, N. (1997) Determination of the entire sequence of turtle CR1: the first open reading frame of the turtle CR1 element encodes a protein with a novel zinc finger motif. *Mol. Biol. Evol.* 14(12): 1206-1217

91. Kapitonov, V.V., and Jurka, J. (2001) Rolling-circle transposons in eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.* 98(15): 8714-8719

92. Kapitonov, V.V., and Jurka, J. (2003) The esterase and PHD domains in CR1-like non-LTR retrotransposons. *Mol. Biol. Evol.* 20(1): 38-46

93. Kapitonov, V.V., and Jurka, J. (2006) Self-synthesizing DNA transposons in eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.* 103(12): 4540-4545

94. Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A.*, et al.* (2003) The UCSC Genome Browser Database. *Nucleic Acids Res.* 31(1): 51-54

95. Kazazian, H.H., Jr. (1998) Mobile elements and disease. *Curr. Opin. Genet. Dev.* 8(3): 343-350

96. Kazazian, H.H., Jr. (2004) Mobile elements: drivers of genome evolution. *Science* 303(5664): 1626-1632

97. Kel, A.E., Gossling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O.V.*, et al.* (2003) MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.* 31(13): 3576-3579

98. Keller, E.F. (1983) *A feeling for the organism: the life and work of Barbara McClintock*. W.H. Freeman, New York

99. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H.*, et al.* (2002) The human genome browser at UCSC. *Genome Res.* 12(6): 996-1006

100. Kitayner, M., Rozenberg, H., Rabinovich, D., and Shakked, Z. (2005) Structures of the DNA-binding site of Runt-domain transcription regulators. *Acta Crystallogr. D Biol. Crystallogr.* 61(Pt 3): 236-246

101. Korenberg, J.R., and Rykowski, M.C. (1988) Human genome organization: Alu, lines, and the molecular structure of metaphase chromosome bands. *Cell* 53(3): 391-400

102. Kouranov, A., Xie, L., de la Cruz, J., Chen, L., Westbrook, J.*, et al.* (2006) The RCSB PDB information portal for structural genomics. *Nucleic Acids Res.* 34(Database issue): D302-305

103. Kreahling, J., and Graveley, B.R. (2004) The origins and implications of *Alu*ternative splicing. *Trends Genet.* 20(1): 1-4

104. Kriegs, J.O., Churakov, G., Jurka, J., Brosius, J., and Schmitz, J. (2007) Evolutionary history of 7SL RNA-derived SINEs in Supraprimates. *Trends Genet.* 23(4): 158-161

105. Kriegs, J.O., Schmitz, J., Makalowski, W., and Brosius, J. (2005) Does the AD7c-NTP locus encode a protein? *Biochim. Biophys. Acta* 1727(1): 1-4

106. Krull, M., Brosius, J., and Schmitz, J. (2005) Alu-SINE exonization: en route to protein-coding function. *Mol. Biol. Evol.* 22(8): 1702-1711

107. Krull, M., Petrusma, M., Makalowski, W., Brosius, J., and Schmitz, J. (2007) Functional persistence of exonized Mammalian-wide Interspersed Repeat elements in protein-coding genes. *Genome Res.* doi:10.1101/gr.6320607

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

109. Kryukov, G.V., Castellano, S., Novoselov, S.V., Lobanov, A.V., Zehtab, O.*, et al.* (2003) Characterization of mammalian selenoproteomes. *Science* 300(5624): 1439-1443

110. Kumar, S., and Hedges, S.B. (1998) A molecular timescale for vertebrate evolution. *Nature* 392(6679): 917-920

111. Kumar, S., Tamura, K., and Nei, M. (2004) MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief. Bioinform.* 5(2): 150-163

112. Lai, F., Chen, C.X., Carter, K.C., and Nishikura, K. (1997) Editing of glutamate receptor B subunit ion channel RNAs by four alternatively spliced DRADA2 double-stranded RNA adenosine deaminases. *Mol. Cell. Biol.* 17(5): 2413-2424

113. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C.*, et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409(6822): 860-921

