BIOINFORMATIC ANALYSIS OF POLYADENYLATION SITE ACTIVITY IN VERTEBRATES

by

ERIC SAU-CHUM HO

A dissertation submitted to the

Graduate School – New Brunswick

Rutgers, The State University of New Jersey

and

The Graduate School of Biomedical Sciences

University of Medicine and Dentistry of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Biochemistry

Written under the direction of

Samuel I. Gunderson

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

OCTOBER, 2010

ABSTRACT OF THE DISSERTATION

BIOINFORMATIC ANALYSIS OF POLYADENYLATION SITE ACTIVITY IN VERTEBRATES

By Eric Sau-chum Ho

Dissertation Director:

Samuel I. Gunderson

Most eukaryotic protein coding precursor messenger RNAs (pre-mRNAs) undergo polyadenylation after transcription. Polyadenylation is a two-step enzymatic reaction, in which the emerging pre-mRNA is cleaved from the transcription complex, and then followed by the polymerization of adenosine nucleotides starting from the cleaved 3' end to form the poly(A) tail. Biologically, poly(A) tail increases mRNA stability, protein translatability, and mRNA nuclear export. Surprisingly, large numbers of protein factors were found to be involved in this apparently simple cleavage and polymerization steps, suggesting that polyadenylation is under complex regulation. Hence in this study, I am interested to investigate the regulatory elements of eukaryotic polyadenylation.

The proposed close species comparison approach revealed an asymmetric selection pressure around the polyadenylation cleavage site (PAS). The region from the PAS to approximately 200 nucleotides (nts) upstream was found to be under a much higher conservation than the downstream region and other part of the 3'UTR. Furthermore, over 2,000 long (>30 nts) conserved fragments at or close to upstream of the PAS were identified through remote species comparison. A substantial portion of them are longer than 100 nts, which is much longer than any known RNA protein recognition sites.

A PAS classifier was built using logistic regression in order to study the characteristics of PAS. Not only it does improve the computational recognition of mammalian PAS than existing methods, it is also helpful in identifying a small number of genes that lack of typical PAS characteristics such as the poly(A) signal and/or the U/GU rich region. These findings provide useful experimental candidates for the study of the still unclear polyadenylation compensatory and/or regulatory elements.

At present, no sequence consensus has been identified for the downstream U/GU enriched region yet. Thus, I have designed a novel rule-based nucleotide sequence motif finding algorithm, called iTriplet, to target long and degenerative motifs with special attention to the PAS downstream sequence. iTriplet has been demonstrated to handle motifs longer than 20 nts, which is still a challenge to existing methods. The utility of iTriplet has been confirmed by showing it accurately predicts PAS downstream elements using a dual Luciferase reporter assay.

ACKNOWLEDGMENTS

I would like to express my earnest gratitude to Dr. Samuel I. Gunderson for his insightful discussions, dedication, and guidance during my graduate study at Rutgers. Especially, I deeply appreciate his meticulousness and open-mindedness in science.

Moreover, I would like to give my deepest thanks to my wife, Flora, and our sons Caleb and Jeshua. The completion of this dissertation will not be possible without their sacrifice, patience, encouragement, and day-to-day support.

DEDICATIONS

To my mother,

my wife Flora, and our sons Caleb and Jeshua

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1

INTRODUCTION

A. Background

The majority of eukaryotic protein-coding messenger RNA precursors (pre-mRNAs) undergo required maturation processing in the nucleus before being exported to the cytoplasm. This maturation process consists of three modifications viz. 5' capping, splicing, and polyadenylation. Although these modifications are often called post-transcriptional processing, they actually occur simultaneously and cooperatively during transcription. RNA modifications serve vital biological functions and are thought to facilitate diversity. Splicing can lead to the production of more than one species (isoform) of mRNA of a single gene, as many as 80% of human genes are detected with alternatively spliced isoforms [reviewed in Matlin et al 2005]. Alternative splicing often alters the protein-coding region of a gene, resulting in different proteins from the same gene without any change in its genome. 5' capping and polyadenylation modify the 5' and 3' ends of the mRNA molecule, respectively. They are critical to mRNA nuclear export, stability, and translatability. Intriguingly, polyadenylation is the only pre-mRNA modification out of the three that is preserved in all domains (super-kingdoms) i.e. prokaryotes, archaea, and eukaryotes. During the three billion years of evolution, additional complexity was selected in the mammalian polyadenylation machinery. Thus in this thesis, my focus is to study the more complicated polyadenylation activity in mammals.

All eukaryotic protein-coding messenger RNAs (mRNAs) are polyadenylated except histones. Polyadenylation consists of two tandem enzymatic reactions i.e. the endonucleolytic cleavage of nascent pre-mRNA emerging from the transcription complex, and the polymerization of adenosine nucleotides to the 3' end of the pre-mRNA. The endonucleolytic cleavage site is called the polyadenylation site (PAS). The choice of PAS is selective even though human and mouse genes are found to possess more than one PAS [Tian et al 2005]. The polyadenosine nucleotides polymerized at the 3' end of the mRNA is collectively called the poly(A) tail. The typical length of the poly(A) tail in mammals is 200-250 nucleotides (nts) long, but lower organisms tends to have a shorter poly(A) tail e.g. it is about 70 nts in yeast, 10-20 nts in bacteria. Polyadenylation is a non-template driven process, in contrast to transcription and DNA replication. It takes place in the nucleus, however not without exception as cytoplasmic polyadenylation can undergo shortening and lengthening in the cytoplasm. Example of cytoplasmic polyadenylation was reported in Xenopus during oocyte maturation and early embryogenesis [Pique et al 2008].

B. Polyadenylation core factors

Illustrated below in Figure 1.1 is the core eukaryotic polyadenylation complex, which consists of six protein factors comprised of fourteen different polypeptides.

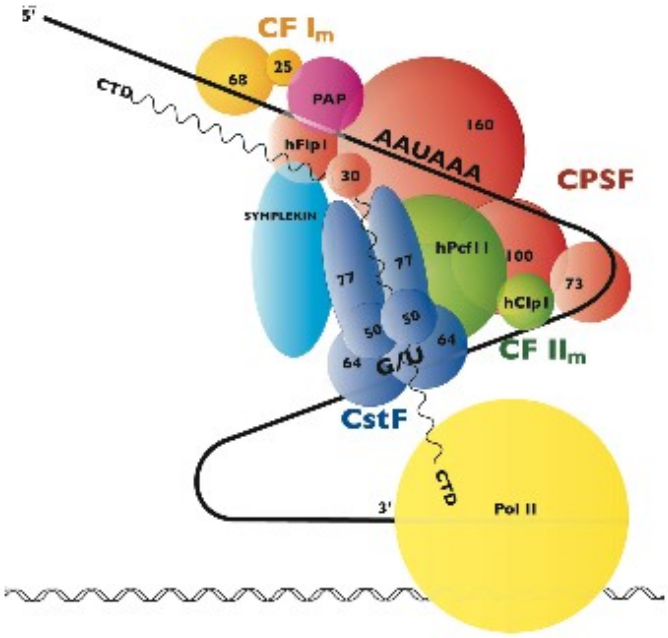


Figure 1.1 Core protein factors of the mammalian polyadenylation complex. The carboxyl terminal domain (CTD) of RNA polymerase II is tightly coupled to the polyadenylation complex. Figure is adopted from [Mandel et al 2008], where the authors suggested that CstF-64 dimerizes at the downstream region.

This complex is highly conserved in eukaryotes. Yeast homlogs can be found in 10 out of 14 mammalian proteins [Mandel et al 2008, Shi et al 2009]. This complex takes about 10 seconds to assemble according to one study [Chao et al 1999]. As mentioned in [Mandel et al 2008], it is surprising that so many proteins are required to perform such a simple cleavage and polymerization

process. In addition, a recent proteomic study has identified as many as 85 different proteins in the polyadenylation complex, including known polyadenylation factors, indicating up to 50 other proteins may influence polyadenylation [Shi et al 2009]. The polyadenylation molecular machinery utilizes two cis elements to recognize the PAS. The upstream element of PAS consists of a highly conserved hexanucleotide, called the poly(A) signal, which is located 10-30 nts from the PAS. The two most prevalent forms of poly(A) signal in vertebrates are AAUAAA and AUUAAA[1], collectively called the canonical poly(A) signal. According to my own and other data [Beaudoing et al 2000, Tian et al 2005], AAUAAA and AUUAAA are found in approximately 66% and 16% of mammalian genes, respectively. The poly(A) signal is recognized by cleavage and polyadenylation specificity factor (CPSF) CPSF-160 during complex formation.

On the contrary, no sequence consensus can be identified for the downstream element (DSE) except that an U and G enriched region is found at ~15 nts downstream from PAS, which is commonly called the U/GU-rich region. The 64-kDa subunit of the cleavage and stimulating factor (CstF), CstF-64, was found to target the U/GU-rich region but not simple $(GU)_n$ repeats through SELEX experiments and NMR study [Takagaki et al 1997, Perez et al 2003]. In addition, experimental data indicated that cleavage and polyadenylation occur deterministically at a fixed location (±10 nts) between the PAS and the U/GU-rich

[1] In this document, uracil (U) and thymine (T) are used interchangeably.

region. A recent computational study of PAS downstream sequences from various metazoans suggested that DSE exhibits a 5' to 3' transition from UG-rich to U-rich [Salibury et al 2006].

Direct binding of the two abovementioned protein factors, CPSF-160 and CstF-64, to the poly(A) signal and DSE, respectively, are inadequate to trigger polyadenylation. Two additional cleavage factors CF-I and CF-II, are reported to increase complex stability, and to enhance CPSF-160 interaction with CstF-64, which results in forming a closed loop in the pre-mRNA substrate between the poly(A) signal and DSE [Takagaki et al 1989, de Vries et al 2000]. Another subunit of the CPSF, CPSF-73, was reported to function as an endonuclease to cleave the pre-mRNA preferentially but not necessarily after dinucleotide 'CA' between the poly(A) signal and the DSE [Mandel et al 2006].

C. Alternative polyadenylation

A substantial portion of human genes were found to possess more than one 3' end [Iseli et al 2002]. With the burgeoning of genomic data, a more recent study has determined that ~54% of human and ~32% of mouse genes were found to have alternative PAS [Tian et al 2005]. Alternative polyadenylation results in the alteration of the 3' UTR, and in some cases, the truncation of the carboxyl terminal of the protein. It is still unknown whether the choice of PAS is stochastic or regulated, as well as its activation or inactivation mechanism. At present, only a few examples are known to delineate its biological function. An example of alteration of the coding region through alternative polyadenylation can be

illustrated by the IgM heavy chain gene, which contains two active polyadenylation sites. Activation of upstream PAS, μ_s, will result in the secretory form of IgM, whereas the activation of the downstream PAS, μ_m, will give rise to the membrane-bound form [Lamson et al 1984, Phillips et al 2001]. The difference in localization is due to the truncation of the coding region in 3' end that encodes the membrane anchor domain. Even though the alternative polyadenylation of 3,108 (22%) human and 898 (8%) mouse genes were detected to alter the protein coding region [Table 3 of Tian et al 1995], so far, only IgM is well studied, indicating the biological function of alternative polyadenylation in many genes is still unknown.

In most situations, alternative polyadenylation affects only the 3' UTR but leaves the coding region intact. It is known that the 3' UTR embodies myriad of regulatory elements such as microRNA targets [Xie et al 2005], mRNA stability elements like AU-rich regions, polyadenylation inhibition elements, U1 binding sites [Gunderson et al 1998], and mRNA localization "ZIPCODE" elements [reviewed in Shav-Tal Y et al 2005]. As a result, mRNA levels may be affected by alternative polyadenylation, and subsequently, affects the protein level as well. Besides the effect on 3' UTRs, an intron enhancer located downstream of exon 4 in the calcitonin gene is also reported to regulate alternative polyadenlyation [Lou et al 1996]. The lengthening of 3' UTR has been revealed to associate with mouse embryonic development [Ji et al 2009] and it is believed that the lengthened transcripts are turned into substrates of other regulatory agents like microRNAs. On the contrary, the shortening of 3' UTR was observed in

oncogene transcripts, which is thought as a mechanism for oncogenes to escape from microRNA repression [Mayr et al 2009].

D. Non-canonical polyadenylation

Despite the fact that the poly(A) signal is highly conserved in vertebrates, a small fraction of genes do not conform to the canonical pattern and yet they are polyadenlyated precisely at the same cleavage site. The first example being reported is the gene poly(A) polymerase gamma (PAPOLG) which has no canonical poly(A) signals but contains multiple copies of conserved UGUAN (N=A is better than U, as better than G,C) in the upstream of PAS [Venkataraman et al 2005]. In that study, the binding of human CF-I to UGUAN sites was shown to stimulate polyadenylation. Note that this study lacked cell-culture data and it failed to exclude the binding of CPSF-160 to a canonical-like poly(A) signal, which was present in PAPOLG. Another example is the DNA polymerase gene of Epstein-Barr virus that contains the non-canonical poly(A) signal, UAUAAA, yet it was shown to be essential for polyadenylation though with less efficiency [Silver Key et al 1997].

The presence of high conservation pressure to preserve the upstream poly(A) signal but not the degenerate downstream U/GU-rich region may indicate only the poly(A) signal is sufficient to trigger polyadenylation. In addition, I have identified many reliable PAS without U and G enriched downstream region (detailed discussion can be found in chapter 3). However, one study has reported the presence of auxiliary G-rich elements further downstream is required to

maintain polyadenylation activity of that gene [Dalziel et al 2007]. The intronless gene MC1R has a canonical poly(A) signal AAUAAA upstream but lacks the U/GU-rich downstream element. Through mutagenesis studies, authors have demonstrated that two downstream G-rich regions serve to rescue normal polyadenylation activity, without which, polyadenylation diminished significantly. Despite that, the UU dinucleotide located 21 nts downstream from PAS, which is the favorite position of the DSE, still remains critically important to maintain polyadenylation as disruption abolishes polyadenylation activity.

With the help of genomic and expression data, there is growing evidence to support the view that the polyadenylation molecular machinery is flexible to tolerate sequence variations of the poly(A) signal and/or the DSE. Such a view is consistent with the discovery of as many as 85 proteins in the polyadenylation complex mentioned above [Shi et al 2009]. Such additional factors may serve as compensatory and regulatory functions. Examples have shown that the weakness of non-canonical poly(A) signal can be compensated by a strong DSE, and vice versa, in the absence of other auxiliary elements. This idea has been illustrated in a recent in-vitro study about the compensatory effect of a non-canonical poly(A) signal and a DSE without any auxiliary element [Nunes et al 2010]. Human MC4R and JunB genes are examples of this type. The intronless human MC4R gene lacks a canonical poly(A) signal but possesses an A-rich upstream region, and an U/GU-rich downstream region. The authors showed that the downstream U/GU-rich region was sufficient to drive polyadenylation activity. Interestingly, though not mentioned in that report, the mouse homolog does

possess the major canonical poly(A) signal AAUAAA and the DSE is quite U/GU-rich too. In addition, the expression of MC4R is quite low in both species, and their 3' ends are not supported by ESTs. Equally interesting is the finding that the gene of the human CPSF-160, which recognizes and binds the poly(A) signal, does not have the canonical poly(A) signal. Based on these few examples, we can understand that the core polyadenylation complex exhibits a wide spectrum of flexibility, and its tolerance to variations is gene-specific. Later, I will discuss examples of genetic disorders due to the slight variations in the flanking region of the PAS.

E. Polyadenylation and transcription termination

Currently, there are two popular views on transcription termination viz. anti-termination and torpedo models. Both models support the interaction between transcription termination and polyadenylation. The anti-termination model proposed that some proteins called anti-termination factors "piggy back" on the transcription complex during the elongation phase. When the transcription complex reaches an active PAS, it will trigger the release of anti-termination factors from the transcription complex thereby causing the destabilization of the RNA polymerase II/DNA complex. An alternate view on termination is called the torpedo model. In this model, after the cleavage of nascent mRNA at the PAS from the RNA polymerase II (Pol II), an 5'->3' exonuclease Xrn2 will degrade the emerging nascent mRNA from the transcription complex until Xrn2 interacts with the complex, which in turn will cause the transcription complex to fall off from the DNA [West et al 2004]. According to my data, 6,000+ of human and 12,000+ of

mouse genes are less than 1,000 nts apart, suggesting proper transcription termination is vital to maintain transcription integrity. Early reports proposed that both the poly(A) signal and the downstream G-rich pause element MAZ were required to cause Pol II transcription termination [Eggermont et al 1993, Yonaha et al 1999, Plant et al 2005, West et al 2006], and transcription termination was suggested to couple with polyadenylation [Yonaha et al 2000]. The nascent mRNA was identified to tether the polyadenylation complex to the Pol II [Rigo et al 2005]. Other studies suggested however that the canonical poly(A) signal alone is sufficient to induce transcription pausing, which may switch Pol II from an elongation state to an abortive state [Orozco et al 2002, Kim et al 2003, Nag et al 2006]. However, my data shows that canonical poly(A) signals are ubiquitous in transcribed regions. In order to support poly(A) signal dependent pausing, factors other than sequence elements must be utilized by the transcription complex to prevent premature loss of possessivity.

F. Evolutionary history of polyadenylation

The origin of mammalian polyadenylation can be traced back to the most primitive organisms in all three domains (super-kingdoms) of life i.e. prokaryotes, archaea, eukaryotes, including organelles like chloroplast and mitochondria. Even though polyadenylation orchestrates quite differently in these three domains in terms of the protein factors, the existence of the poly(A) signal, and the sequence characteristics surrounding the PAS, a common biological role has been preserved through evolution, which is the turnover of mRNA molecules. This observation suggests that mRNA turnover is the ancestral function of

polyadenylation. Thus polyadenylation should be viewed as the counterpart of transcription, where the former helps to recycle ribonucleotides for the latter.

Escherichia coli (E. coli) will be used as the model organism to illustrate prokaryotic polyadenylation. The 3' end of most E.coli transcript is marked by a stem-loop structure, which helps to resist 3'-exonucleolytic degradation. Endonucleases such as RNase E try to remove the stem-loop by attacking its base so as to allow 3'->5' degradation. However this reaction is slow. Apart from exonucleolytic degradation, the exposed 3' end of the RNA is also available for polyadenylation by the poly(A) polymerase pcnB. When the poly(A) tail is formed at the 3' end of the transcript, it is thought to serve as a 'toehead' for another enzyme polynucleotide phosphorylase (PNPase), which works synergetically with RNase E to stimulate 3'->5' exonucleolytic degradation [Xu et al 1995, Cohen 1995]. This mechanism was reported to account for the regulation of plasmid copy in E.coli [Xu et al 1993, 2002, He et al 1993]. Most prokaryotic poly(A) tail was found to be 10-20 nts long, and only 2-60% of the mRNA of a gene were detected to have a poly(A) tail [Taljanidisz et al 1987, Karnik et al 1987]. The stimulatory role of the poly(A) tail were also found in archaea, chloroplast, and mitochondria [Rott et al 2003, Slomovic et al 2005, Portnoy et al 2006]. Several good reviews of prokaryotic polyadenylation can be found in [Sarkar 1997, Edmonds 2002, Slomovic et al 2006].

On the other hand, additional components were selected in eukaryotes during the course of evolution. These include the presence of the poly(A) signal, distinct nucleotide composition flanking the PAS, and the multimeric

polyadenylation complex. These indicate that new functions are incorporated in eukaryotic polyadenylation in addition to its ancestral mRNA turnover role. One critical distinction between prokaryotes and eukaryotes is the non-covalent circularization of mRNA. Circularization enhances mRNA stability and protein translation capability. But it requires additional players to bring 5', and 3' ends together. The birth of 5' capping enzyme fulfilled such 5' role. The capping enzyme produces a cap structure (m^7Gppp) in the 5' end of the pre-mRNA by attaching a guanosine to the 5' most nucleotide through an usual 5'-to-5' triphosphate linkage. Regarding the 3' end, a highly conserved poly(A) binding protein (PABP), which binds to the poly(A) tail, is found in eukaryotes. These two terminal modifications help the mRNA molecule to resist exosome degradation in the nucleus, which is essential for mRNA stability. Moreover, the nuclear export pathway also uses these two modifications to gauge the export of mRNA to cytoplasm for translation. Before translation, the 5' cap interacts with the PABP-bound poly(A) tail through the mediation of translation initiation factors eIF4E and eIF4G. The circularization structure is shown to facilitate multiple rounds of translation per mRNA molecule.

The comparison between prokaryotic and eukaryotic polyadenylation not only provides additional understanding about this process, but also how little is known about nucleus formation. Even by comparing the two unicellular organisms E.coli and yeast, the vast difference between their polyadenylation mechanisms is still puzzling. So far, little evidence is known about the intermediate for the transition from non-nucleus to nucleus. Further investigation

is needed to fill the missing knowledge between the two domains of life during evolution.

G. Polyadenylation and diseases

Several studies have shown that genomic variation flanking the PAS can be detrimental. Examples of disease-related genomic variations in regions surrounding the PAS will be discussed here. It has been reported that aged-related macular degeneration (AMD) is associated with the deletion-insertion (indel) of an upstream region of PAS of gene ARMS2 [Fritsche et al 2008]. AMD causes diminishing of central retinal vision, and 50% of AMD patients are accounted by indel genetic variation [Edwards et al 2005, Haines et al 2005, Hageman et al 2005]. Genotyping of AMD patients indicated a 43-nt fragment, which carries the poly(A) signal, being replaced by a 54-nt fragment with two non overlapping AU-rich pentamers. Homologous ARMS2 can only be found within the primate lineage, and the biological function of ARMS2 still remains unknown. The loss of the poly(A) signal compounded with the two extra AU-rich pentamers not only hampers polyadenylation activity, but also reduces mRNA stability. As a result, the protein level of ARMS2 drops drastically in retina of affected patients.