114. Landry, J.R., Medstrand, P., and Mager, D.L. (2001) Repetitive elements in the 5' untranslated region of a human zinc-finger gene modulate transcription and translation efficiency. *Genomics* 76(1-3): 110-116

115. Lareau, L.F., Inada, M., Green, R.E., Wengrod, J.C., and Brenner, S.E. (2007) Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature* 446(7138): 926-929

116. Lee, H.J., Sorimachi, H., Jeong, S.Y., Ishiura, S., and Suzuki, K. (1998) Molecular cloning and characterization of a novel tissue-specific calpain predominantly expressed in the digestive tract. *Biol. Chem.* 379(2): 175-183

117. Leib-Mosch, C., Haltmeier, M., Werner, T., Geigl, E.M., Brack-Werner, R.*, et al.* (1993) Genomic distribution and transcription of solitary HERV-K LTRs. *Genomics* 18(2): 261-269

118. Lev-Maor, G., Sorek, R., Shomron, N., and Ast, G. (2003) The birth of an alternatively spliced exon: 3' splice-site selection in Alu exons. *Science* 300(5623): 1288-1291

119. Li, J., Williams, B.L., Haire, L.F., Goldberg, M., Wilker, E.*, et al.* (2002) Structural and functional versatility of the FHA domain in DNA-damage signaling by the tumor suppressor kinase Chk2. *Mol. Cell* 9(5): 1045-1054

120. Li, T.H., and Schmid, C.W. (2004) Alu's dimeric consensus sequence destabilizes its transcripts. *Gene* 324: 191-200

121. Li, W.H., Gu, Z., Wang, H., and Nekrutenko, A. (2001) Evolutionary analyses of the human genome. *Nature* 409(6822): 847-849

122. Li, X.N., Shu, Q., Su, J.M., Adesina, A.M., Wong, K.K.*, et al.* (2007) Differential expression of survivin splice isoforms in medulloblastomas. *Neuropathol. Appl. Neurobiol.* 33(1): 67-76

123. Liew, C.K., Simpson, R.J., Kwan, A.H., Crofts, L.A., Loughlin, F.E.*, et al.* (2005) Zinc fingers as protein recognition motifs: structural basis for the GATA-1/friend of GATA interaction. *Proc. Natl. Acad. Sci. U. S. A.* 102(3): 583-588

124. Lindblad-Toh, K., Wade, C.M., Mikkelsen, T.S., Karlsson, E.K., Jaffe, D.B.*, et al.* (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* 438(7069): 803-819

125. Linhoff, M.W., Wright, K.L., and Ting, J.P. (1997) CCAAT-binding factor NF-Y and RFX are required for in vivo assembly of a nucleoprotein complex that spans 250

base pairs: the invariant chain promoter as a model. *Mol. Cell Biol.* 17(8): 4589-4596

126. Liu, W.M., Chu, W.M., Choudary, P.V., and Schmid, C.W. (1995) Cell stress and translational inhibitors transiently increase the abundance of mammalian SINE transcripts. *Nucleic Acids Res.* 23(10): 1758-1765

127. Loflin, J., Lopez, N., Whanger, P.D., and Kioussi, C. (2006) Selenoprotein W during development and oxidative stress. *J. Inorg. Biochem.* 100(10): 1679-1684

128. Lorenc, A., and Makalowski, W. (2003) Transposable elements and vertebrate protein diversity. *Genetica* 118(2-3): 183-191

129. Lovell, S.C. (2003) Are non-functional, unfolded proteins ('junk proteins') common in the genome? *FEBS Lett.* 554(3): 237-239

130. Lower, R., Lower, J., and Kurth, R. (1996) The viruses in all of us: characteristics and biological significance of human endogenous retrovirus sequences. *Proc. Natl. Acad. Sci. U. S. A.* 93(11): 5177-5184

131. Lu, C., Qiu, F., Zhou, H., Peng, Y., Hao, W.*, et al.* (2006) Identification and characterization of selenoprotein K: an antioxidant in cardiomyocytes. *FEBS Lett.* 580(22): 5189-5197