Other more subtle polymorphisms surrounding PAS were also found to be disease related, though they were not as drastic as losing the poly(A) signal. Their main adverse effect is the alteration of polyadenylation efficiency. One example is the single nucleotide polymorphism (SNP) at the PAS of prothrombin or coagulation factor II gene (F2). Two SNPs have been discovered immediately

5’ upstream of F2’s PAS viz. rs72550707 C→T, and rs1799963 G→A. C→T is mostly found in Afro-Americans and Afro-Caribbeans, whereas G→A is almost exclusively found in Caucasians [Danckwardt et al 2006]. The G→A polymorphism was reported to elevate mRNA level of the F2 gene [Sachchithananthan et al 2005, Danckwardt et al 2004, 2006] due to the increase of polyadenylation efficiency but not translatability [Gehring et al 2001]. As blood coagulation is a sensitive and responsive physiological process, the boosting of polyadenylation efficiency increases the level of prothrombin protein in the plasma that will result in venous thrombophilia. C→T polymorphism also contributes to thrombophilia and complications of pregnancy.

Besides polymorphisms at the poly(A) signal and the PAS, variation in the downstream U/GU-rich region was also found to upset thrombosis. The fibrinogen gamma gene (FGG) consists of 10 exons and two PAS. The upstream PAS (PA1) is located in intron 9. The use of PA1 produces the shorter isoform of FGG (γ’), whereas the use of downstream PAS (PA2) will produce the longer FGG (γA). γA contains four more amino acids “AGDV” than γ’ at the carboxyl terminal. The last four amino acids are involved in platelet-binding. A mixture of γ’ and γA are found in the blood stream, where γ’ usually consists of 7-15% of total FGG level. Maintaining the γ’ to γA ratio in blood is physiologically important. A C→T SNP located at the U/GU-rich region 3’ downstream of PA2 was found in patients suffering from deep venous thrombosis (DVT) [Uitte de Willige et al 2005, 2007]. The same study discovered an elevated mRNA level of γA in DVT patients. Thus, the C→T variation was believed to strengthen PA2, which led to

the lowering of γ' and the ratio between γ' and total FGG. The strengthening of PA2 is thought to be contributed by making the PAS downstream region more U-rich, which may facilitate polyadenylation factor CstF-64 binding.

Besides genetic variation, utilizing alternative the polyadenylation mechanism to shorten 3' UTR was observed in six oncogenes in cancer cells that led to changes in protein products [Mayr et al 2009]. The shortening of 3' UTR allows oncogenes to escape microRNA-mediated repression. In addition, one of the oncogenes with shortened 3' UTR was IMP-1, which was found to promote oncogenic transformation. Regarding the mechanism to activate 5' upstream or 3' downstream PAS, it is still unknown. Through previous microarray comparative study, one group of the authors speculated that the elevated level of CPSF-160 (CPSF1) and CstF-64 (CSTF20) may favor the usage of 5' PAS even though the sequence propensity is suboptimal.

H. Polyadenylation and oligonucleotide-based therapeutics

Oligonucleotide-based, or simply oligo-based, drugs like most existing drugs are antagonists. Currently, there are two main categories of oligo-based therapeutic methods viz. antisense oligonucleotide (ASO), and RNA interference (RNAi). Their main difference lies in the use of different endogenous mRNA degradation pathways. In the last two decades, growing attention has been given to harness these mRNA degradation pathways as the therapeutic method for diseases such as cancer, familial hypercholestrolaemia, malaria etc. [Melnikova 2008]. With the advances in nucleic acid chemistry, delivery mechanism, and voluminous

genomic data, the momentum of oligo-based therapeutics is growing even larger. The key advantage of oligo-based drugs compared with traditional small compound drugs is in the discovery of potent interacting sites between the antagonist and the target. In the traditional drug discovery process, identification of the active site of the target protein requires structural information, which may be a daunting task for some protein families such as membrane proteins. Once the target site is decided, the next step is to develop an assay in testing the potency of small compounds from a chemical library. Required working knowledge is completely different from one target to the next. Synthesis of small compounds also varies from one drug to the others as well. However, the screening process is more streamlined for oligo-based drug discovery. All one needs is to screen for one or more unique and accessible target sequences in the mRNA of the target gene. As variation in sequence pattern usually does not affect the biochemical property and synthesis of the oligonucleotides, the screening process does not depend on the target protein. In addition, oligo-based drugs make personalized medicine more probable than traditional approaches, as the personalization of an oligonucleotide is much easier than a small compound. Similar advantages apply to the combat of drug resistance due to the evolution of targets.

In addition to the above two oligo-based methods, a new method has been invented recently which takes on a different mRNA degradation pathway i.e. the inhibition of polyadenylation via the U1snRNP splicing factor. Previous studies have demonstrated that direct interaction between the U1-70k subunit of

splicing factor U1 snRNP, and poly(A) polymerase (PAP) can inhibit polyadenylation after cleavage [Gunderson et al 1998, Vagner et al 2000]. This inhibition mechanism was engineered as a post-transcription gene silencing tool, namely U1 silencing, as shown in Figure 1.2 below:

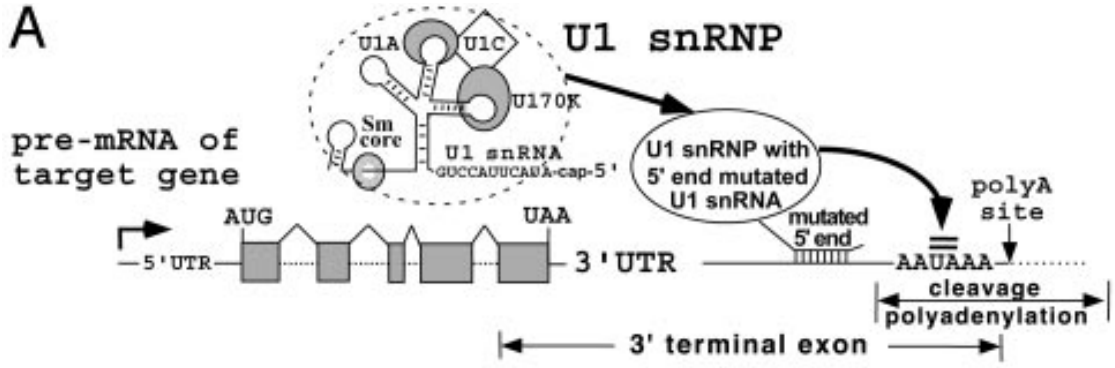


Figure 1.2 U1 silencing. U1 snRNP consists of U1 snRNA and 10 other proteins. A 10-nt sequence at the 5' end of U1 snRNA targets the 5' splice site (5'ss) during splicing. The 10-nt sequence in the mutated U1 snRNA is changed to basepair with the target gene. The above figure is adopted from [Forte et al 2003]

The idea of U1 silencing is to tether the U1 snRNP to the upstream of PAS in the terminal exon via a mutated U1 snRNA, where its natural 10-nt long 5' end targeting sequence is changed to form a duplex with an unique site flanking the PAS in the target gene as illustrated in Figure 1.2 above. Various research groups have demonstrated successes in applying this method to silence genes in different cell lines by transfecting cells with the mutated U1 snRNA [Beckley et al 2001, Fortes et al 2003, Akum et al 2004, Abad et al 2008, Jankowska et al

2008]. Recently, a significant improvement has been made to improve this method by the concept of U1 Adaptor [Goraczniak et al 2009].

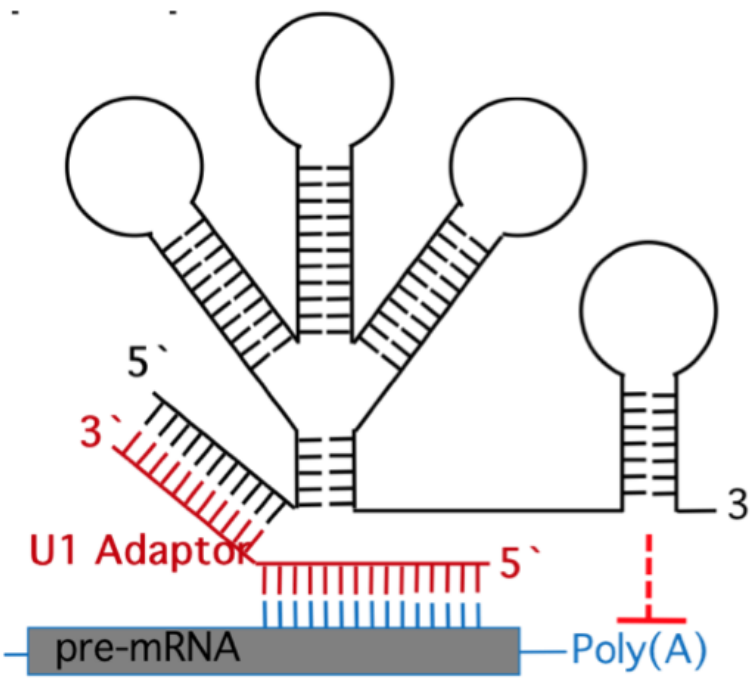


Figure 1.3 U1 Adaptor technology. Endogenous U1 snRNA is labeled in black, U1 Adaptor is labeled in red. Adopted from [Goraczniak et al 2009].

Instead of customizing U1 snRNA, a short adaptor oligonucleotide known as the U1 Adaptors is used to tether U1 snRNP to the terminal exon that contains the PAS. U1 Adaptor is a synthetic oligonucleotide of about 28-33 nucleotides in length and comprised of a 5' segment, the Target Domain, which binds within the terminal exon of the target pre-mRNA, and a 3' segment, the U1 Domain, which binds to the 5' end of U1 snRNA [Goraczniak et al 2009]. U1 Adaptor tethers U1 snRNP, via its U1 snRNA subunit, to a sequence near the PAS of the targeted gene. The U1 Domain design is relatively simple as its role is to bind as strongly as possible to U1 snRNP via base pairing to U1 snRNA. In contrast, the Target

Domain design is a balance between high affinity to the target and low affinity to non-targeted pre-mRNAs. A key aspect that, in part, explains the specificity of the Adaptor method is that inhibition only occurs when the Adaptor:U1 snRNP complex is bound in the terminal exon. Thus, Adaptor:U1 snRNP complex binding to upstream introns or exons of either the target gene or non-targeted genes has no effect. Even though a robust algorithm to select the U1 Adaptor target site is still under development, several genes have been silenced by this technology.

I. Summary

As discussed above, a seemingly straightforward two-step enzymatic reaction turns out to be far more complex that it should. During the course of evolution, variations of polyadenylation factors and the surrounding PAS bring in advantageous functions as well as complexity to this modification step. Such additional complexity is likely associated with regulatory functions. Hence I am interested to discover the regulatory role of regions flanking the PAS. In this report, I provide an extensive bioinformatic study to identify polyadenylation regulatory elements and to determine how widespread they are in mammals using bioinformatic, machine learning, and statistical techniques.

I have found an unusual asymmetric conservation pressure upstream of the PAS but not downstream of the PAS. Around 2,000+ of highly conserved fragments, at least 30 nts long, are found in the upstream region of remote species. Their discovery may reveal important and yet unknown activity

associated with these conserved fragments. Furthermore, I conducted an extensive study to identify the features that constitute strong and weak polyadenylation sites. Hence, I have used a supervised learning method to construct a polyadenylation site classifier. The classifier not only allows us to make prediction of PAS in novel genomes, but also assist in the identification of atypical polyadenylation sites. Such polyadenylation site outliers provide excellent examples to investigate less understood factors of polyadenylation. Finally, the degenerate nature of downstream U/GU-rich elements has prompted me to develop a new motif finding algorithm that is specifically capable of identifying long and degenerate motifs, which are commonly found in RNA.

CHAPTER 2

CONSERVATION OF POLY(A) SITE FLANKING REGION

A. Introduction

The existence of cis polyadenylation elements both upstream and downstream of the poly(A) signal has been studied experimentally and bioinformatically. Bioinformatic analysis discovered the enrichment of certain hexamers upstream, up to 100 nucleotides (nts), in human [Hu et al 2005], or downstream, up to 60 nts, of polyadenylation sites (PAS) in metazons [Salisbury et al 2006]. Through experimental studies, various functions have been attributed to other cis regulatory elements including, but not limited to, the inhibition of polyadenylation through a U-rich region downstream of the PAS [Zhu et al 2006], stabilization of the polyadenylation complex by U-rich elements upstream of the PAS [Kaufmann et al 2004, Danckwardt et al 2007], alteration of polyadenylation by U/GU-rich elements downstream of the PAS [Liu et al 2008], stimulation of the cleavage step through proximal and distal G-rich elements downstream of the PAS [Phillips et al 2004, Dalziel et al 2007], and U1A autoregulation through polyadenylation inhibition element (PIE) [Boelens et al 1993, Gunderson et al 1994, 1997]. So far, these studies have emphasized the presence of short (<15 nts) cis regulatory elements flanking (up to 100 nts upstream) the PAS. Furthermore, other related studies largely ignored the possibility that highly conserved elements could be effecting 3' end processing [Siepel et al 2005]. This chapter attempts to establish, first, the existence of selection pressure in the

farther upstream region (up to 200 nts) of the PAS, and second, the existence and prevalence of longer (>30 nts) conserved fragments (CFs) in distant mammalian species specifically, human, mouse, cow and platypus. Last but not least, the biological implications of these conserved regions will be discussed at the end.

B. Close species comparison reveals selection pressure on the farther region 200nt upstream of poly(A) sites

Polyadenylation is required for expression of all eukaryotic genes (except histone). It has long been understood that there is a strong selection pressure to maintain the poly(A) signal upstream near the PAS. In contrast, it is not understood whether selection pressure extends beyond the poly(A) signal and at what range of distance from the PAS. In order to answer these questions, the mutation rate near the PAS was measured. However, a simple comparison of PAS flanking sequences among different species is not feasible because, unlike ORFs, 3' UTRs are generally not conserved. Furthermore, nucleotide sequence comparison suffers from the homoplasy effect, i.e. recent mutation(s) can revert a mutated nucleotide to its ancestral form over a long evolutionary time. To overcome this issue, the approach to harness close species genomes was adopted to examine the existence of selection pressure flanking the PAS. Two pairs of close species were used: viz. human-chimpanzee and mouse-rat. The human and chimpanzee genomes are almost 99% identical [Chimpanzee genome sequencing consortium 2005], and the genome between mouse and rat is close to 90% identical [Rat genome sequencing consortium 2004]. Results

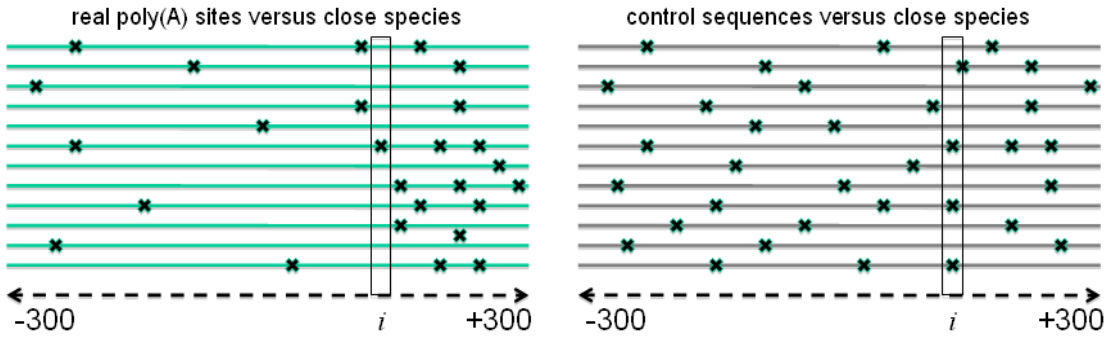
suggest that the proposed method is capable of pairing up orthologous (based on ORF) PAS regions even in less conserved 3' UTRs of close species.

1. Methods

For a given genomic region, in the absence of selection pressure, one would expect mutations to be distributed evenly along the genome; otherwise, mutations are either localized or depleted in that region. Based upon this intuition, the following procedure was devised to reveal the extent of selection pressure flanking the PAS.

1. obtain 17,080 human and 8,799 mouse PAS from our EST-based PAS database (described in the Appendix B)
2. consider regions [-300,+300] (see note below)
3. use NCBI-BLASTN [Camacho et al 2009] to identify chimpanzee and rat homologous PAS of human and mouse, respectively
4. remove genes with 3' UTRs shorter than 500 nts so as to eliminate the conservation effect caused by the ORF
5. choose two control data sets that are of the same length and same number as the sequences from step 1. These two control sequences were taken from random locations in the intergenic region and in the ORF
6. examine the mismatch ratio (explained below) for each position among homologous pairs in [-300,+300] (see note below) of the PAS

Note: [-M,+N] denotes M nts upstream and N nts downstream of the PAS.



$$mismatch_i = \frac{mismatch\_poly(A)_i}{mismatch\_control_i}$$

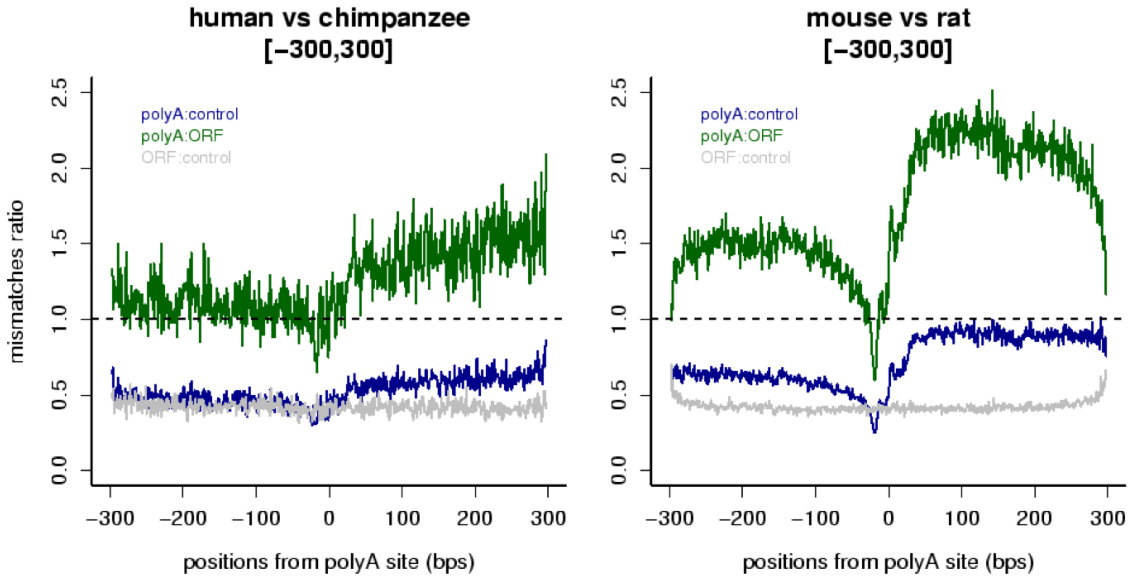
Figure 2.1 Mismatch ratio. Green lines on the left denote 600-nt long real PAS sequences supported by EST data. Grey lines on the right represent control sequences. Cross symbol represents mismatch. Mismatch ratio is computed for each position, denoted by *i*.

Mismatch ratio. 16,835 and 8,604 pairs of homologous PAS were found between human-chimpanzee, and mouse-rat, respectively, using NCBI-BLAST. For both real and control result sets, the number of mismatches were counted between each pair of species for each position in the [-300,+300] region. Then the two mismatch counts were combined into a ratio per position as shown in Figure 2.1. (Note: the mismatch ratio was set to undefined during plotting if the number of mismatches in control sequences was zero. Since large number of PAS regions were used, this situation were only found to happen in the first and last three positions at either ends, thereby it would not affect the overall analysis.) The mismatch ratio reflects the comparative mutation rate in PAS

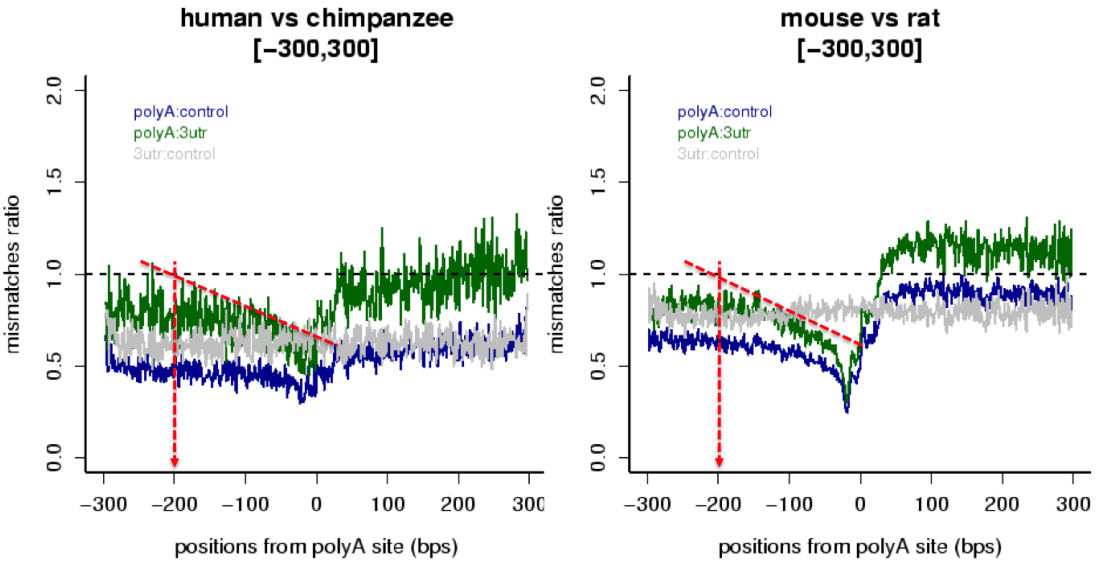
regions versus control sequences. A value close to 1, >1, and <1 indicates neutral, faster, and lower mutation rates in the PAS region versus control. Regarding the choice of control sequences, the decision is based on the assumption that intergenic sequence is subjected to the least selection pressure, whereas the strongest pressure is on the ORF. The comparison of the PAS flanking region with these two extremes enables us to understand the magnitude of selection pressure. Besides the PAS flanking region, other types of genomic sequences such as 5' splicing sites, part of the 3' UTR and introns were included in this study in order to confirm the validity of this method. The degree and the extent of conservation of the region flanking the PAS were examined by plotting the mismatch ratio for these two pairs of close species.

2. Results

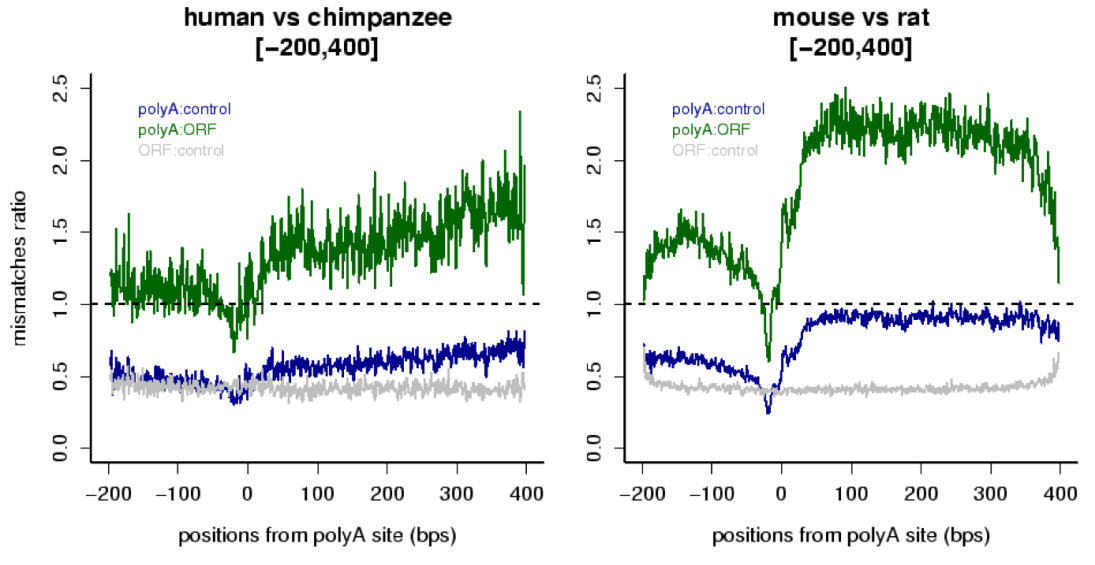
a) Selection pressure in human-chimpanzee and mouse-rat



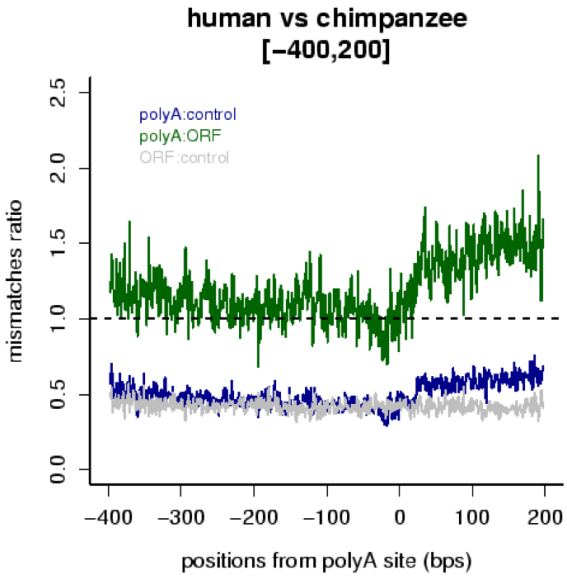
A B


human vs chimpanzee
[−300,300]
polyA:control
polyA:3utr
3utr:control
mismatches ratio
positions from polyA site (bps)
mouse vs rat
[−300,300]
polyA:control
polyA:3utr
3utr:control
mismatches ratio
positions from polyA site (bps)

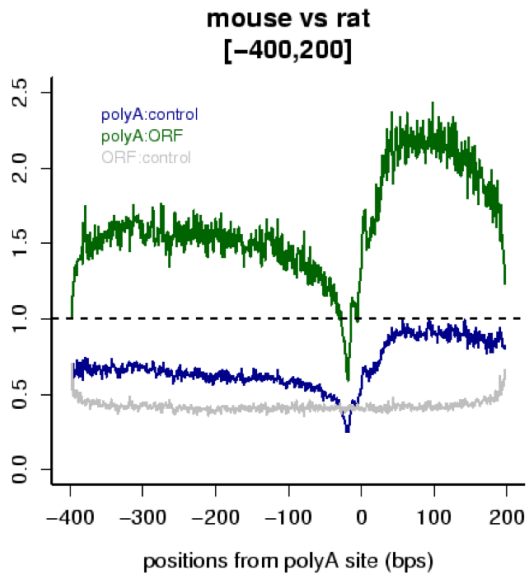
C D


human vs chimpanzee
[−200,400]
polyA:control
polyA:ORF
ORF:control
mismatches ratio
positions from polyA site (bps)
mouse vs rat
[−200,400]
polyA:control
polyA:ORF
ORF:control
mismatches ratio
positions from polyA site (bps)

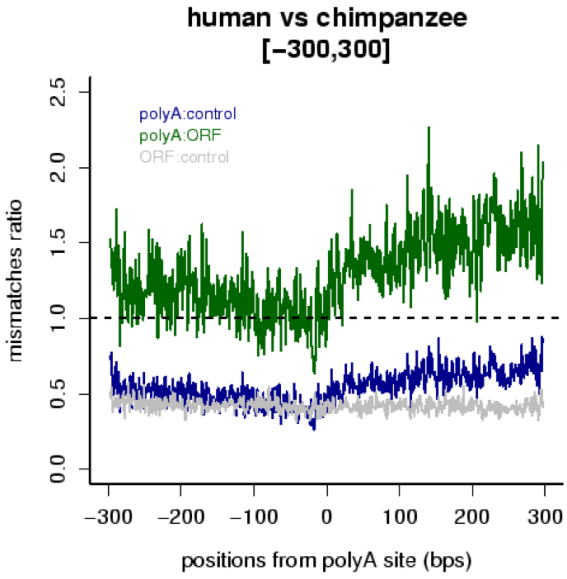
E F


human vs chimpanzee
[−400,200]
polyA:control
polyA:ORF
ORF:control
mismatches ratio
0.0
0.5
1.0
1.5
2.0
2.5
−400
−300
−200
−100
0
100
200
positions from polyA site (bps)

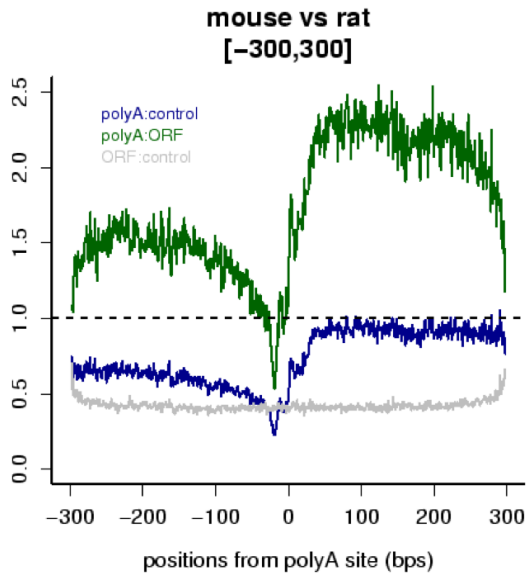
G


mouse vs rat
[−400,200]
polyA:control
polyA:ORF
ORF:control
0.0
0.5
1.0
1.5
2.0
2.5
−400
−300
−200
−100
0
100
200
positions from polyA site (bps)

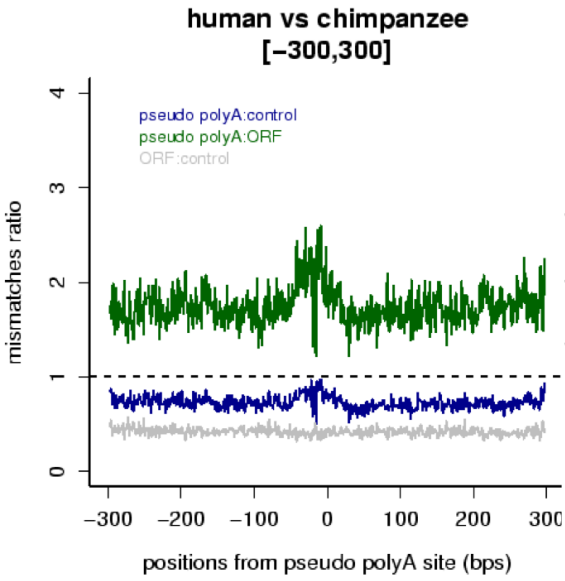
H


human vs chimpanzee
[−300,300]
polyA:control
polyA:ORF
ORF:control
mismatches ratio
0.0
0.5
1.0
1.5
2.0
2.5
−300
−200
−100
0
100
200
300
positions from polyA site (bps)

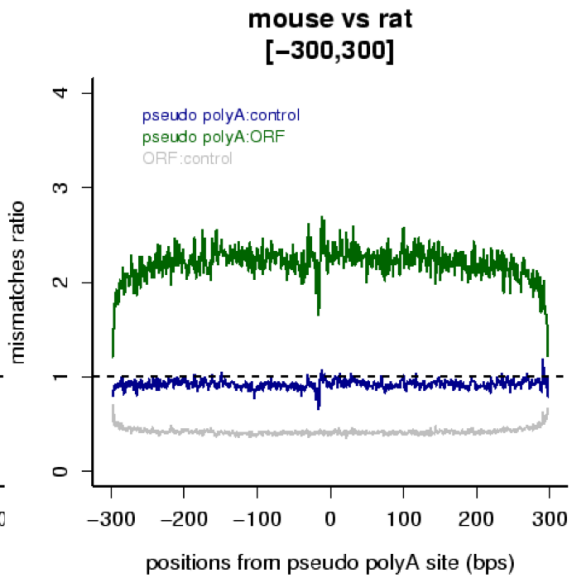
I


mouse vs rat
[−300,300]
polyA:control
polyA:ORF
ORF:control
0.0
0.5
1.0
1.5
2.0
2.5
−300
−200
−100
0
100
200
300
positions from polyA site (bps)

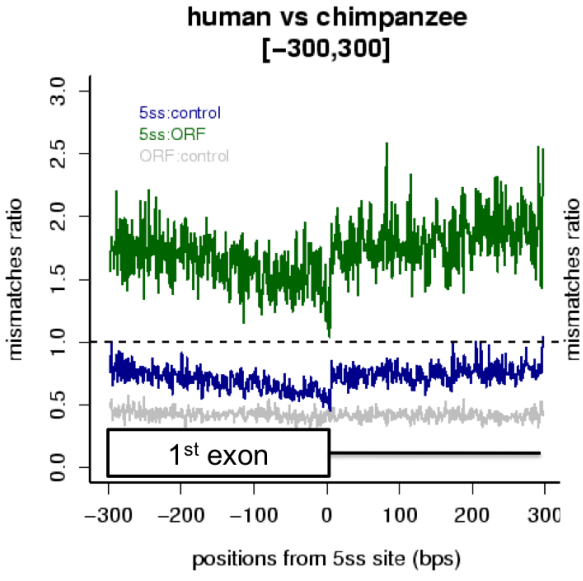
J


human vs chimpanzee
[−300,300]
pseudo polyA:control
pseudo polyA:ORF
ORF:control
mismatches ratio
positions from pseudo polyA site (bps)

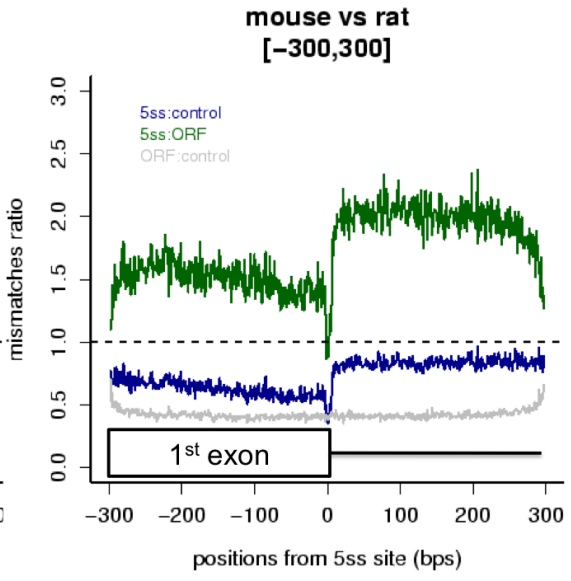
K


mouse vs rat
[−300,300]
pseudo polyA:control
pseudo polyA:ORF
ORF:control
mismatches ratio
positions from pseudo polyA site (bps)

L


human vs chimpanzee
[−300,300]
5ss:control
5ss:ORF
ORF:control
mismatches ratio
1st exon
positions from 5ss site (bps)

M


mouse vs rat
[−300,300]
5ss:control
5ss:ORF
ORF:control
mismatches ratio
1st exon
positions from 5ss site (bps)

N

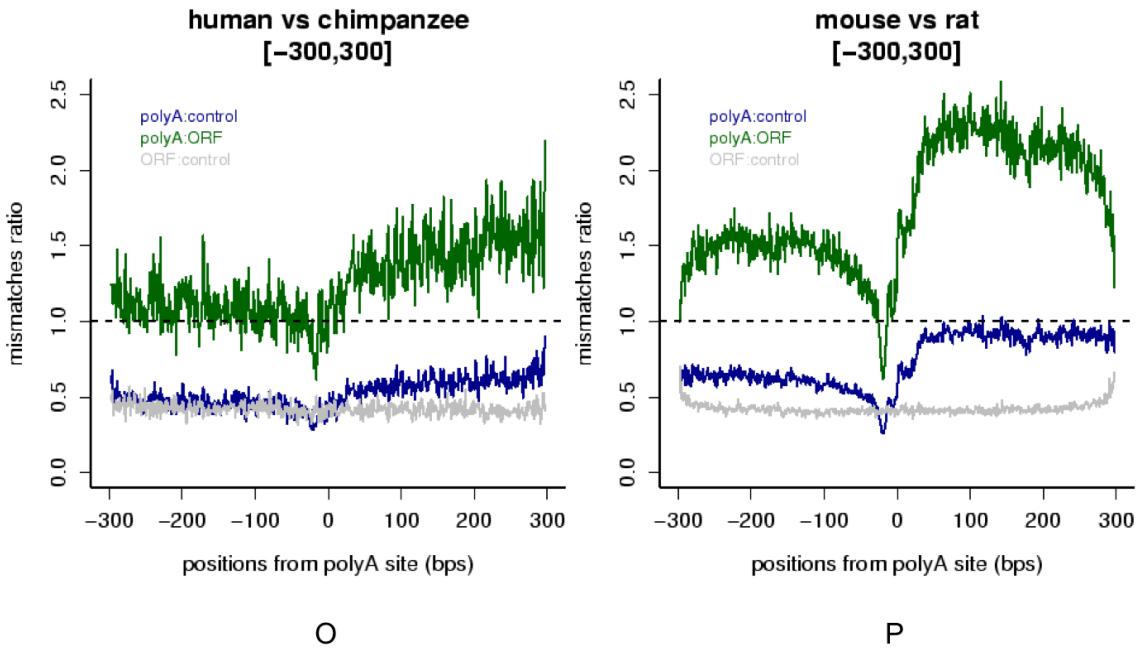


Figure 2.2 Mismatch ratio in PAS flanking region between close species. A-B) Mismatch ratio variation for region [-300,+300], C-D) the PAS flanking region versus 3' UTR, E-F) mismatch ratio variation for region from 200 nts upstream to 400 nts downstream, G-H) mismatch ratio variation for region from 400 nts upstream to 200 nts downstream, I-J) PAS flanking region for single PAS genes only, K-L) pseudo PAS intronic sequences, M-N) mismatch ratio variation at the first splicing donor site (5' ss), O-P) analysis of non-overlapping genes.

In Figure 2.2, the blue line represents the mismatch ratio between the real PAS and the intergenic control sequence, similarly, for the green line except that

the control is changed to the ORF. The grey line represents the comparison between the two types of control sequences i.e. ORF versus intergenic.

As shown in Figure 2.2A, the mismatch ratio of real PAS sequence versus intergenic sequence (blue line) is <1 for the entire region indicating a stronger selection pressure in the PAS sequences than in the intergenic sequences. However, the experienced selection pressure is not as strong as the pressure to preserve the ORF (green line) except for the region ~30 nts upstream of the PAS, which is the preferred location of the poly(A) signal. Such a pattern becomes more explicit in the comparison between mouse and rat plotted in Figure 2.2B as mouse and rat diverged about 18 million years ago (mya) [Rat genome sequencing 2004] while human and chimpanzee diverge only 6 mya. In addition, the region upstream of the poly(A) signal not only experienced a stronger selection pressure than the region downstream but also a wider range as the downstream selection pressure vanishes after ~50 nts from the PAS as shown in Figure 2.2B. This asymmetrical pressure is not caused by any possible uneven selection pressure in the two types of control sequences along the considered region because the mismatch ratio line (grey line) for ORF versus intergenic stays at a steady level (~0.5) across the entire region. In order to determine the range of the selection pressure on the upstream region starting from the poly(A) signal, the first 600 nts of 3' UTR was chosen as control rather than ORF. The reason to support the use of 3' UTR is that the PAS flanking region is, in fact, part of the 3' UTR therefore it should be subjected to similar selection pressure. One assumption is that any difference observed in the region

flanking the poly(A) signal, no matter high or low, is related to PAS activity. As shown in Figure 2.2C and D, when the 3'UTR is taken as the control, the mismatch ratio (green) line asymptotically approaches 1 in the upstream direction and becomes flat by ~200 nts upstream of the PAS. The mismatch ratio between the 3' UTR and the intergenic region (Figure 2.2C and D) is similar to ORF versus intergenic in Figure 2.2A and B indicating the 3' UTR does not exhibit uneven selection pressure across the considered region. The data also indicate 3' UTRs do experience a lower mutation rate than intergenic sequences, in agreement with prior studies that many expression related regulatory elements are located in the 3' UTR [Xie et al 2005] but with less clear positional preference.

b) Justification of close species comparison method

Although the above close species analysis supports the existence of selection pressure flanking the PAS, it is prudent to do several types of control analysis to rule out alternative explanations such as artifacts inherent in the computation methods and alternative biological mechanisms. One well-known artifact is the NCBI-BLAST algorithm favors alignment of sequences in the middle of an alignment over sequences near the edges. To examine this, figures 2.2E to H were generated that repeated the A to B plots but with the region of interest shifted upstream or downstream by 200 nts. As the pattern in plots E to H remains largely unchanged, alignment bias can be ruled out in this study. To examine whether the selection pressure pattern depends on proximal repeats of PAS, only the single PAS genes were selected to produce figure 2.2I and J. As

shown, the same pattern persists in both close species pairs. Another possible reason for the selection pressure pattern may be caused by the highly conserved poly(A) signal AWUAAA. To examine this, a set of 600-nt long intronic sequences (17,080 from human, 8,799 from mouse) with AWUAAA positioned ~270 nts from the 5' end was randomly sampled. We dub this the pseudo PAS sequence set and more details on how to collect them can be found in Appendix C. Analysis of this sequence data set is shown in Figure 2.2K and L, where it is clear that these sequences have no selection pressure pattern. The spike located 30 nts near the middle indicates the aligned poly(A) signals AWUAAA at position 270. Thus, the poly(A) signals themselves failed to reproduce the same pattern exhibited by the real PAS flanking region in plots A and B. Moreover, if the distinct mismatch ratio pattern were solely caused by the highly conserved poly(A) signal, figure 2.2A and B should show a symmetric pattern too. The same analysis was also applied to the 5' splice site (5' ss) region found in the first exon as it is well documented that 5'ss recognition is facilitated by the presence of short sequence elements located immediately upstream of the 5'ss [Fairbrother et al 2002, Wang et al 2004]. These sequence elements, commonly known as exonic splicing enhancers, are targets of serine-rich proteins (SR proteins) [Graveley 2000]. Since 5'ss splicing enhancers are essential for pre-mRNA processing, they must be subjected to positive selection pressure. As shown in Figure 2.2M and N, the mismatch ratio has the lowest value just upstream of the 5'ss, and then rises abruptly immediately after the exon-intron junction in the 5' to 3' direction. Finally, 30% and 38% of human and mouse genes were found to

overlap (<1000nt separation) with a neighboring gene. To examine whether such a gene overlap influences this analysis, the overlapping genes were removed from the initial dataset leaving 12,195 and 5,553 pairs of human-chimpanzee and mouse-rat homologous poly(A) regions. As shown in Figure 2.2O and P, there is no observable difference in the variation of mismatch ratio with respect to the unfiltered sequences (Figure 2.2A and B). Thus this battery of analysis has identified that there is positive selection pressure on sequences within 0-200 nts upstream of the PAS.