132. Lundwall, A.B., Wetsel, R.A., Kristensen, T., Whitehead, A.S., Woods, D.E.*, et al.* (1985) Isolation and sequence analysis of a cDNA clone encoding the fifth complement component. *J. Biol. Chem.* 260(4): 2108-2112

133. Macbeth, M.R., Schubert, H.L., Vandemark, A.P., Lingam, A.T., Hill, C.P.*, et al.* (2005) Inositol hexakisphosphate is bound in the ADAR2 core and required for RNA editing. *Science* 309(5740): 1534-1539

134. Mahotka, C., Krieg, T., Krieg, A., Wenzel, M., Suschek, C.V.*, et al.* (2002) Distinct in vivo expression patterns of survivin splice variants in renal cell carcinomas. *Int. J. Cancer* 100(1): 30-36

135. Mahotka, C., Wenzel, M., Springer, E., Gabbert, H.E., and Gerharz, C.D. (1999) Survivin-deltaEx3 and survivin-2B: two novel splice variants of the apoptosis inhibitor survivin with different antiapoptotic properties. *Cancer Res.* 59(24): 6097-6102

136. Makalowski, W., Mitchell, G.A., and Labuda, D. (1994) Alu sequences in the coding regions of mRNA: a source of protein variability. *Trends Genet.* 10(6): 188-193

137. Makalowski, W. (1995) SINEs as a genomic scrap yard: an essay on genomic evolution. In *The impact of short interspersed elements (SINEs) on the Hpst genome* (Maraia, R.J., ed), 81-104, RG Landes, Austin, TX

138. Makalowski, W. (2000) Genomic scrap yard: how genomes utilize all that junk. *Gene* 259(1-2): 61-67

139. Makalowski, W. (2001) The human genome structure and organization. *Acta Biochim. Pol.* 48(3): 587-598

140. Malik, H.S., Burke, W.D., and Eickbush, T.H. (1999) The age and evolution of non-LTR retrotransposable elements. *Mol. Biol. Evol.* 16(6): 793-805

141. Malik, H.S., Henikoff, S., and Eickbush, T.H. (2000) Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. *Genome Res.* 10(9): 1307-1318

142. Manning, J.E., Schmid, C.W., and Davidson, N. (1975) Interspersion of repetitive and nonrepetitive DNA sequences in the Drosophila melanogaster genome. *Cell* 4(2): 141-155

143. McCarthy, E.M., and McDonald, J.F. (2003) LTR_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics* 19(3): 362-367

144. McClintock, B. (1951) Chromosome organization and genic expression. *Cold Spring Harb. Symp. Quant. Biol.* 16: 13-47

145. McClintock, B. (1956) Controlling elements and the gene. *Cold Spring Harb. Symp. Quant. Biol.* 21: 197-216

146. McClure, M.A. (1991) Evolution of retroposons by acquisition or deletion of retrovirus-like genes. *Mol. Biol. Evol.* 8(6): 835-856

147. McDonald, J.H., and Kreitman, M. (1991) Adaptive protein evolution at the Adh locus in Drosophila. *Nature* 351(6328): 652-654

148. Mitchell, G.A., Labuda, D., Fontaine, G., Saudubray, J.M., Bonnefont, J.P.*, et al.* (1991) Splice-mediated insertion of an Alu sequence inactivates ornithine delta-aminotransferase: a role for Alu elements in human mutation. *Proc. Natl. Acad. Sci. U. S. A.* 88(3): 815-819

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

150. Moghadaszadeh, B., Petit, N., Jaillard, C., Brockington, M., Roy, S.Q.*, et al.* (2001) Mutations in SEPN1 cause congenital muscular dystrophy with spinal rigidity and restrictive respiratory syndrome. *Nat. Genet.* 29(1): 17-18

151. Nei, M. (1969) Gene duplication and nucleotide substitution in evolution. *Nature* 221(175): 40-42

152. Nei, M., and Gojobori, T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* 3(5): 418-426

153. Nekrutenko, A., and Li, W.H. (2001) Transposable elements are found in a large number of human protein-coding genes. *Trends Genet.* 17(11): 619-621