3. Discussion

Results show that close species comparison is useful in revealing the different degree of conservation in generally non-alignable regions in remote species. Selection pressure is found to be higher in 3' UTR than intergenic (grey line of Figure 2.2C and D) and intronic sequences (grey line of Figure 2.2K and L). Such selection pressure is uniform for the whole 3' UTR except for the region flanking the PAS. This observation indicates the conservation of position independent sequence motifs and/or nucleotide composition along the 3' UTR. On the other hand, the comparison between mouse and rat (Figure 2.2D) shows the presence of an asymmetrical selection pressure localized in the [-200,+50] region. A similar pattern is reconfirmed in the comparison between ORF and PAS flanking region as shown in Figure 2.2B. Such a finding reveals a longer upstream and a shorter downstream region that may be involved in polyadenylation than reported previously [Legendre et al 2003, Tian et al 2005, Hu et al 2005]. Even though the requirement of upstream poly(A) signal and

downstream U/GU-rich region are well established, the asymmetrical selection pressure presence in up to 200 nts upstream of the PAS suggests the existence of other unknown cis elements. Unlike 5'ss sequences, a sharp fall in the mismatch ratio is not observed in the upstream region (Figure 2.2M and N). Three possible explanations may account for the lack of a sharp fall. First, the upstream binding factor(s) (not CPSF-160) is flexible in acting at a distance. Second, the selection pressure for the region [-200,-100] is gene specific rather than basal and thus can only be seen when comparing orthologous genes as done here. Third, unlike frameshift mutations caused by mis-splicing, no severe drawback would be expected if cleavage occurs at a slightly different position. According to previous studies [Legendre et al 2003, Tian et al 2005, Hu et al 2005], one characteristic of the upstream region is the gradual elevation of uracil composition in the 5' to 3' direction in the region [-100,-30]. The maximum increment is about 5% which happens immediately 5' of the poly(A) signal. A stronger PAS possesses higher uracil content upstream than the weaker one. However, the entire human and mouse 3' UTRs, except the region 50 nts immediately after the stop codon and the last 100 nts at 3' the end, are evenly enriched with uracil (~29%) and adenine (~27%) (Appendix D). A similar observation has also been reported in diverse species [Graber et al 1999]. If the polyadenylation machinery solely relies on a uracil-rich signal, false signals in the 3' UTR should appear more frequent than the real one. Even taking the two canonical poly(A) signals into account to enhance specificity, such an idea helps little to improve the recognition of PAS as poly(A) signals occur ubiquitously.

Close to 3.4 and 2.2 million canonical poly(A) signals were found in human and mouse introns, respectively. Examination of the region [-500,+500] in those intronic sequences show they contain 30% A and T, which is similar to the 3' UTR in terms of nucleotide composition. Hence, additional gene-specific cis elements may be needed to define the PAS. (Details about the recognition of true PAS will be discussed in chapter 3).

In summary, close species comparison has revealed biased selection pressure flanking the PAS, which is the highest within the entire 3' UTR. The proximity of such selection pressure surrounding the PAS has inevitably led us to associate it to polyadenylation. This result leads us to investigate further into the extent of conservation among distant species at the level of the individual gene.

C. Identification of conserved fragments (CFs) in human, mouse, cow, and platypus

Previous attempts were made to identify enriched short sequence motifs (6-10 nts) in the <100 nt upstream region of PAS across all genes [Graber et al 1999, Hu et al 2005, Hutchins et al 2008]. The majority of these upstream elements (USEs) were of low complexity in composition and their function was proposed be related to the 3' end processing/polyadenylation. However, their potency was also found to be position dependent such as U-rich elements [Danckwardt et al 2007] that can regulate polyadenylation for up to 100 nts upstream of the PAS [Zhu et al 2007], features consistent with the conspicuous enrichment of uracil within 40 nts upstream of the PAS [Legendre et al 2003,

Tian et al 2005]. The close species comparison presented earlier, revealed the presence of selection pressure farther than 100 nts upstream, namely up to 200 nts, from the PAS, supporting the existence of other non-repetitive cis elements upstream of the PAS. Although previous approaches were successful in capturing the enrichment of short and fixed size sequence motifs at the 3' end of the transcript, such approaches neglect gene-specific elements. Here, I report on gene specific USEs in several diverse mammalian species. Four evolutionarily distant mammalian species were chosen for this study viz. human, mouse, cow and platypus. Results show that long conserved fragments (CFs) (30-500 nts) flanking the PAS are widespread. But little is known about their biological function. This finding will help to identify novel experimental targets, which may shed light on the regulatory role of these conserved PAS flanking regions in PAS choice and polyadenylation regulation.

1. Methods

Four species were chosen in this analysis viz. human, mouse, cow and platypus. Gene homologous information (based on ORF) of human, mouse and cow was obtained from the NCBI HomoloGene database [HomoloGene 2009]. As the genome of platypus was completed only recently, little expression data is available to obtain its homologous information with other species. To circumvent this, human PAS flanking sequences were used to search against the platypus genome in order to identify homologous regions in platypus. Since two different ways were used to obtain the homologous information, the four mammalian species were divided into two homologous groups, namely HMC, which was

composed of human, mouse and cow, and HMCP, which contained all four species.

To explore the conservation of the region that spans the region [-500,+500] while avoiding the influence of the ORF, genes possessing 3'UTRs shorter than 500 nts were dropped from the dataset. Low complexity and repeat fragments were removed from the analysis using RepeatMasker [Smit et al 2004]. The multiple sequence alignment tool T-COFFEE [Notredame et al 2000] was then used to align the PAS flanking regions for each orthologous group. A score value, in the range of 0 to 100, was returned for each alignment, where 0 and 100 represents no and perfect alignment, respectively. Based on the alignment report, CF was extracted from each orthologous gene group, and duplicated fragments were eliminated if the gene possesses multiple closely-spaced PAS at the 3'UTR. A 15-nt sliding window was used to scan the alignment base by base. A "good" alignment was defined to be ≤3 mismatches (80% identity) and overlapping of good windows were then stitched together to form the CF.

2. Results

a) Percentage of alignment of poly(A) flanking regions among remote mammalian species

The multiple alignment program T-COFFEE was used to align 10,765 and 5,362 orthologous gene groups in HMC and HMCP, respectively. The relationship between the percentage of alignment by position was plotted

separately by alignment score as shown in Figure 2.3 below. Two alignment score thresholds were used viz. 50 and 70. According to my experience, alignment score above 50 generally indicates the presence of long fragments (>30 nts). Note that higher alignment scores are often associated with longer and/or multiple CFs.

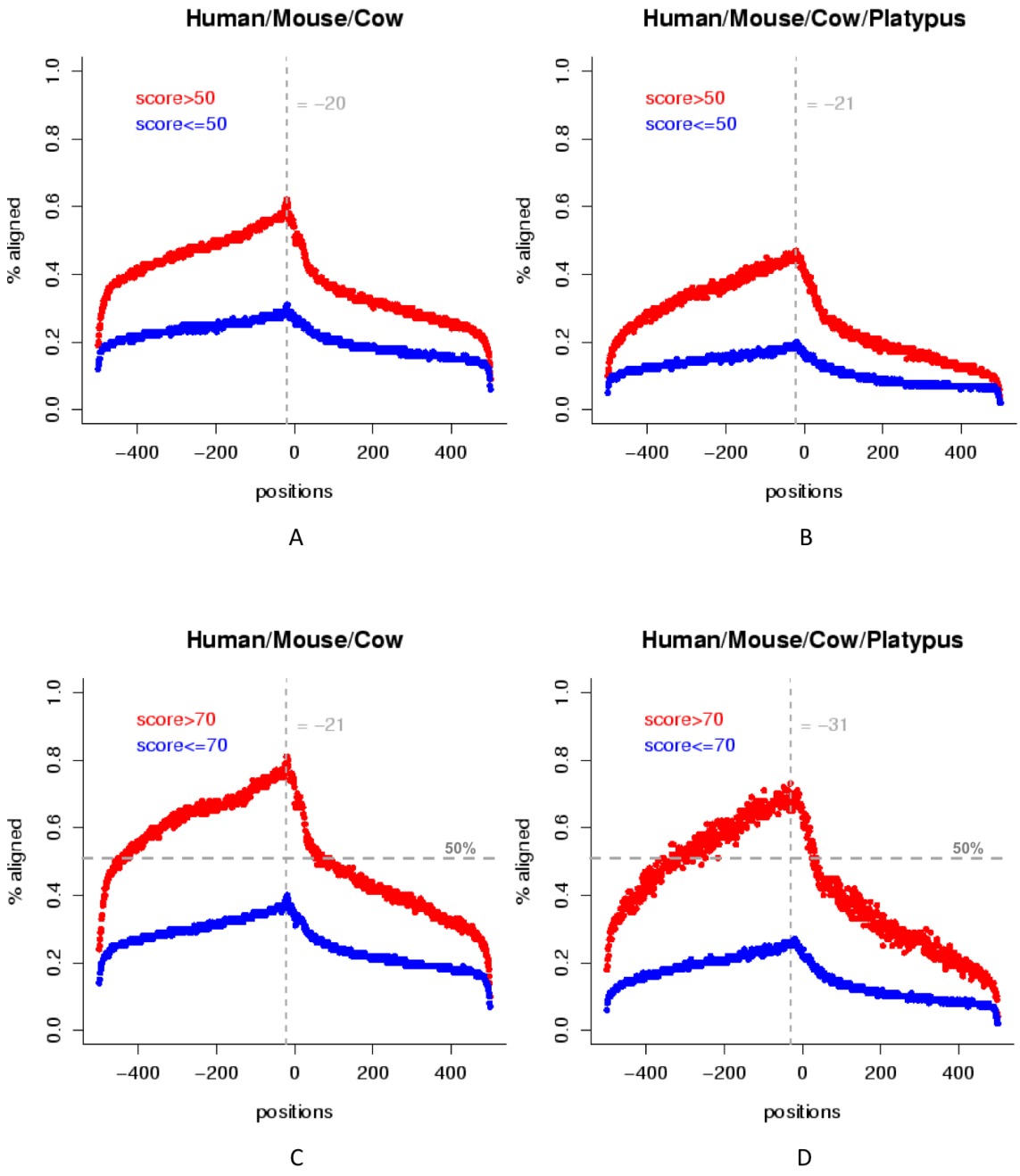


Figure 2.3 Percentage of alignment along the flanking positions at around PAS. Red and blue lines denote high and low scoring groups, respectively. A) HMC group with threshold 50, B) HMCP with threshold 50, C) HMC with threshold 70, D) HMCP with threshold 70.

Red and blue lines denote high and low scoring groups respectively. Each line represents the variation in percentage of genes containing the same nucleotide as human along the flanking region of PAS.

5,261 out of 10,765 genes or 49% were found to achieve higher than 50 alignment score in the HMC group (Figure 2.3A). In the HMCP group, 2,668 out 5,362 genes or 50%, similar to the HMC group were found to exceed alignment score 50. When a more stringent threshold, 70, was adopted, the number of genes dropped to 2,160 (20%) for the HMC group and the HMCP group dropped even more to 629 genes (12%). But raising the threshold resulted in higher percentage of alignment (compare Figure 2.3A and C or between B and D).

Not surprisingly, for both high and low scoring groups, the best alignment was attained at around 21 nts upstream from the PAS, which is the preferred location of the poly(A) signal. Even the peak occurred at 31 nts instead of 21 nts upstream in the HMCP group with threshold 70 (Figure 2.3D), the percentages of alignment between them differ by 3 percentage points only. The trend of the plot resembles that of the close species comparison method where selection pressure is asymmetrical, i.e. higher in strength and range in the upstream than the downstream region. However, the degree of alignment seems to extend farther than 200 nts upstream for a subset of high scoring genes as revealed in Figure 2.3 C and D. 1,080 of 2,160 orthologous HMC-group genes show a high degree of alignment, but not necessarily in one continuous stretch, for up to 400 nts upstream. This observation provides intriguing indication to look into the conservation of the non-coding sequence of each gene.

b) Identification of Conserved Fragments

Two independent methods presented here suggest the conservation pressure is prominent upstream rather than downstream of the PAS, thus the analysis concentrated on the upstream region only. Based on the multiple alignment results, CFs were extracted from genes with alignment scores >50, longer than 30 nts, and limited to one fragment per gene. Altogether, 3,315 and 1,265 non-redundant conserved upstream fragments were discovered in HMC and HMCP groups, respectively. The distribution of their lengths is shown in Figure 2.4.

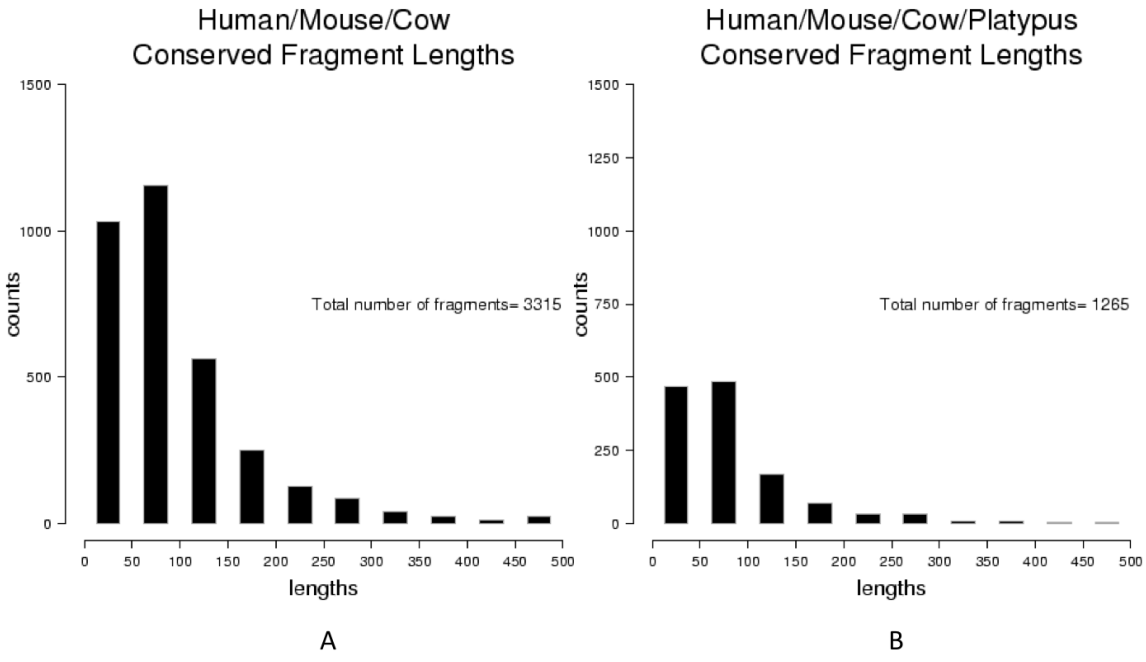


Figure 2.4 Distribution of length of human conserved upstream fragments. A) in HMC group, B) in HMCP group.

As shown in Figure 2.4A, almost two-thirds of the CFs was between 30-100 nts long in the HMC group. Several CFs were found that are 400-500 nts long (Figure 2.4A and B). As expected, smaller numbers of CFs were found in the HMCP group however both groups exhibit similar distribution (Figure 2.4A and B). Next, CF distance (based on 3' end of CF) from the PAS, the relationship between fragment length, and proximity to the PAS were examined. Figure 2.5 below displays the distribution of the distance of these human CFs from the PAS in both the HMC and HMCP groups.

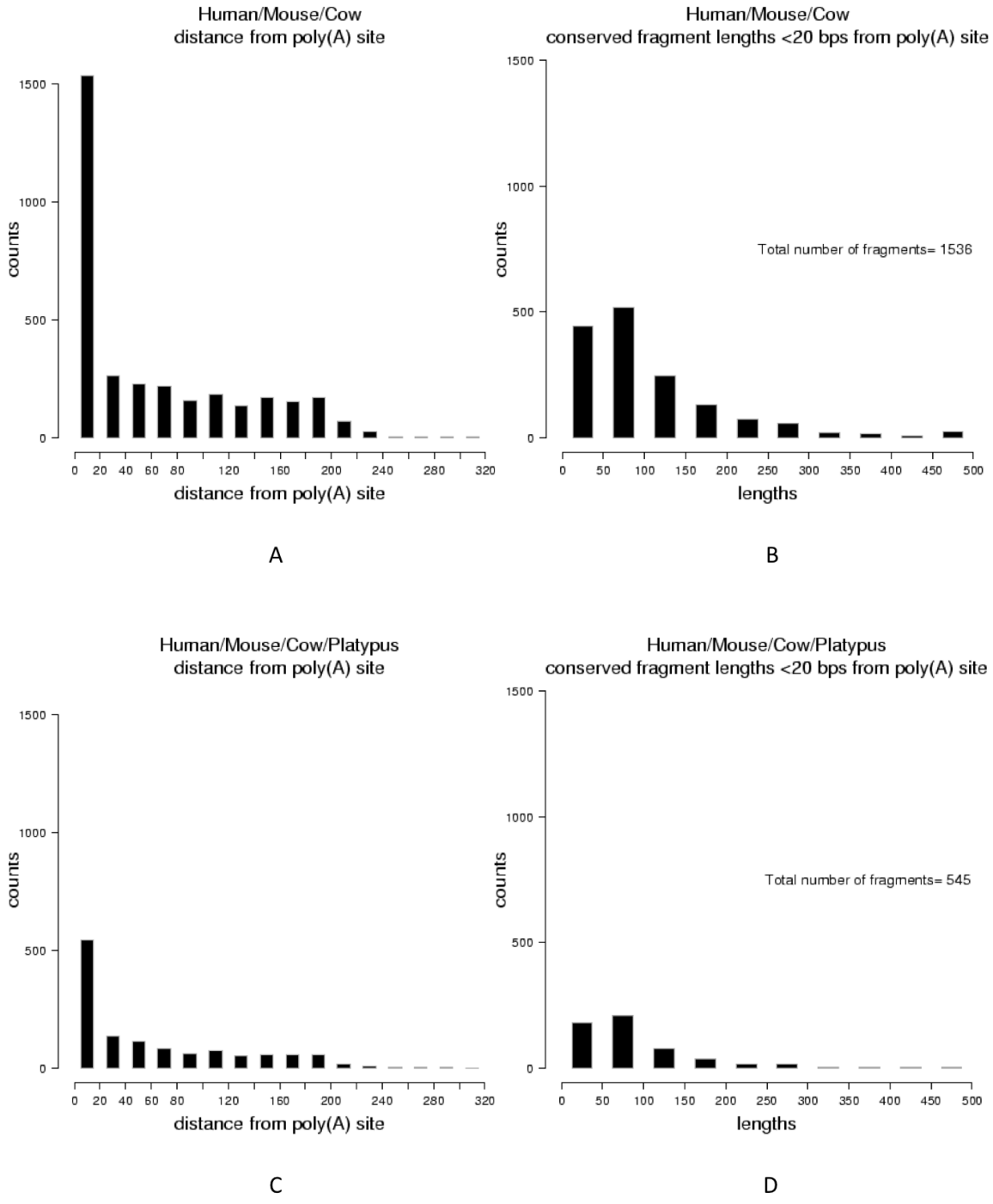
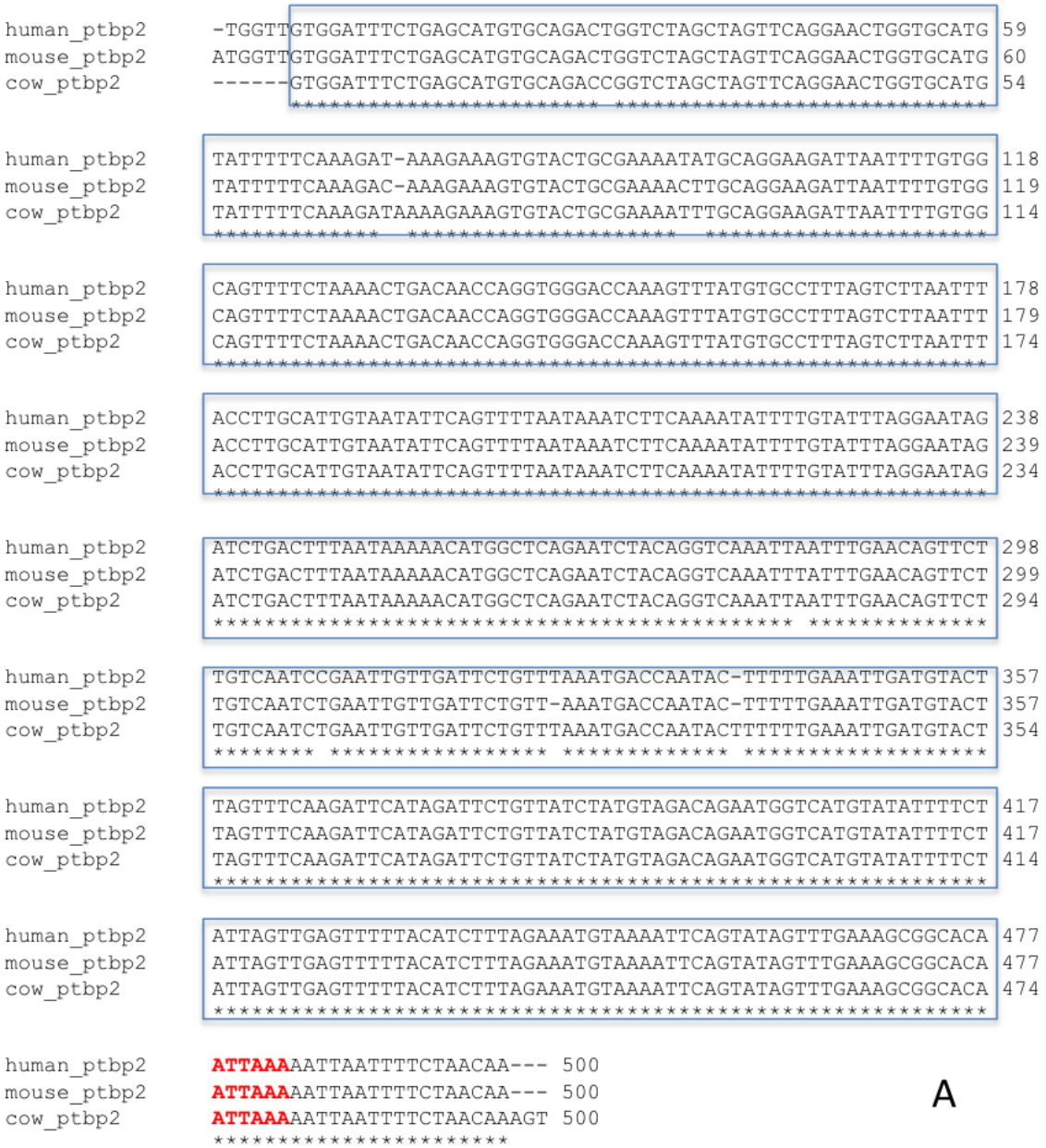


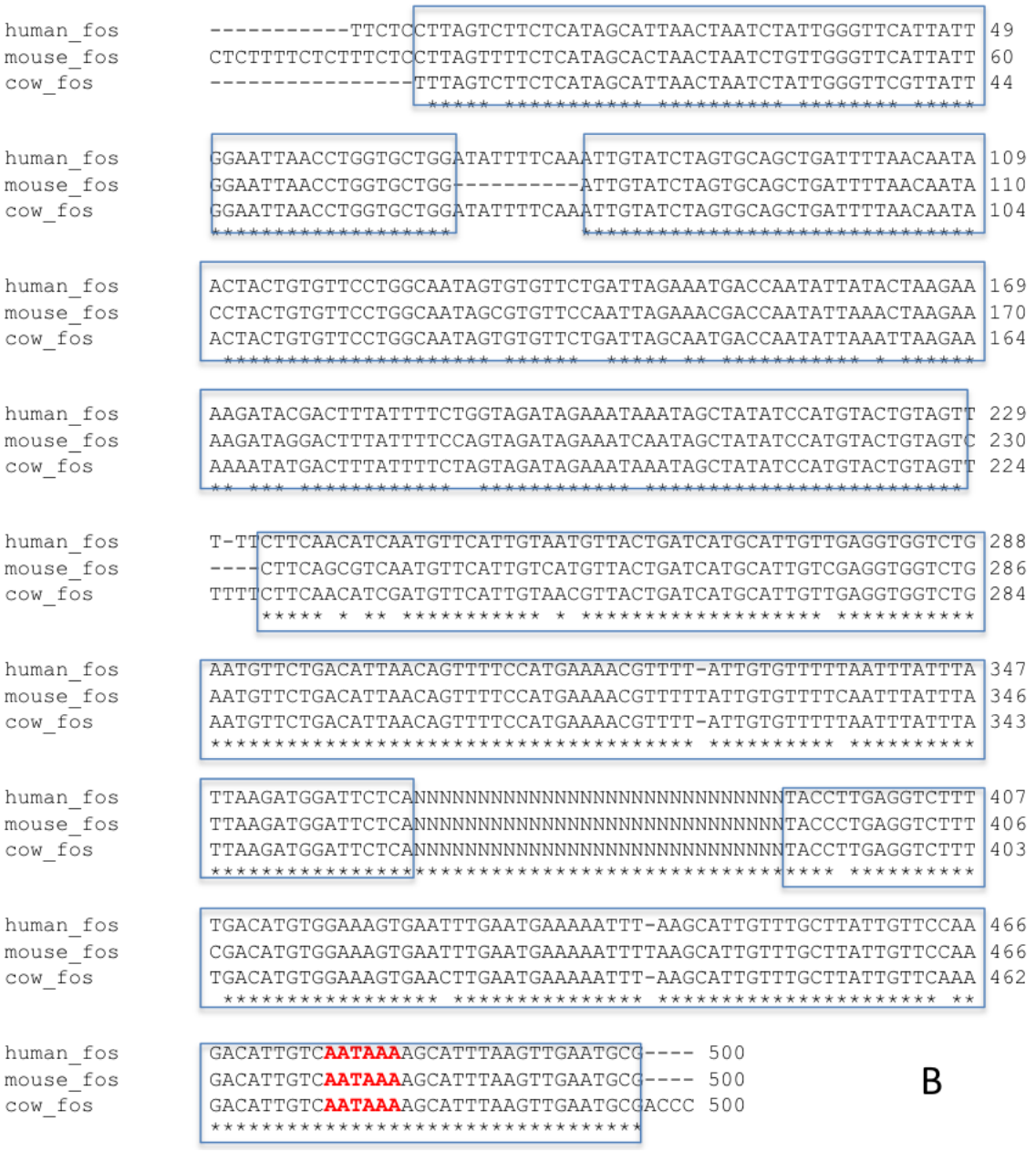
Figure 2.5 Distance of human CFs (based on 3' end of CF) from the PAS. A) distance of CF from PAS in the HMC group, B) length of CF <20 nts from the PAS in HMC group, C) distance of CF from PAS in the HMCP group, D) length of CF <20 nts from PAS in HMCP group.

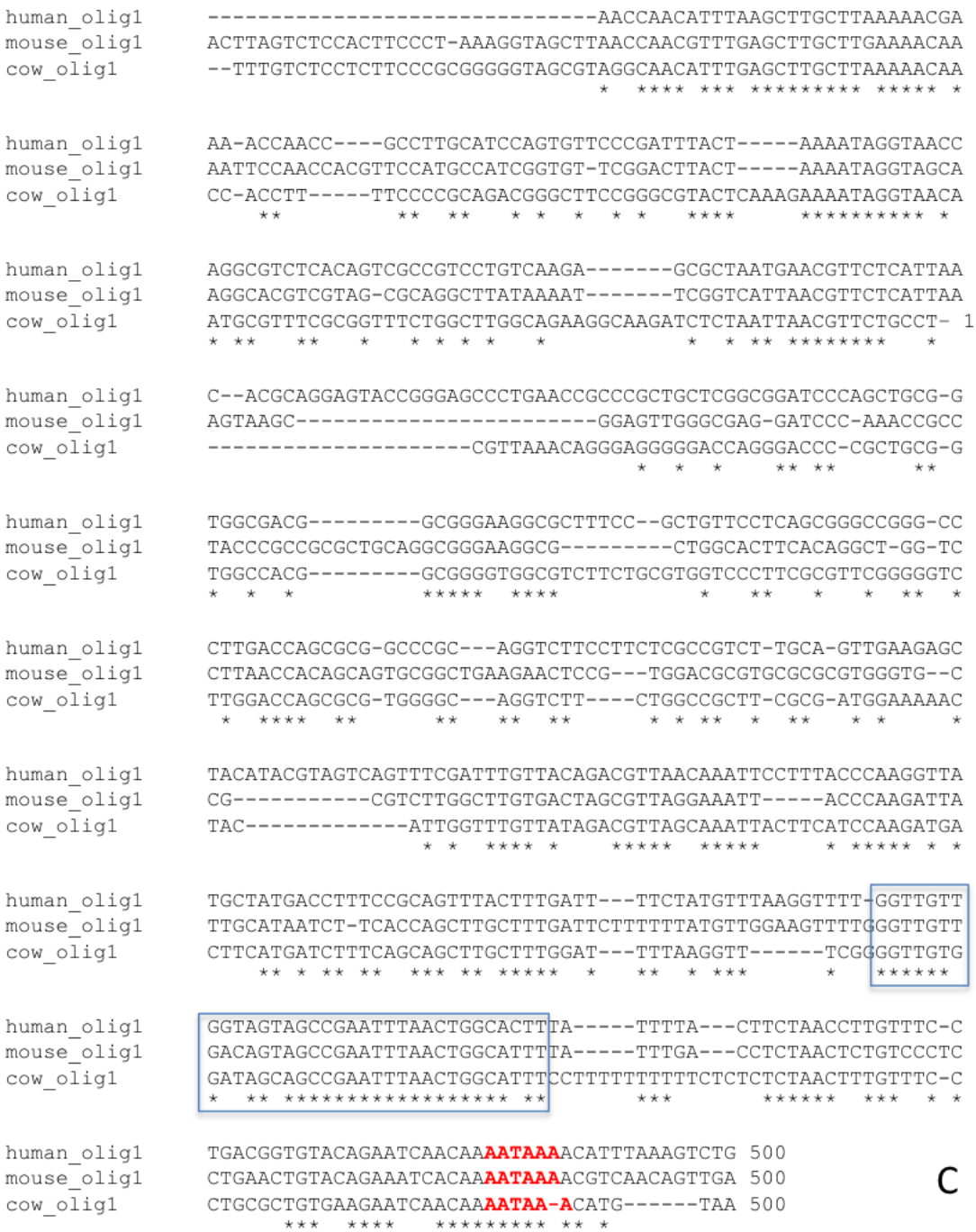
Almost half of the CFs were found to reside within 20 nts from the PAS in the HMC group (Figure 2.5A), and the remaining CFs were uniformly distributed along the upstream region, suggesting there is no particular relation between the size of the CF and proximity to the PAS. A consistent picture is found in both the HMC and HMCP groups (Figure 2.5C). Furthermore, the length of those CFs that were within 20 nts from the PAS were analyzed as shown in Figure 2.5B and D. Their distribution closely resembles the overall distribution of CFs where the majority of them were between 30-100 nts long.

c) Examples of Conserved Fragments

A sample of alignments and CFs for three genes will be illustrated viz. polypyrimidine tract binding protein 2 (PTBP2), FBJ murine osteosarcoma viral oncogene homolog (FOS), and oligodendrocyte transcription factor 1 (OLIG1). These three genes manifest different degrees of conservation near the PAS like PTBP2 and FOS are extreme examples as they contain 400 to nearly 500-nt long CFs starting from the PAS in the 5' direction. PTBP2 is reported to control the assembly of other splicing regulatory proteins. It binds to intronic polypyrimidine tracts during splicing. PTBP2 is similar to PTBP1 except for the fact that it is abundant mainly in brain. In Figure 2.6A, it is evident there is a continuous stretch of CFs among human, mouse and cow including the poly(A) signal. It is rich in A and T but not of low complexity as repeated and low complexity regions were removed before alignment. The conservation is amazing which is even higher than the coding sequence.


human_ptbp2 -TGGTTGTGGATTTCTGAGCATGTGCAGACTGGTCTAGCTAGTTCAGGAACTGGTGCATG 59
mouse_ptbp2 ATGGTTGTGGATTTCTGAGCATGTGCAGACTGGTCTAGCTAGTTCAGGAACTGGTGCATG 60
cow_ptbp2 ------GTGGATTTCTGAGCATGTGCAGACCGGTCTAGCTAGTTCAGGAACTGGTGCATG 54
human_ptbp2 TATTTTTCAAAGAT-AAAGAAAGTGTACTGCGAAAATATGCAGGAAGATTAATTTTGTGG 118
mouse_ptbp2 TATTTTTCAAAGAC-AAAGAAAGTGTACTGCGAAAACTTGCAGGAAGATTAATTTTGTGG 119
cow_ptbp2 TATTTTTCAAAGATAAAAGAAAGTGTACTGCGAAAATTTGCAGGAAGATTAATTTTGTGG 114
human_ptbp2 CAGTTTTCTAAAACTGACAACCAGGTGGGACCAAAGTTTATGTGCCTTTAGTCTTAATTT 178
mouse_ptbp2 CAGTTTTCTAAAACTGACAACCAGGTGGGACCAAAGTTTATGTGCCTTTAGTCTTAATTT 179
cow_ptbp2 CAGTTTTCTAAAACTGACAACCAGGTGGGACCAAAGTTTATGTGCCTTTAGTCTTAATTT 174
human_ptbp2 ACCTTGCATTGTAATATTCAGTTTTAATAAATCTTCAAAATATTTTGTATTTAGGAATAG 238
mouse_ptbp2 ACCTTGCATTGTAATATTCAGTTTTAATAAATCTTCAAAATATTTTGTATTTAGGAATAG 239
cow_ptbp2 ACCTTGCATTGTAATATTCAGTTTTAATAAATCTTCAAAATATTTTGTATTTAGGAATAG 234
human_ptbp2 ATCTGACTTTAATAAAAACATGGCTCAGAATCTACAGGTCAAATTAATTTGAACAGTTCT 298
mouse_ptbp2 ATCTGACTTTAATAAAAACATGGCTCAGAATCTACAGGTCAAATTTATTTGAACAGTTCT 299
cow_ptbp2 ATCTGACTTTAATAAAAACATGGCTCAGAATCTACAGGTCAAATTAATTTGAACAGTTCT 294
human_ptbp2 TGTCAATCCGAATTGTTGATTCTGTTTAAATGACCAATAC-TTTTTGAAATTGATGTACT 357
mouse_ptbp2 TGTCAATCTGAATTGTTGATTCTGTT-AAATGACCAATAC-TTTTTGAAATTGATGTACT 357
cow_ptbp2 TGTCAATCTGAATTGTTGATTCTGTTTAAATGACCAATACTTTTTTGAAATTGATGTACT 354
human_ptbp2 TAGTTTCAAGATTCATAGATTCTGTTATCTATGTAGACAGAATGGTCATGTATATTTTCT 417
mouse_ptbp2 TAGTTTCAAGATTCATAGATTCTGTTATCTATGTAGACAGAATGGTCATGTATATTTTCT 417
cow_ptbp2 TAGTTTCAAGATTCATAGATTCTGTTATCTATGTAGACAGAATGGTCATGTATATTTTCT 414
human_ptbp2 ATTAGTTGAGTTTTTACATCTTTAGAAATGTAAAATTCAGTATAGTTTGAAAGCGGCACA 477
mouse_ptbp2 ATTAGTTGAGTTTTTACATCTTTAGAAATGTAAAATTCAGTATAGTTTGAAAGCGGCACA 477
cow_ptbp2 ATTAGTTGAGTTTTTACATCTTTAGAAATGTAAAATTCAGTATAGTTTGAAAGCGGCACA 474
human_ptbp2 ATTAAAAATTAATTTTCTAACAA--- 500
mouse_ptbp2 ATTAAAAATTAATTTTCTAACAA--- 500
cow_ptbp2 ATTAAAAATTAATTTTCTAACAAAGT 500
A


human_fos ----------TTCTCCTTAGTCTTCTCATAGCATTAACTAATCTATTGGGTTCATTATT 49
mouse_fos CTCTTTTCTCTTTCTCCTTAGTTTTCTCATAGCACTAACTAATCTGTTGGGTTCATTATT 60
cow_fos ----------TTTAGTCTTCTCATAGCATTAACTAATCTATTGGGTTCGTTATT 44
human_fos GGAATTAACCTGGTGCTGGATATTTTCAAATTGTATCTAGTGCAGCTGATTTTAACAATA 109
mouse_fos GGAATTAACCTGGTGCTGG----------ATTGTATCTAGTGCAGCTGATTTTAACAATA 110
cow_fos GGAATTAACCTGGTGCTGGATATTTTCAAATTGTATCTAGTGCAGCTGATTTTAACAATA 104
human_fos ACTACTGTGTTCCTGGCAATAGTGTGTTCTGATTAGAAATGACCAATATTATACTAAGAA 169
mouse_fos CCTACTGTGTTCCTGGCAATAGCGTGTTCCAATTAGAAACGACCAATATTAAACTAAGAA 170
cow_fos ACTACTGTGTTCCTGGCAATAGTGTGTTCTGATTAGCAATGACCAATATTAAATTAAGAA 164
human_fos AAGATACGACTTTATTTTCTGGTAGATAGAAATAAATAGCTATATCCATGTACTGTAGTT 229
mouse_fos AAGATAGGACTTTATTTTCCAGTAGATAGAAATCAATAGCTATATCCATGTACTGTAGTC 230
cow_fos AAAATATGACTTTATTTTCTAGTAGATAGAAATAAATAGCTATATCCATGTACTGTAGTT 224
human_fos T-TTCTTCAACATCAATGTTCATTGTAATGTTACTGATCATGCATTGTTGAGGTGGTCTG 288
mouse_fos ----CTTCAGCGTCAATGTTCATTGTCATGTTACTGATCATGCATTGTCGAGGTGGTCTG 286
cow_fos TTTTCTTCAACATCGATGTTCATTGTAACGTTACTGATCATGCATTGTTGAGGTGGTCTG 284
human_fos AATGTTCTGACATTAACAGTTTTCCATGAAAACGTTTT-ATTGTGTTTTTAATTTATTTA 347
mouse_fos AATGTTCTGACATTAACAGTTTTCCATGAAAACGTTTTTATTGTGTTTTCAATTTATTTA 346
cow_fos AATGTTCTGACATTAACAGTTTTCCATGAAAACGTTTT-ATTGTGTTTTTAATTTATTTA 343
human_fos TTAAGATGGATTCTCANNNNNNNNNNNNNNNNNNNNNNNNNNNNNTACCTTGAGGTCTTT 407
mouse_fos TTAAGATGGATTCTCANNNNNNNNNNNNNNNNNNNNNNNNNNNNNTACCCTGAGGTCTTT 406
cow_fos TTAAGATGGATTCTCANNNNNNNNNNNNNNNNNNNNNNNNNNNNNTACCTTGAGGTCTTT 403
human_fos TGACATGTGGAAAGTGAATTTGAATGAAAAATTT-AAGCATTGTTTGCTTATTGTTCCAA 466
mouse_fos CGACATGTGGAAAGTGAATTTGAATGAAAAATTTTAAGCATTGTTTGCTTATTGTTCCAA 466
cow_fos TGACATGTGGAAAGTGAACTTGAATGAAAAATTT-AAGCATTGTTTGCTTATTGTTCAAA 462
human_fos GACATTGTCAATAAAAGCATTTAAGTTGAATGCG---- 500
mouse_fos GACATTGTCAATAAAAGCATTTAAGTTGAATGCG---- 500
cow_fos GACATTGTCAATAAAAGCATTTAAGTTGAATGCGACCC 500
B


human_olig1 ----------AACCAACATTTAAGCTTGCTTAAAAACGA 29
mouse_olig1 ACTTAGTCTCCACTTCCCT-AAAGGTAGCTTAACCAACGTTTGAGCTTGCTTGAAAACAA 59
cow_olig1 --TTTGTCTCCTCTTCCCGCGGGGGTAGCGTAGGCAACATTTGAGCTTGCTTAAAAACAA 58
human_olig1 AA-ACCAACC----GCCTTGCATCCAGTGTTCCCGATTTACT-----AAAATAGGTAACC 79
mouse_olig1 AATTCCAACCACGTTCCATGCCATCGGTGT-TCGGACTTACT-----AAAATAGGTAGCA 113
cow_olig1 CC-ACCTT-----TTCCCCGCAGACGGGCTTCCGGGCGTACTCAAAGAAAATAGGTAACA 112
human_olig1 AGGCGTCTCACAGTCGCCGTCCTGTCAAGA-------GCGCTAATGAACGTTCTCATTAA 132
mouse_olig1 AGGCACGTCGTAG-CGCAGGCTTATAAAAT-------TCGGTCATTAACGTTCTCATTAA 165
cow_olig1 ATGCGTTTCGCGGTTTCTGGCTTGGCAGAAGGCAAGATCTCTAATTAACGTTCTGCCT- 170
human_olig1 C--ACGCAGGAGTACCGGGAGCCCTGAACCGCCCGCTGCTCGGCGGATCCCAGCTGCG-G 189
mouse_olig1 AGTAAGC----------GGAGTTGGGCGAG-GATCCC-AAACCGCC 199
cow_olig1 ----------CGTTAAACAGGGAGGGGGACCAGGGACCC-CGCTGCG-G 207
human_olig1 TGGCGACG---------GCGGGAAGGCGCTTTCC--GCTGTTCCTCAGCGGGCCGGG-CC 237
mouse_olig1 TACCCGCCGCGCTGCAGGCGGGAAGGCG---------CTGGCACTTCACAGGCT-GG-TC 248
cow_olig1 TGGCCACG---------GCGGGGTGGCGTCTTCTGCGTGGTCCCTTCGCGTTCGGGGGTC 258
human_olig1 CTTGACCAGCGCG-GCCCGC---AGGTCTTCCTTCTCGCCGTCT-TGCA-GTTGAAGAGC 291
mouse_olig1 CTTAACCACAGCAGTGCGGCTGAAGAACTCCG---TGGACGCGTGCGCGCGTGGGTG--C 303
cow_olig1 TTGGACCAGCGCG-TGGGGC---AGGTCTT----CTGGCCGCTT-CGCG-ATGGAAAAAC 308
human_olig1 TACATACGTAGTCAGTTTCGATTTGTTACAGACGTTAACAAATTCCTTTACCCAAGGTTA 351
mouse_olig1 CG----------CGTCTTGGCTTGTGACTAGCGTTAGGAAATT-----ACCCAAGATTA 347
cow_olig1 TAC----------ATTGGTTTGTTATAGACGTTAGCAAATTACTTCATCCAAGATGA 355
human_olig1 TGCTATGACCTTTCCGCAGTTTACTTTGATT---TTCTATGTTTAAGGTTTT-GGTTGTT 407
mouse_olig1 TTGCATAATCT-TCACCAGCTTGCTTTGATTCTTTTTTATGTTGGAAGTTTTGGGTTGTT 406
cow_olig1 CTTCATGATCTTTCAGCAGCTTGCTTTGGAT---TTTAAGGTT------TCGGGGTTGTG 406
human_olig1 GGTAGTAGCCGAATTTAACTGGCACTTTA-----TTTTA---CTTCTAACCTTGTTTC-C 458
mouse_olig1 GACAGTAGCCGAATTTAACTGGCATTTTA-----TTTGA---CCTCTAACTCTGTCCCTC 458
cow_olig1 GATAGCAGCCGAATTTAACTGGCATTTCCTTTTTTTTTTCTCTCTCTAACTTTGTTTC-C 465
human_olig1 TGACGGTGTACAGAATCAACAAAATAAAACATTTAAAGTCTG 500
mouse_olig1 CTGAACTGTACAGAAATCACAAAATAAAACGTCAACAGTTGA 500
cow_olig1 CTGCGCTGTGAAGAATCAACAAAATAA-ACATG------TAA 500
C

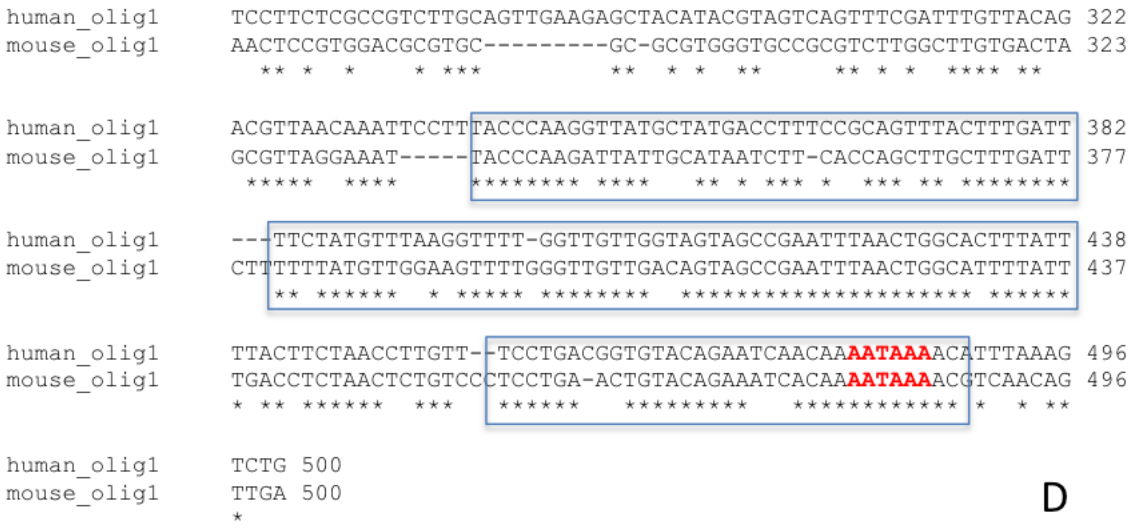


Figure 2.6 Examples of CF. A) polypyrimidine tract binding protein 2 (PTBP2), B) FBJ murine osteosarcoma viral oncogene homolog oncogene (FOS), C) oligodendrocyte transcription factor 1 (OLIG1), D) alignment between human and mouse OLIG1.