154. Ni, J.Z., Grate, L., Donohue, J.P., Preston, C., Nobida, N.*, et al.* (2007) Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. *Genes Dev.* 21(6): 708-718

155. Nirenberg, M.W., and Matthaei, J.H. (1961) The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides. *Proc. Natl. Acad. Sci. U. S. A.* 47: 1588-1602

156. Nirenberg, M., Caskey, T., Marshall, R., Brimacombe, R., Kellogg, D.*, et al.* (1966) The RNA code and protein synthesis. *Cold Spring Harb. Symp. Quant. Biol.* 31: 11-24

157. Nowak, R. (1994) Mining treasures from 'junk DNA'. *Science* 263(5147): 608-610

158. Oei, S.L., Babich, V.S., Kazakov, V.I., Usmanova, N.M., Kropotov, A.V.*, et al.* (2004) Clusters of regulatory signals for RNA polymerase II transcription associated with Alu family repeats and CpG islands in human promoters. *Genomics* 83(5): 873-882

159. Oh, Y.S., Lee, S., Won, C., and Warnock, D.G. (2001) An Alu cassette in the human epithelial sodium channel. *Biochim. Biophys. Acta* 1520(1): 94-98

160. Ohno, S. (1970) *Evolution by gene duplication*. Springer-Verlag, Berlin

161. Ohno, S. (1972) So much 'junk' DNA in our genome. *Brookhav. Symp. Biol.* 23: 366-370

162. Ono, K., Suga, H., Iwabe, N., Kuma, K., and Miyata, T. (1999) Multiple protein tyrosine phosphatases in sponges and explosive gene duplication in the early evolution of animals before the parazoan-eumetazoan split. *J. Mol. Evol.* 48(6): 654-662

163. Ono, M. (1986) Molecular cloning and long terminal repeat sequences of human endogenous retrovirus genes related to types A and B retrovirus genes. *J. Virol.* 58(3): 937-944

164. Ono, M., Kawakami, M., and Takezawa, T. (1987) A novel human nonviral retroposon derived from an endogenous retrovirus. *Nucleic Acids Res.* 15(21): 8725-8737

165. Ono-Koyanagi, K., Suga, H., Katoh, K., and Miyata, T. (2000) Protein tyrosine phosphatases from amphioxus, hagfish, and ray: divergence of tissue-specific isoform genes in the early evolution of vertebrates. *J. Mol. Evol.* 50(3): 302-311

166. Orgel, L.E., and Crick, F.H. (1980) Selfish DNA: the ultimate parasite. *Nature* 284(5757): 604-607

167. Ostertag, E.M., Goodier, J.L., Zhang, Y., and Kazazian, H.H., Jr. (2003) SVA elements are nonautonomous retrotransposons that cause disease in humans. *Am. J. Hum. Genet.* 73(6): 1444-1451

168. Ovchinnikov, I., Troxel, A.B., and Swergold, G.D. (2001) Genomic characterization of recent human LINE-1 insertions: evidence supporting random insertion. *Genome Res.* 11(12): 2050-2058

169. Papp, L.V., Lu, J., Holmgren, A., and Khanna, K.K. (2007) From selenium to selenoproteins: synthesis, identity, and their role in human health. *Antioxid. Redox Signal.* 9(7): 775-806

170. Paulson, K.E., Matera, A.G., Deka, N., and Schmid, C.W. (1987) Transcription of a human transposon-like sequence is usually directed by other promoters. *Nucleic Acids Res.* 15(13): 5199-5215

171. Pavlicek, A., Clay, O., and Bernardi, G. (2002) Transposable elements encoding functional proteins: pitfalls in unprocessed genomic data? *FEBS Lett.* 523(1-3): 252-253

172. Peng, K., Obradovic, Z., and Vucetic, S. (2004) Exploring bias in the Protein Data Bank using contrast classifiers. *Pac. Symp. Biocomput.* 9: 435-446