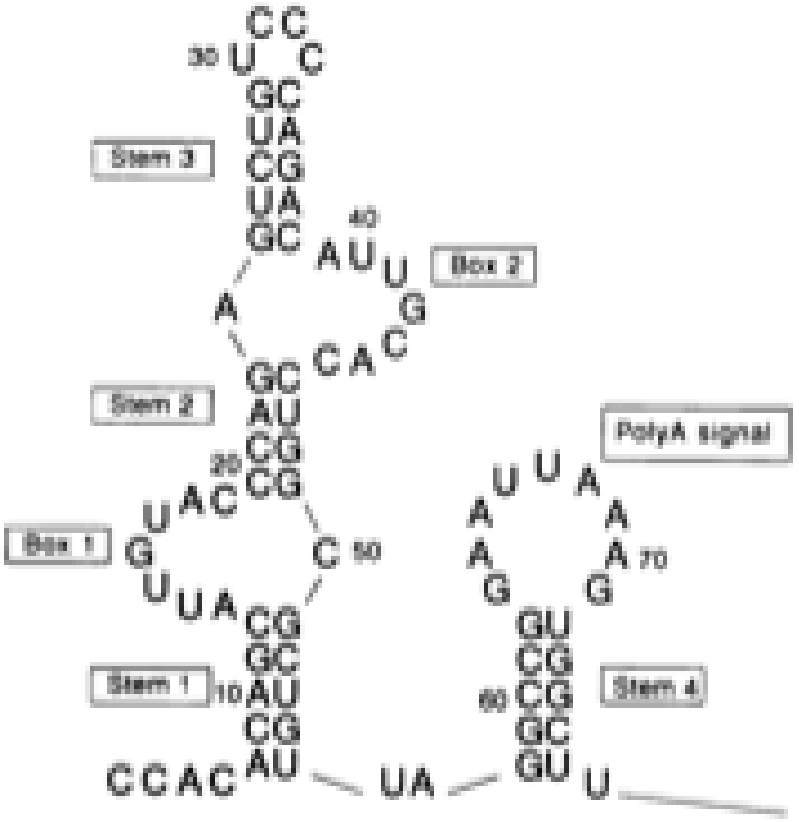
Another example is FOS, which is a well-studied oncogene. It regulates cell proliferation, differentiation and transformation. The total conserved region, excluding the repeat masked fragment, is about 400 nts.

Not all CFs discovered here include the poly(A) signal like PTBP2 and FOS, however, many of them are close to the poly(A) signal. For instance, in Figure 2.6 C above, a 34-nt long CF was found to locate ~100 nts upstream from the PAS. OLIG1 is a transcription factor in oligodendrocyte development [Lu et al 2001] and it plays a role in remyelination after injury [Labombarda et al 2009].

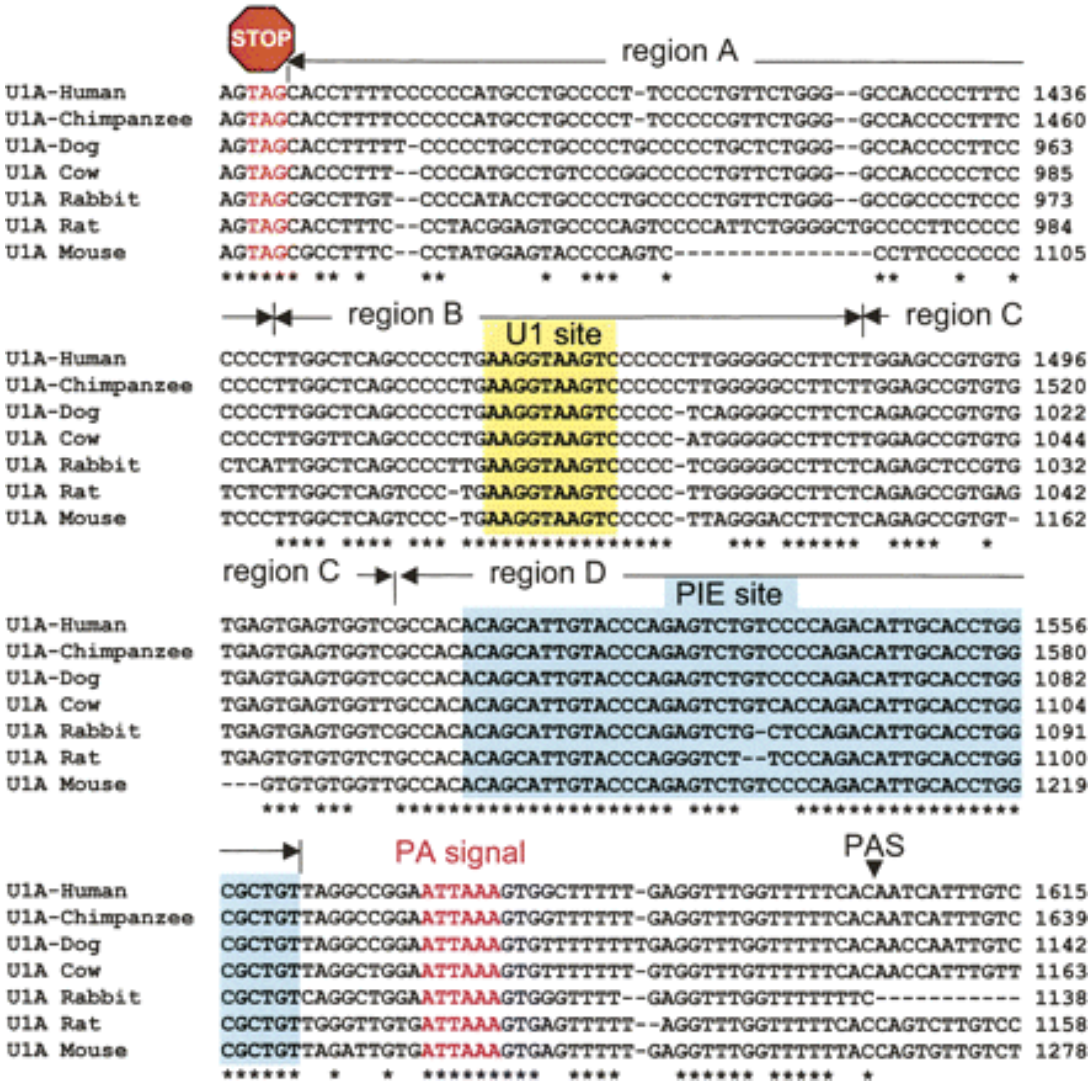
The presence of such a CF is unusual, suggestive of a regulatory function yet to be discovered. Especially, the region of conservation between human and mouse expands significantly as shown in Figure 2.6 D. A full list of alignments of the upstream region among the four mammalian species can be found in Appendix M.

Do these CFs share sequence similarity? To examine this, an exhaustive pairwise comparison was performed among the CFs in order to cluster them into groups by sequence similarity. However, no significant similarity was found among them except for three pairs viz. MORF4L1/MORF4L2, RPL27AP6/RPL27A, and TUBA3C/TUBA4A. Each pair shares about 100+ nts long of highly similar fragments. For these three pairs, their similarity is probably due to gene duplication rather than shared regulatory pathway because their protein sequences share 77-97% identity even though the conservation pressure is extended beyond the coding region.

Besides these examples, the CF of one gene that has been studied experimentally by the Gunderson group is U1A, which is a subunit of the splicing factor U1 snRNP. U1A binds to a specific stem-loop secondary structure in the U1 snRNA. Intriguingly, similar sequence pattern and secondary structure is found in the PAS flanking region of U1A itself as shown in Figure 2.7A and B [van Gelder et al 1993]. An approximately 53-nt long conserved fragment, called the polyadenylation inhibition element (PIE), is conserved among mammalian U1A genes (highlighted in blue in Figure 2.7B).


Stem 3
Box 2
Stem 2
PolyA signal
Box 1
Stem 1
Stem 4
30
40
50
60
70
10

A



B

Figure 2.7 Conservation of U1A PAS flanking region among mammals. A) Secondary structure of the PIE element. Adopted from [van Gelder et al 1993], B) Multiple alignment of U1A PAS flanking regions in seven mammals. Adopted from [Guan F, Coratozzolo R, Goraczniak R, Ho ES, Gunderson SI 2007]

PIE has been reported to serve an auto-regulatory role in U1A expression [Boelens et al 1993, Gunderson et al 1993, 1997]. Two molecules of U1A bind to PIE in its own mRNA to exert inhibition activity on poly(A) polymerase (PAP). NMR and biochemical methods show the inhibition activity was delivered through the helix C located at the N-terminal of the U1A's RNA binding domain [Gunderson et al 1997, Varani et al 2000].

In addition to the highly conserved PIE, a shorter (11 nts) but conserved 5'ss-like fragment was found upstream of PIE (Figure 2.7B, highlighted in yellow) [Guan et al 2007]. This CF is dubbed the U1 site in order to differentiate it from splicing function. As discussed previously in Chapter 1, the binding of U1 snRNP to the U1 site in the 3' UTR of a gene can inhibit polyadenylation via the U1-70K subunit, which leads to the degradation of pre-mRNA in the nucleus. Owing to that, the Gunderson group has studied the role and relationship of the two distinct repression elements in U1A. The conserved U1 site was suggested to be with a secondary RNA structure in the stem part of a stem-loop structure. PIE alone was able to exhibit inhibition activity however the U1 site alone was not. When PIE was disrupted, the binding of U1 snRNP to the U1 site alone manifested weak inhibitory effect. When both PIE and U1 sites were active, inhibition was stronger than PIE alone, indicating a synergetic effect of the two sites. The cooperative work by the PIE and the U1 sites may entail evolutionary advantage in repression as compared to using a single site.

D. Discussion

By taking advantage of close and distant genomic information, the presence of asymmetrical evolutionary pressure flanking the PAS is revealed. The preserved 200-nt upstream region but not the downstream region is likely to function as a transcription termination signal. Although previous work has shown the downstream (~800 nts after PAS) G-rich pause element MAZ_4 in human C2, and co-transcriptional cleavage (CoTC) in human β-globin are essential for transcription termination [Gromak et al 2006], these are likely to be gene-specific functions. Except for the two highly conserved poly(A) signals, no sequence consensus can be found in the upstream region besides a high elevation of uracil content. By aligning the PAS flanking region of orthologous genes among four distant mammalian species, 3,315 and 1,265 evolutionarily conserved non coding fragments (>30 nts long), one per gene, were identified in HMC and HMCP groups, respectively. They represent 31% and 24% of the orthologous genes in the HMC and HMCP groups respectively. As shown in Figure 2.4, large numbers of them are longer than the well-studied AU-rich, U-rich, G-rich and C-rich regions, which regulate mRNA stability within their target proteins.

The approach discussed here complements previous work to search for overrepresented short and fixed length cis elements of polyadenylation [Graber et al 1999, Hu et al 2005, Hutchins et al 2008]. Previous work may be predisposed with the model that these cis elements are binding targets of one or two factors. But the long CF reported here may play a different role as RNA protein recognition sites are usually short. A recent study has shown nucleosome

depletion at around [-100,+100] region [Spies et al 2009]. Double-stranded homopolymeric stretches of deoxyadenosine (10-20 nts) [Segal et al 2009], poly(A) signal and T-rich content are suggested for the diminishing of nucleosomes for both high and low usage PAS. Another important insight comes from the study of ultraconserved elements (UCEs). By comparing human, mouse and rat genomes, 481 identical genomic segments longer than 200 nts were found, and they are also highly conserved in chicken and dog [Bejerano et al 2004]. Some of them function as long-range enhancers [Pennacchio et al 2006], driving development [Woolfe et al 2005], regulating splicing [Lareau et al 2007, Ni et al 2007], and epigenetic modification [Bernstein et al 206, Lee et al 2006]. At present, only one report said the deletion of UCEs, postulated as enhancers, could yield viable mice [Ahituv et al 2007]. Even though the CFs discovered here cannot be considered as ultraconserved, their conservation among distant mammalian species is so high and long that it is perplexing if they happen by pure chance during the course of evolution.

What may be the possible roles of these CFs? It is well established that the presence of a highly conserved poly(A) signal at ~20 nts upstream and a U/GU-rich region at ~15 nts downstream from the PAS is sufficient to cause the polyadenylation machinery to cleave the nascent pre-mRNA from the transcription complex. Many of these CF are located less than 20 nts from the PAS (Figure 2.5) and they lack significant sequence similarity except for the three probably duplicated genes. These observations indicate that genes with CFs do not regulate by common factor. One supporting evidence is the

synergetic effect of the evolutionarily conserved U1 site and the PIE site in mammalian U1A gene.

Half of the CFs were found closer than 20 nts upstream of the PAS, suggesting that they may be correlated to polyadenylation activity, otherwise there is no reason to support their biased proximity to the PAS. However, even with such positional preference, one cannot exclude the possibility that these CFs are required by other biological processes, such as mRNA stability and, microRNA mediated translation regulation. Even though CFs longer than 100 nts are unusual, one should not overlook the rest of the 30-100 nts long CFs as known RNA protein recognition sites are short. In conclusion, the biological function of these CFs reported here is largely unknown. Novel gene specific regulatory mechanism may be attributed to their conservation. The pursuit described in this chapter may contribute in the discovery of intriguing experimental targets. Further validation is needed to confirm whether the disruption of these CFs could cause any negative impact on polyadenylation.

CHAPTER 3

PAS CLASSIFIER USING LOGISTIC REGRESSION

A. Introduction

There is a growing attention in the regulatory role of 3' UTR. Protein families such as Puf [Wickens et al 2002], Hu [Hinman et al 2008], ARE-BP [Chen et al 2001] regulate mRNA stability post-transcriptionally through 3' UTR binding. In addition, a multitude of both conserved and unconserved microRNAs target sites have been discovered recently in plants, insects, and mammals [reviewed in Bartel 2004, Griffiths-Jones et al 2008]. Many of them are attributed to cell proliferation [Sandberg et al 2008], development [Aravin et al 2003], translation regulation [Lim et al 2005], and differentiation [Chen et al 2004]. A substantial portion of human (54%) and mouse (34%) genes possess more than one PAS [Tian et al 2005], which leads to alternative 3' UTRs, or even ORFs for some genes. Studies have shown that alternative polyadenylation serves crucial biological functions such as T or B-cell differentiation [Takagaki et al 1998, Chuvpilo et al 1999] and embryonic development [Ji et al 2009]. Shortening of global 3' UTR was found to be widespread in activating oncogenes in cancer cells [Mayr et al 2009]. On the contrary, lengthening of 3' UTR was observed during embryonic development in mouse [Ji et al 2009]. Mutations located immediately downstream of PAS were found to increase polyadenylation efficiency in F2 [Gehring et al 2001, Danckwardt et al 2004] and FGG [Uitte et al 2007] genes, leading to venous thromboembolic events. The development of

emerging post-transcriptional gene silencing technologies triggers mRNA degradation through duplex formation in the 3' UTR [Brown et al 2008, Goraczniak et al 2009]. While the 5' end start of the 3'UTR is obvious to all, the 3' end is often unclear and evenly mistakenly mapped even in well-curated databases such as NCBI RefSeq. According to my data, I have found that the 3' end of only 27% of human and 17% of mouse cDNA entries reviewed in the NCBI RefSeq database are supported by polyadenylated ESTs. Therefore, a better method is needed to accurately predict the 3' end of the transcripts so that the whole 3' UTR can be studied for its essential regulatory role and therapeutic application.

Beside the proposed close species comparison method mentioned in the previous chapter, this chapter will discuss the construction of a polyadenylation site classifier (PAS classifier) using a supervised machine learning method named logistic regression. Such a PAS classifier can complement commonly used expression-data-based methods such as ESTs and next generation sequencing to mark the 3' end. Moreover, constructing a classifier involves the identification of distinctive features of active PAS that will enrich current understanding of their intrinsic properties. Inevitably, some active PAS will be found that do not share the typical characteristics possessed by the majority. Such outliers are valuable in expanding our existing model of polyadenylation that may lead to the discovery of compensatory factors related to polyadenylation. Furthermore, mutations flanking the PAS have been known to have health implications due to the alteration of polyadenylation activity. It will be

interesting to use the PAS classifier as a tool to score the impact of such mutations.

Previous work has been done to predict PAS. Examples of three PAS classifiers are Polyadq [Tabaska et al 1999], ERPIN [Legendre et al 2003], and polya_svm [Cheng et al 2006]. Polyadq used two weight matrices to capture position scores, one for poly(A) signals upstream, the other for the downstream U/GU-rich region. To determine the threshold for a real PAS, it used a set of real and false 150-nucleotide (nt) long PAS sequences to train two quadratic discriminant functions (QDF). Instead of using two weight matrices, ERPIN used only one weight matrix to cover 300 nts upstream and downstream of the PAS, hereinafter denoted by [-300,+300]. The values of weight matrix are the log-odd ratio of real to false PAS. The most recent example is polya_svm. By examining [-100,+100] region, the authors identified 15 distinguishing cis elements of a PAS and used these to construct 15 position-specific matrices. In this method, each sequence yielded a feature vector consisting of 15 values. A set of feature vectors converted from real and false PAS sequences were used to train the support vector machine in order to determine the boundary support vectors.

This chapter will describe a logistic regression based PAS classifier. One advantage of logistic regression is that the returned model is more interpretable than other methods because the relative contribution of each feature can be measured, leading to a better understanding of their biological importance for a PAS. The analysis below is broken into sections as follows: 1) describe the method used to collect good quality real PAS sequences, 2) present the training

procedure, 3) assess the prediction performance, 4) compare with the two existing PAS classifiers, ERPIN and polya_svm, mentioned above, and 5) application to analyze already collected PAS data.

B. Classifier Construction

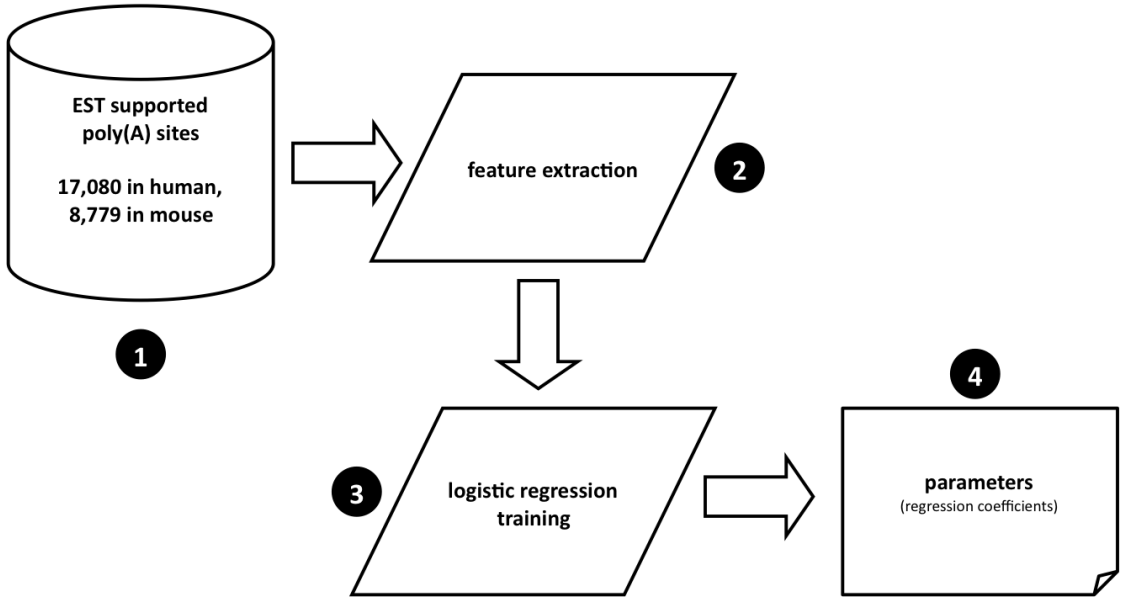


Figure 3.1 Workflow of logistic regression PAS classifier construction.