173. Petit, N., Lescure, A., Rederstorff, M., Krol, A., Moghadaszadeh, B.*, et al.* (2003) Selenoprotein N: an endoplasmic reticulum glycoprotein with an early developmental expression pattern. *Hum. Mol. Genet.* 12(9): 1045-1053

174. Pikkarainen, S., Tokola, H., Kerkela, R., and Ruskoaho, H. (2004) GATA transcription factors in the developing and adult heart. *Cardiovasc. Res.* 63(2): 196-207

175. Poulter, R., Butler, M., and Ormandy, J. (1999) A LINE element from the pufferfish (fugu) Fugu rubripes which shows similarity to the CR1 family of non-LTR retrotransposons. *Gene* 227(2): 169-179

176. Pradet-Balade, B., Boulme, F., Beug, H., Mullner, E.W., and Garcia-Sanz, J.A. (2001) Translation control: bridging the gap between genomics and proteomics? *Trends Biochem. Sci.* 26(4): 225-229

177. Price, A.L., Eskin, E., and Pevzner, P.A. (2004) Whole-genome analysis of Alu repeat elements reveals complex evolutionary history. *Genome Res.* 14(11): 2245-2252

178. Pritham, E.J., and Feschotte, C. (2007) Massive amplification of rolling-circle transposons in the lineage of the bat Myotis lucifugus. *Proc. Natl. Acad. Sci. U. S. A.* 104(6): 1895-1900

179. Pritham, E.J., Putliwala, T., and Feschotte, C. (2007) Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. *Gene* 390(1-2): 3-17

180. Qiu, P., Ding, W., Jiang, Y., Greene, J.R., and Wang, L. (2002) Computational analysis of composite regulatory elements. *Mamm. Genome* 13(6): 327-332

181. Quentin, Y. (1994) Emergence of master sequences in families of retroposons derived from 7sl RNA. *Genetica* 93(1-3): 203-215

182. Reiss, D., Quesneville, H., Nouaud, D., Andrieu, O., and Anxolabehere, D. (2003) Hoppel, a P-like element without introns: a P-element ancestral structure or a retrotranscription derivative? *Mol. Biol. Evol.* 20(6): 869-879

183. Renard, M., Varela, P.F., Letzelter, C., Duquerroy, S., Rey, F.A.*, et al.* (2005) Crystal structure of a pivotal domain of human syncytin-2, a 40 million years old endogenous retrovirus fusogenic envelope gene captured by primates. *J. Mol. Biol.* 352(5): 1029-1034

184. Robertson, H.M. (2002) Evolution of DNA transposons in eukaryotes. In *Mobile DNA II* (Craig, N.L.*, et al.*, eds), 1093-1110, AMS Press, Washington, D.C.

185. Rogozin, I.B., Wolf, Y.I., Sorokin, A.V., Mirkin, B.G., and Koonin, E.V. (2003) Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr. Biol.* 13(17): 1512-1517

186. Rojas, F.J., Brush, M., and Moretti-Rojas, I. (1999) Calpain-calpastatin: a novel, complete calcium-dependent protease system in human spermatozoa. *Mol. Hum. Reprod.* 5(6): 520-526

187. Saitou, N., and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4(4): 406-425

188. Schmid, C.W., and Deininger, P.L. (1975) Sequence organization of the human genome. *Cell* 6(3): 345-358

189. Schmid, C.W., and Rubin, C.M. (1995) Alu: what's the use? In *The impact of short interspersed elements (SINEs) on the Hpst genome* (Maraia, R.J., ed), 105-123, RG Landes,

190. Schmid, C.W. (1998) Does SINE evolution preclude Alu function? *Nucleic Acids Res.* 26(20): 4541-4550

191. Schmid, C.W. (2003) Alu: a parasite's parasite? *Nat. Genet.* 35(1): 15-16

192. Shen, L., Wu, L.C., Sanlioglu, S., Chen, R., Mendoza, A.R.*, et al.* (1994) Structure and genetics of the partially duplicated gene RP located immediately upstream of the complement C4A and the C4B genes in the HLA class III region. Molecular cloning, exon-intron structure, composite retroposon, and breakpoint of gene duplication. *J. Biol. Chem.* 269(11): 8466-8476

193. Shi, B., Triebe, D., Kajiji, S., Iwata, K.K., Bruskin, A.*, et al.* (1999) Identification and characterization of baxepsilon, a novel bax variant missing the BH2 and the transmembrane domains. *Biochem. Biophys. Res. Commun.* 254(3): 779-785

194. Singer, M.F. (1982a) Highly repeated sequences in mammalian genomes. *Int. Rev. Cytol.* 76: 67-112

195. Singer, M.F. (1982b) SINEs and LINEs: highly repeated short and long interspersed sequences in mammalian genomes. *Cell* 28(3): 433-434

196. Sinnett, D., Richer, C., Deragon, J.M., and Labuda, D. (1991) Alu RNA secondary structure consists of two independent 7 SL RNA-like folding units. *J. Biol. Chem.* 266(14): 8675-8678

197. Smalheiser, N.R., and Torvik, V.I. (2006) Alu elements within human mRNAs are probable microRNA targets. *Trends Genet.* 22(10): 532-536

198. Smit, A.F. (1996) The origin of interspersed repeats in the human genome. *Curr. Opin. Genet. Dev.* 6(6): 743-748

199. Smit, A.F. (1999) Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* 9(6): 657-663

200. Sorek, R., Ast, G., and Graur, D. (2002) Alu-containing exons are alternatively spliced. *Genome Res.* 12(7): 1060-1067

201. Sorimachi, H., Imajoh-Ohmi, S., Emori, Y., Kawasaki, H., Ohno, S.*, et al.* (1989) Molecular cloning of a novel mammalian calcium-dependent protease distinct from both m- and mu-types. Specific expression of the mRNA in skeletal muscle. *J. Biol. Chem.* 264(33): 20106-20111

202. Speek, M. (2001) Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. *Mol. Cell. Biol.* 21(6): 1973-1985

203. Staalesen, V., Falck, J., Geisler, S., Bartkova, J., Borresen-Dale, A.L.*, et al.* (2004) Alternative splicing and mutation status of CHEK2 in stage III breast cancer. *Oncogene* 23(52): 8535-8544

204. Stankiewicz, P., and Lupski, J.R. (2002) Genome architecture, rearrangements and genomic disorders. *Trends Genet.* 18(2): 74-82

205. Suzuki, Y., Yamashita, R., Nakai, K., and Sugano, S. (2002) DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. *Nucleic Acids Res.* 30(1): 328-331

206. Sverdlov, E.D. (1998) Perpetually mobile footprints of ancient infections in human genome. *FEBS Lett.* 428(1-2): 1-6

207. Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22(22): 4673-4680

208. Thornburg, B.G., Gotea, V., and Makalowski, W. (2006) Transposable elements as a significant source of transcription regulating signals. *Gene* 365: 104-110

209. Tulko, J.S., Korotkov, E.V., and Phoenix, D.A. (1997) MIRs are present in coding regions of human genes. *DNA Seq.* 8(1-2): 31-38

210. van de Lagemaat, L.N., Landry, J.R., Mager, D.L., and Medstrand, P. (2003) Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends Genet.* 19(10): 530-536

211. van Rossum, D.B., Patterson, R.L., Sharma, S., Barrow, R.K., Kornberg, M.*, et al.* (2005) Phospholipase Cgamma1 controls surface expression of TRPC3 through an intermolecular PH domain. *Nature* 434(7029): 99-104

212. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J.*, et al.* (2001) The sequence of the human genome. *Science* 291(5507): 1304-1351

213. Vidal, F., Mougneau, E., Glaichenhaus, N., Vaigot, P., Darmon, M.*, et al.* (1993) Coordinated posttranscriptional control of gene expression by modular elements including Alu-like repetitive sequences. *Proc. Natl. Acad. Sci. U. S. A.* 90(1): 208-212

214. Voytas, D.F., and Boeke, J.D. (2002) Ty1 and Ty5 of *Saccharomyces cerevisiae*. In *Mobile DNA II* (Craig, N.L.*, et al.*, eds), 631-662, ASM Press, Washington, D.C.