As illustrated in Figure 3.1, the whole procedure of PAS classifier construction consists of four major steps:

Step 1. Polyadenylated EST sequences were used to locate PAS in genomes, where each PAS had to be supported by at least three ESTs. In order to avoid a "garbage-in-garbage-out situation", different measures were used to avoid false priming and erroneous directionality. A detailed procedure to identify EST-supported PAS can be found in Appendix B. As human has almost doubled

amount of ESTs than mouse (the next closest species) i.e. 8 millions in human versus 4 millions in mouse, therefore more EST-supported PAS were found in human than in mouse.

Step 2. Analysis of PAS collected from step 1 identified 10 distinguishing features of a real PAS (will be discussed later). Based on these features, each sequence was encoded as a vector of 10 numeric values, named feature vector. The training step also included the learning of unreal PAS, which were sourced from intronic sequences with the canonical poly(A) signal (AWTAAA), intergenic regions, ORFs, and simulated sequences. Both positive (real PAS) and negative (unreal PAS) feature vectors were then passed to the next step.

Step 3. Based on the positive and negative feature vectors, the logistic regression function searched for a set of coefficients such that the overall misclassification was minimum. This step also calculated several performance coefficients by using the classifier to make prediction for unseen samples.

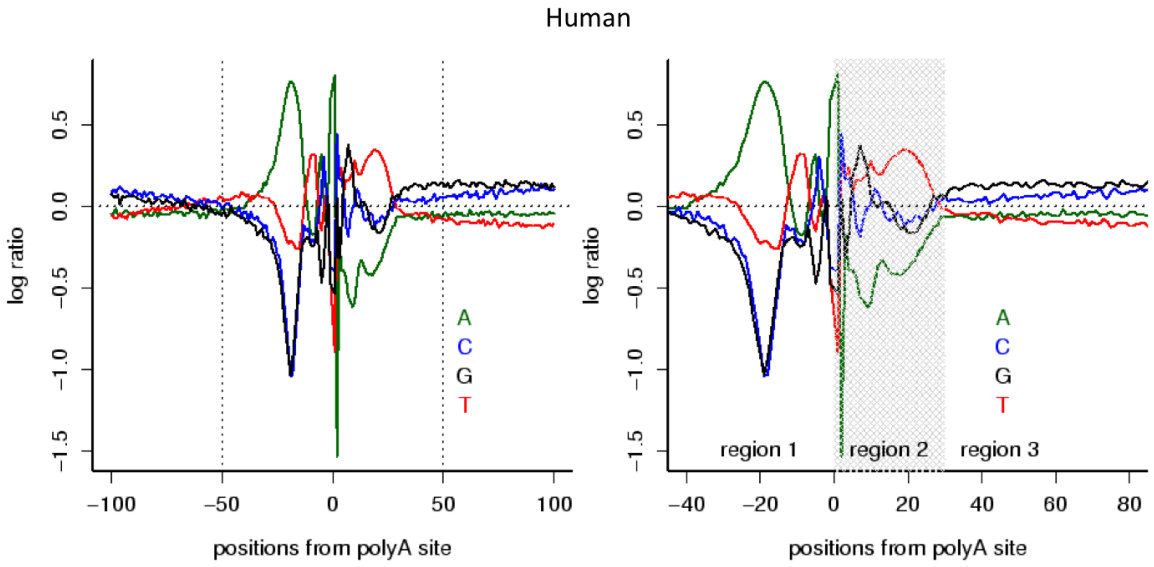
Step 4. The set of coefficients from the best model were then used to make predictions.

C. Features Selection

Based upon the procedure briefly discussed above, 17,080 human and 8,799 mouse PAS were compiled that became the primary data set to identify features of active PAS. Two broad aspects of these PAS were analyzed viz. nucleotide profile and enrichment of *k*mers (*k*-sized oligomers).

1. Nucleotide profile

Not only do 3' UTRs contain elevated levels of A and T nucleotides as discussed in the previous chapter, they also exhibit signatory nucleotide distribution surrounding the PAS. By using the human and mouse PAS sequences, the overall nucleotide profile across region [-100,+100] was plotted in Figure 3.2 below.



A B

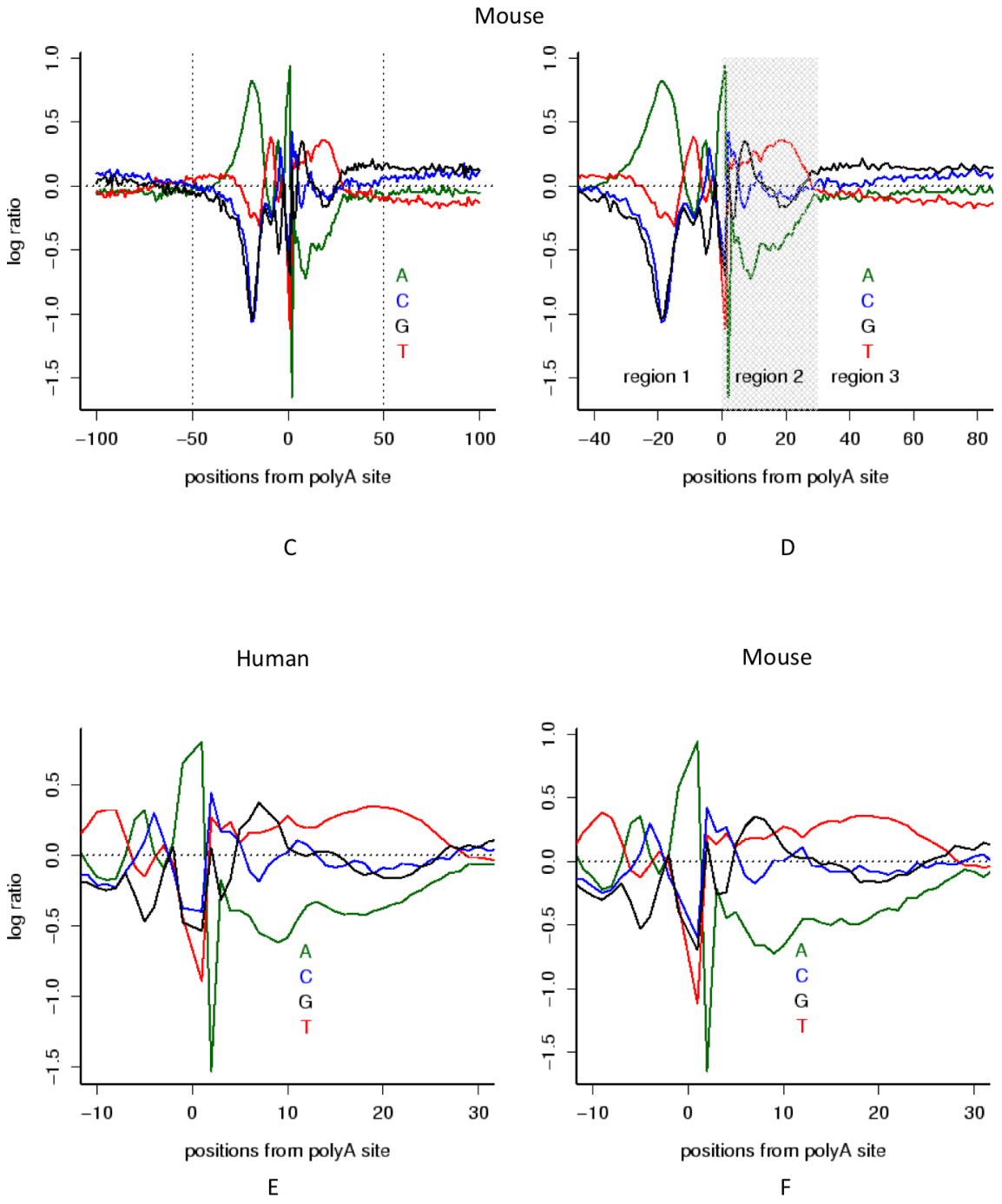


Figure 3.2 Nucleotide profiles of the PAS flanking region. A) Human region [-100,+100], B) zoomed into region [-40,+80], C) Mouse region [-100,+100], D) zoomed into region [-40,+ 80], E-F) zoomed into region [-10,+30] in human and mouse, respectively.

Let $N_i \in \{A,C,G,T\}$, and $P_j \in \{-100,-99,\ldots,-1,1,2,\ldots,100\}$. The y-axis in Figure 3.2 is the log ratio between the actual number of N_i observed and average N_i per position, i.e. log(observed/average), where the average is the occurrence of N_i in all PAS sequences divided by the total number of positions. Using average N_i per position is better than assuming equal proportion of all four nucleotides because it avoids the mistake of taking simply A and/or T rich sequence as PAS. Thus this method is designed to reward A/T at appropriate positions only. It is observed that each nucleotide exhibits a distinctive profile:

1. Nucleotide profiles highly resemble each other between human and mouse, consistent with the fact that polyadenylation factors are largely conserved in mammals.
2. The log ratio of all nucleotides stays flat at level zero in region [-50,0] indicating the classifier should focus on nucleotide distribution in the region [-50,+100].
3. Adenine exhibits the most dramatic localization pattern among all nucleotides along [-50,+100] as it is highly enriched in the region [-40,-10], which reflects the localization of poly(A) signals. There is an adenine spike near the cleavage site [-5,+1] followed downstream by a depletion of adenine in the region [+1,+30].
4. Cytosine has higher presence in the downstream region [+25,+80] but mainly after +30.

5. Guanine concentrates in two disjoint downstream regions [0,+10], and [+30,+80].
6. Thymine has two spikes: one after adenine in the upstream region [-20,-10] and a second broader spike at around [+1,+25] after the peaks of C and G in the downstream region, an observation already reported in [Salisbury et al 2006].

Several trials were done to minimize the size of the flanking region without sacrificing prediction accuracy and this demonstrated region [-40,+80] was sufficient to yield good prediction. A position weight matrix captured the nucleotide profile of this region. As polyadenylation is required for all genes except histones one may expect the PAS exhibits a high degree of variation such that some genes conform only partly to the above characteristics, for instance, high A-rich upstream but poor T-rich downstream. In this case a few genes with unfavorable features could cancel out significant findings in the main population of genes. Hence Table 3.1 below summarizes eight separate scores that can reflect the distinct patterns exhibited by different nucleotides at different regions.

Features	Remarks
A1	Sum of log ratio of nucleotide A in [-40,-10]
A2	Sum of log ratio of nucleotide A in [-5,-1]
A3	Sum of log ratio of nucleotide A in [+1,+30]
C1	Sum of log ratio of nucleotide C in [+25,+80]

G1	Sum of log ratio of nucleotide G in [+1,+10]
G2	Sum of log ratio of nucleotide G in [-+25,+80]
T1	Sum of log ratio of nucleotide T in [-15,-5]
T2	Sum of log ratio of nucleotide T in [+1,+30]

Table 3.1 Features from nucleotide profile analysis.

2. Enriched *k*mers

Aside from nucleotide profiling, a second broad aspect was to identify the enrichment of certain *k*mers by location. A dimensionality reduction method called singular value decomposition (SVD) was used to reveal *k*mer enrichment at specific positions relative to the PAS as well as *k*mer enrichment in the overall PAS region.

Given a set of sequences, each *l* nts long, using a sliding window of size *k*, each sequence was broken into (*l*-*k*+1) overlapping *k*mers. Positions of *k*mers were recorded in a position-by-*k*mer matrix M with m rows, n columns where m=*l*-*k*+1 and n=4^k. For convenience, it is assumed m<n, i.e. the length of sequence is less than the total possible *k*mers. For instance, given forty 100-nt long sequences, and the size of *k*mer is set to 4, then M is a 97-by-256 matrix. Applying SVD to M will factorize it into the product of three matrices represented by the equation below:

$$M = U\Sigma V^t$$

where *U* and *V* are orthogonal matrices with dimension m-by-m and n-by-n, respectively. Σ is an m-by-n diagonal matrix with the m eigenvalues ($\sigma 1, \sigma 2, \sigma 3, \ldots, \sigma_m$) of M, where $\sigma 1 \geq \sigma 2 \geq \sigma 3 \geq \ldots \geq \sigma_m$, along the diagonal.

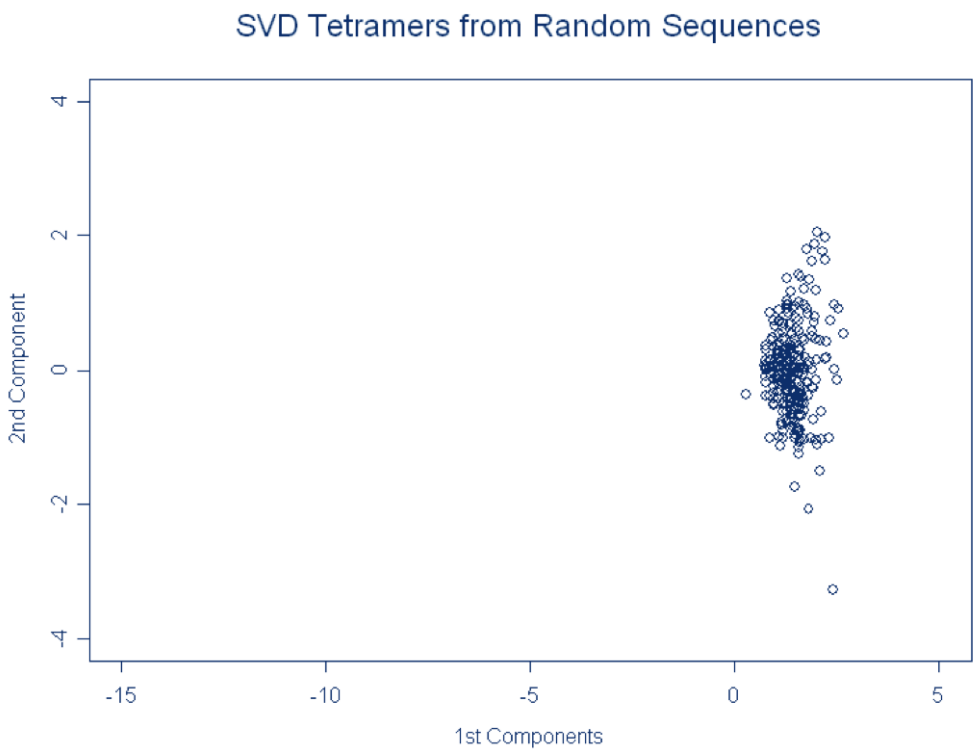
The physical interpretation of SVD is that matrix M captures the distribution of *k*mers, in terms of occurrences and positions, in the sequences. Each row represents the occurrences of different *k*mers at a position. To probe for the localization of *k*mers, the distribution of the m rows (one for each position) in an n-dimensional hyperspace were examined. If some *k*mers do prefer to stay at particular position(s) in the sequences, the overall distribution of these n-dimensional position vectors should be asymmetrical; otherwise its distribution is hyperspherical. In other circumstances, if *k*mers (which correspond to certain binding sites) are flexible in terms of location then they are enriched but without a constraint in position. In similar manner, it is possible to examine the distribution of n columns in the m-dimensional hyperspace where each column represents the abundance and/or localization properties of a *k*mer. But it is hard to measure asymmetry for high dimensional data. By SVD, the original high dimensional matrix can be approximated in the least mean squared error by a two dimensional matrix such that visualization becomes possible.

After factorizing M into three matrices, the distribution of positions can be projected onto a 2-D plane by selecting the first two columns of an m-by-n matrix $U\Sigma$, i.e.

$$M_2 = U_2 \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}$$

where M_2 and U_2 are m-by-2 matrices, $\sigma 1$ and $\sigma 2$ are the two largest eigenvalues of M, and $\sigma 1 > \sigma 2$. Mathematically, M2 is the best m-by-2 approximation of the m-by-n matrix M. The reader can refer to [Meyers 2001] for a more detailed mathematical discussion of SVD.

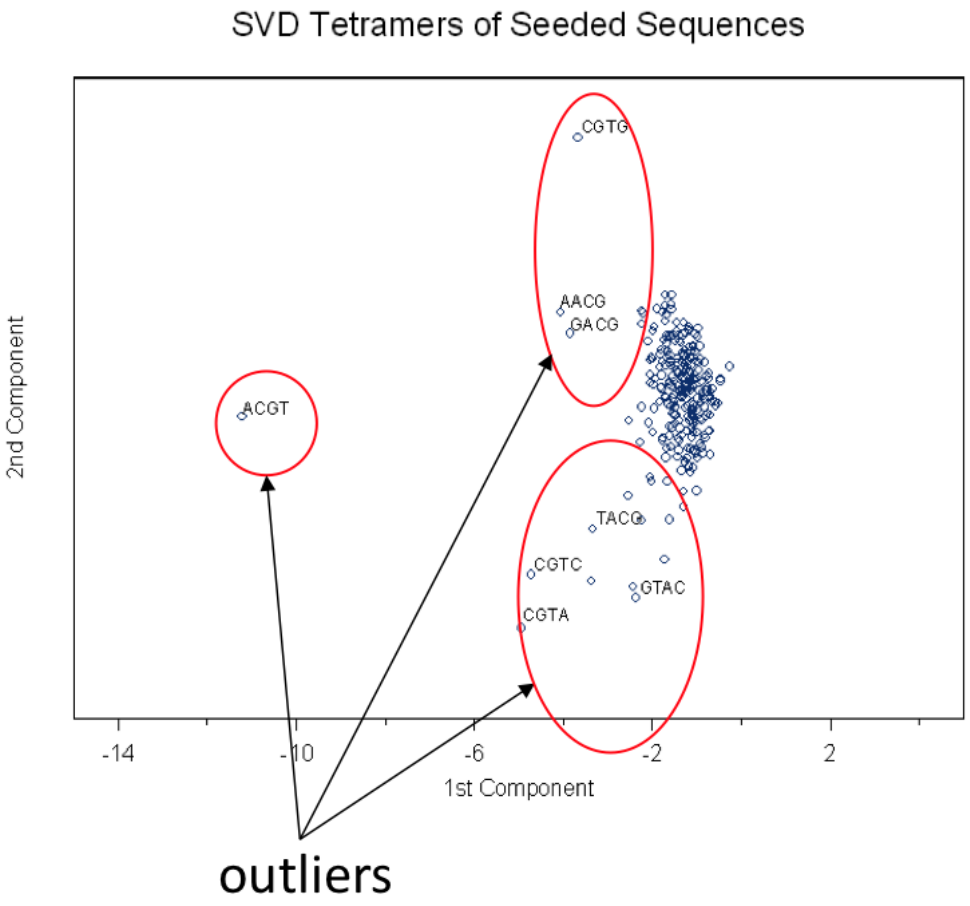
The distribution of *k*mers can be assessed by the selection of the first two columns of $V\Sigma$. To illustrate the idea, one hundred 40-nt long sequences were generated with A, C, G and T in equal abundance. These sequences were encoded into a position-by-*k*mer matrix M as discussed. M was factorized by SVD and the *k*mer distribution, i.e. $V\Sigma$, was projected onto a 2-D plane as shown in Figure 3.3A below. Each dot in the diagram represents a *k*mer, and there are 256 (4^4) of them. As shown, all *k*mers are clustered together, meaning that no particular *k*mer is enriched in the sequence set.