215. Wagner, E., and Lykke-Andersen, J. (2002) mRNA surveillance: the perfect persist. *J. Cell Sci.* 115(Pt 15): 3033-3038

216. Wang, L., Wu, Q., Qiu, P., Mirza, A., McGuirk, M.*, et al.* (2001) Analyses of p53 target genes in the human genome by bioinformatic and microarray approaches. *J. Biol. Chem.* 276(47): 43604-43610

217. Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F.*, et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420(6915): 520-562

218. Watson, J.D., and Crick, F.H. (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* 171(4356): 737-738

219. Wells, J.M., Ellingson, J.L., Catt, D.M., Berger, P.J., and Karrer, K.M. (1994) A small family of elements with long inverted repeats is located near sites of developmentally regulated DNA rearrangement in *Tetrahymena thermophila*. *Mol. Cell. Biol.* 14(9): 5939-5949

220. Wessler, S.R. (2006) Transposable elements and the evolution of eukaryotic genomes. *Proc. Natl. Acad. Sci. U. S. A.* 103(47): 17600-17601

221. Zhou, Z., Ying, K., Dai, J., Tang, R., Wang, W.*, et al.* (2001) Molecular cloning and characterization of a novel peptidylprolyl isomerase (cyclophilin)-like gene (PPIL3) from human fetal brain. *Cytogenet. Cell Genet.* 92(3-4): 231-236

222. Zukunft, J., Lang, T., Richter, T., Hirsch-Ernst, K.I., Nussler, A.K.*, et al.* (2005) A Natural CYP2B6 TATA Box Polymorphism (-82T-> C) Leading to Enhanced Transcription and Relocation of the Transcriptional Start Site. *Mol. Pharmacol.* 67(5): 1772-1782

VITA

Valer Gotea

EDUCATION

8/2001 – 8/2007	**The Pennsylvania State University**, University Park, Pennsylvania, USA PhD in Biology; Advisor: Dr. Wojciech Makałowski
8/2000 – 8/2001	**Louisiana State University**, Baton Rouge, Louisiana, USA MS in Wildlife (not finished; transferred to Penn State in 8/2001).
10/1994 - 6/1999	**Transilvania University**, School of Silviculture and Forest Engineering, Brasov, Romania BS in Silviculture; Advisor: Prof. Dr. Ing. Aurel Negrutiu
9/1990 - 6/1994	**The Forestry Highschool**, Gurghiu, Romania

SELECTED PUBLICATIONS

Mount, S.M., **V. Gotea**[*], C.-F. Lin, K. Hernandez, W. Makałowski (**2007**) Spliceosomal small nuclear RNA genes in 11 insect genomes. *RNA* **13**(1): 5-14

Gotea, V., W. Makałowski (**2006**) Do transposable elements really contribute to proteomes? *Trends in Genetics* **22**(5): 260-267

Thornburg, B.G., **V. Gotea**[*], W. Makałowski (**2006**) Transposable elements as a significant source of transcription regulating signals. *Gene* 365: 104-110

Gotea, V., V. Veeramachaneni, W. Makałowski (**2003**) Mastering seeds for genomic size nucleotide BLAST searches. *Nucleic Acids Research* **31**(23): 6935-6941

Note: * - first co-author

SELECTED AWARDS / GRANTS

12/2006: **2007 Alumni Association Dissertation Award**, The Graduate School, The Pennsylvania State University, University Park, PA, USA

1/2006: Department of Biology **Graduate Assistant Excellence in Teaching Award** for 2005, The Pennsylvania State University, University Park, PA, USA

5/2003 – 7/2003: Grant from The Pennsylvania State University through its **Worldwide Universities Network (WUN)** program, for training stage in protein modeling at **University of Manchester**, Manchester, England

8/2001 – 5/2003: **The Braddock Fellowship** from the Eberly College of Science, The Pennsylvania State University, University Park, USA