SVD Tetramers from Random Sequences
2nd Component
4
2
0
-2
-4
-15
-10
-5
0
5
1st Components

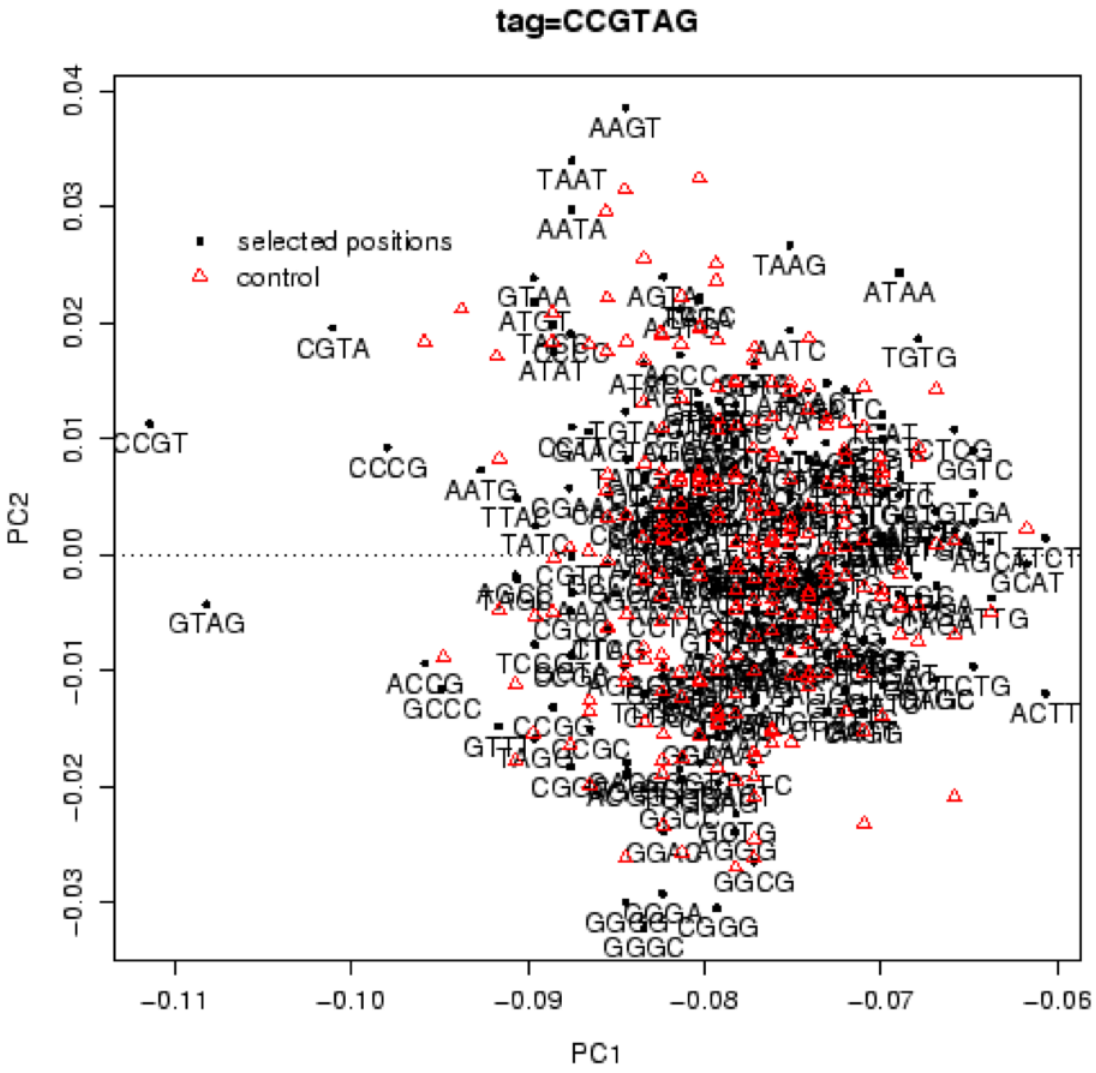
A

AGGCTGATCAGGCCAGAATATGCCGACGTTTGTCTAGGGCGCAGCATCTA
CGCTTGGAAGCAACGAACGTGGAGATACGCAGGGGGTTCCTAGTATATGG
ACGTACACTTAGAAACCAAGCCAGATGCTGCAGCCCTGCCATGCAGTATG
AATGTTGCTTAGTAGTAACGCTAGCACCGTACACGTAGCCACCTTATCAC
TTGATAATTCTGATACACGTTGCGGAGTGCCTTGTACCCACACCTTTCGC

B



C


tag=CCGTAG
selected positions
control
AAGT
TAAT
AATA
TAAG
ATAA
GTAA
ATGT
CGTA
AATC
TGTG
ATAT
CCGT
CCCG
AATG
TTAT
TATC
GGTC
GTGA
GCAT
GTAG
ACCG
GCCC
ACTT
GGCG
PC1
PC2
−0.11
−0.10
−0.09
−0.08
−0.07
−0.06
0.04
0.03
0.02
0.01
0.00
−0.01
−0.02
−0.03

D

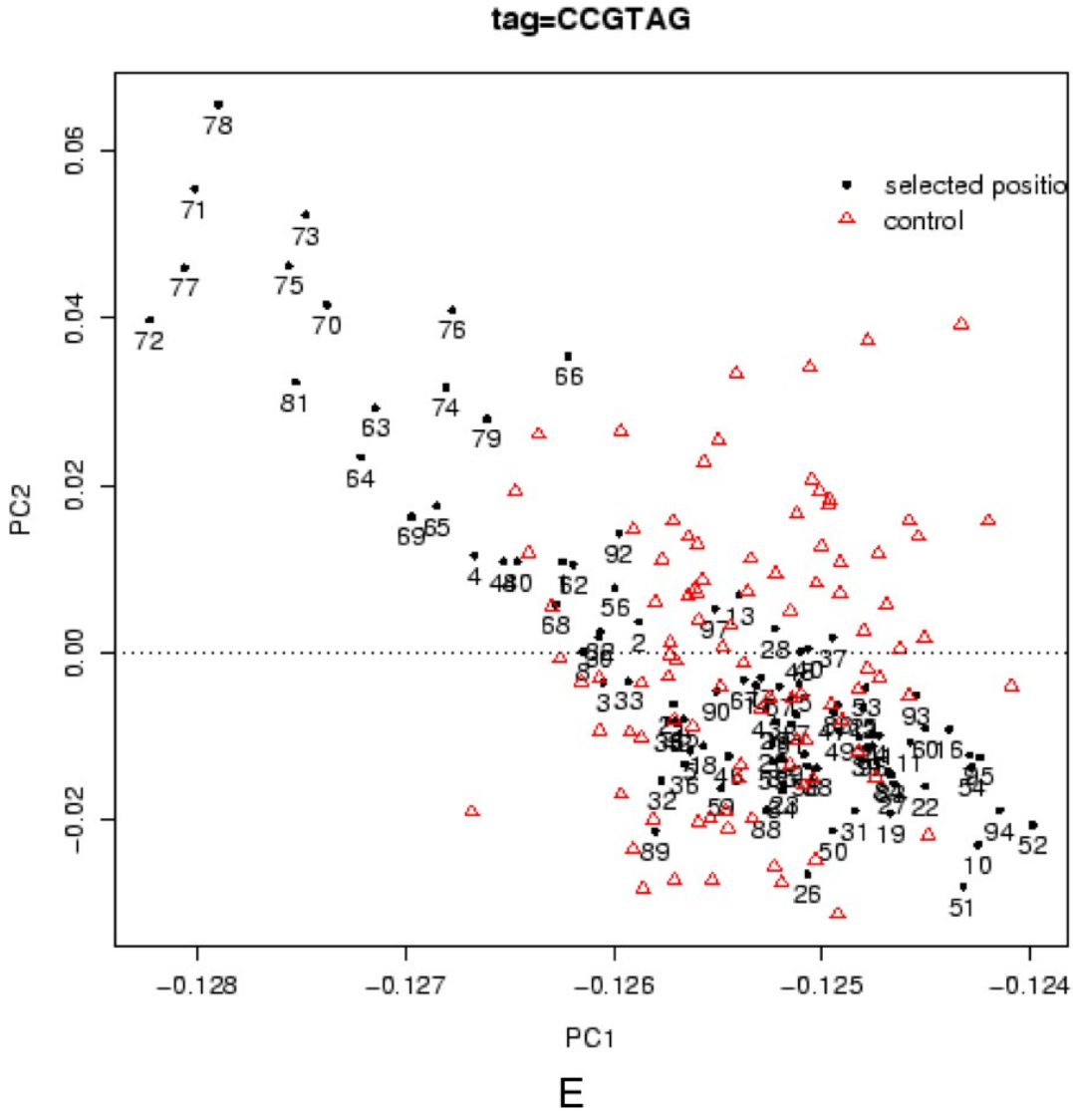
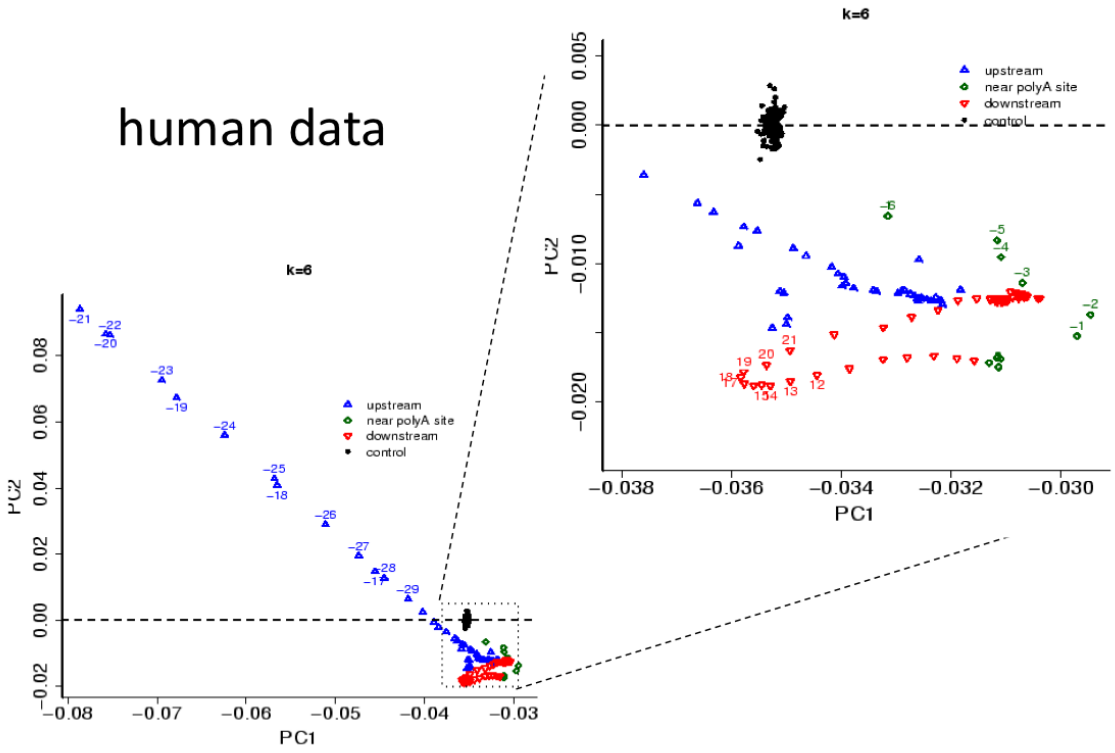


Figure 3.3 SVD of simulated sequences. A) simulated sequences with no motif planting, B) simulated sequences planted with ACGT (marked in red) at random locations, C) SVD analysis of simulated sequences in B, D) *k*mer SVD projection of simulated sequences planted with CCGTAG with one mutation, E) position SVD projection of the same set of sequences from D.

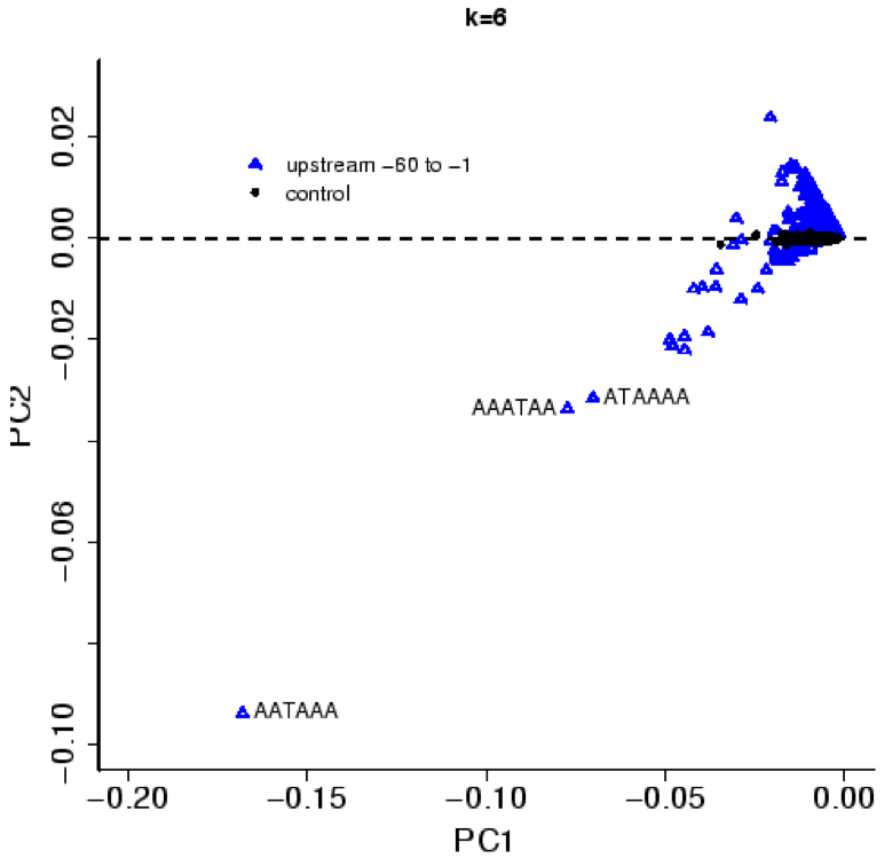
In the next test, a short fragment ACGT was planted at random locations in the sequences (Figure 3.3B), factorized M by SVD, and projected *k*mer distribution in a 2-D plane. In Figure 3.3C, it is clear that ACGT is the farthest *k*mer from the rest, also for its overlapping neighbors such as CGTA, CGTC, etc. To illustrate how SVD can help to discover the localization of a slightly varied sequence elements, a short fragment CCGTAG carrying one mutation at any position was planted in the region spanning 60 to 80 in each 100-nt long simulated sequence. The *k*mer and position projections are depicted in Figure 3.3D and E, where it is evident that a handful of tetramers are located away from the other tetramers such as CCGT, CGTA, and GTAG, although the signature pattern is not as conspicuous as in Figure 3.3C. Examination of the position SVD projection indicates positions from 63 to 79 are located far from the rest. The combination of *k*mer and position SVD projections results in an accurate localization and identification of a possible longer mutated sequence motif, a result that cannot be obtained by simple counting of over or under-represented *k*mers in the presence of mutations.

Below are results of application of SVD to the PAS sequences in order to detect enriched *k*mers and their localization information. Figure 3.4A is a position SVD projection for the region 100 nts upstream (in red), ±5 nts around cleavage site (in green), and downstream (in blue) of PAS where k=6. Regarding the choice of k, i.e. the size of *k*mer, a large k will create a sparse matrix, which cannot be factorized by SVD. On the other hand, a small k will make it hard to differentiate signaling motif from background sequence. By trying a range of k

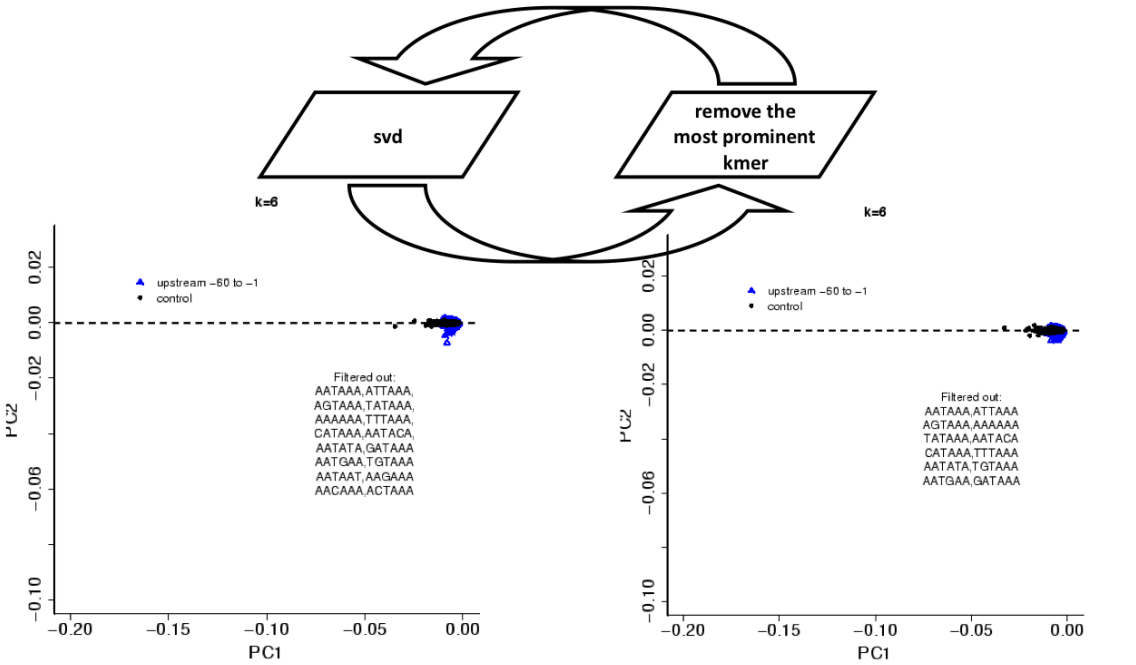
from 4 to 8, *k*=6 produced the clearest picture for position and *k*mer analyses along both upstream and downstream PAS sequences. However, it will be discussed later that a smaller *k* may produce a better result when the analysis is narrowed down to the downstream region only.



A


k=6
upstream −60 to −1
control
AAATAA
ATAAAA
AATAAA
PC2
PC1
0.02
0.00
−0.02
−0.06
−0.10
−0.20
−0.15
−0.10
−0.05
0.00

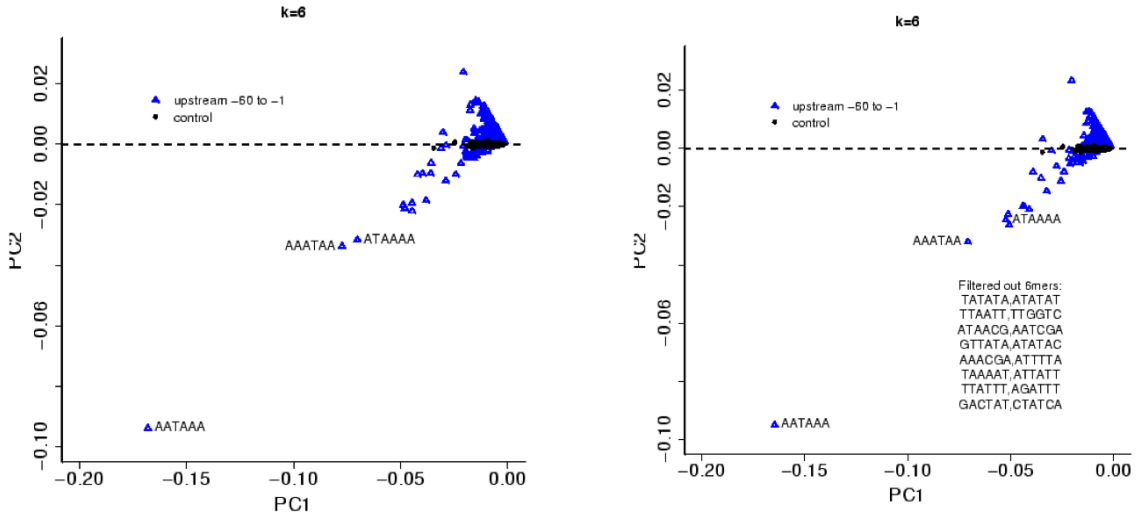
B



filtered out the 16 most remote 6mers in human

filtered out the 12 most remote 6mers in mouse

C



no filtering

filtered 16 irrelevant 6mers

D

Figure 3.4 Feature extraction by SVD projection. A) position SVD projection of human 100 nts upstream and downstream from PAS, B) *k*mer SVD projection of human upstream region, C) iteration of *k*mer SVD for human and mouse upstream regions, D) removal of irrelevant hexamers does not cause the positive data cloud (blue) to shrink.

In Figure 3.4A, the analysis exhibits strong position bias versus control denoted by the black dots. The upstream region [-11,-28] indicates unusual localization of certain hexamers and positions [+12,+21] downstream of PAS are also located away from the majority though to a lesser extent than upstream.

To identify hexamers enriched in the upstream of these PAS sequences, *k*mer SVD analysis was applied to the upstream region alone (Figure 3.4B). As expected, the most common canonical poly(A) signal AATAAA is the obvious outlier, however whether other outlying hexamers such as AAATAA and ATAAAA are autonomous or simply part of the canonical poly(A) signal is difficult to judge. To examine this, an iterative *k*mer SVD analysis procedure was used as follows. After each round of *k*mer SVD analysis, the most pronounced outlier *k*mer from the input sequences was removed and then the *k*mer SVD step was repeated. Iteration of this process was done until the positive data shrunk into the control data. Application of the iteration process as illustrated in Figure 3.4C (left) resulted in identification of 16 hexamers that are enriched in the upstream region. Analysis of mouse PAS sequences identified 16 hexamers of which 12 matched the 16 human hexamers (Figure 3.4C right). In order to eliminate the

possibility that shrinking is due to the removal *k*mers regardless of whether they are enriched, the iterative *k*mer SVD analysis was repeated using 16 randomly selected hexamers not in the 16 enriched hexamers. As shown in Figure 3.4D, removal of such random *k*mers from sequences failed to cause any shrinkage.

Previous work has identified overrepresented *k*mers in the upstream region [Beaudoing et al 2000,Tian et al 2005] permitting a comparison of the finding herein with the findings with the 13 hexamers reported by the Tian lab in Table 3.2.

	SVD rank	Tian rank [Tian et al 2005]	Beaudoing rank [Beaudoing et al 2000]
AATAAA	1	1	1
ATTAAA	2	2	2
AGTAAA	3	4	4
TATAAA	4	3	3
AAAAAA	5	-	-
TTTAAA	6	11	9
CATAAA	7	8	6
AATACA	8	7	8
AATATA	9	6	5
GATAAA	10	9	7
AATGAA	11	10	11
TGTAAA	12	-	-
AATAAT	13	-	-
AAGAAA	14	5	10
AACAAA	15	-	-
ACTAAA	16	12	13
AATAGA	-	13	12

Table 3.2 Comparison of upstream hexamers discovered by kmer SVD analysis and two existing reports.

The two sets share 12 common hexamers and our set does not have the least common hexamer, AATAGA from Tian and Beaudoing labs. The four new hexamers are AAAAAA, TGTAAA, AATAAT, and AACAAA.

In contrast, the analysis of the downstream region showed hexamer was not a good size to produce a clear visualization result, presumably because of the more degenerate nature of the downstream region. Such argument is supported by Figure 3.4A that the downstream position bias (red) exhibits a smaller scale when compared to the upstream region (blue). By using the same procedure but with *k* being set to 4, tetramers TTTT, GTGT, TGTG and their one point mutant variations were found to be enriched in the downstream region.

Based on these hexamers and tetramers, a position weight matrix of log ratio values was done that was similar to the one used for nucleotide profile with the only difference being that hexamers and tetramers were used instead of single nucleotides. Two feature values one for upstream (pscore1), and the other for downstream (pscore3) were calculated. (pscore 2 is a feature value for the region [-5,+5], as the CA dinucleotide has been reported to be enriched at the cleavage site. However, during the training step, pscore2 did not contribute much to prediction and so it was dropped from the final logistic model.) Thus, the final feature vector contains 10 values as shown below:

feature vector = (A1, A2, A3, C1, G1, G2, T1, T2, pscore1, pscore3)

D. Logistic Regression

To determine the likelihood of a sequence to be a PAS, I chose to assume this was a binary classification problem using logistic regression as the underlying model. The logistic regression function is defined to be:

$$p(y) = \frac{1}{1 + e^{-y}}$$

$y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \beta_3 * x_3 + \ldots \beta_{10} * x_{10}$, β_i i=0 to 10, are the regression coefficients to be determined by the training procedure. x_1, x_2,… ,x_{10} are feature values from the feature vector. The value of p(y) is in the range of 0 to 1 and was called logistic score.

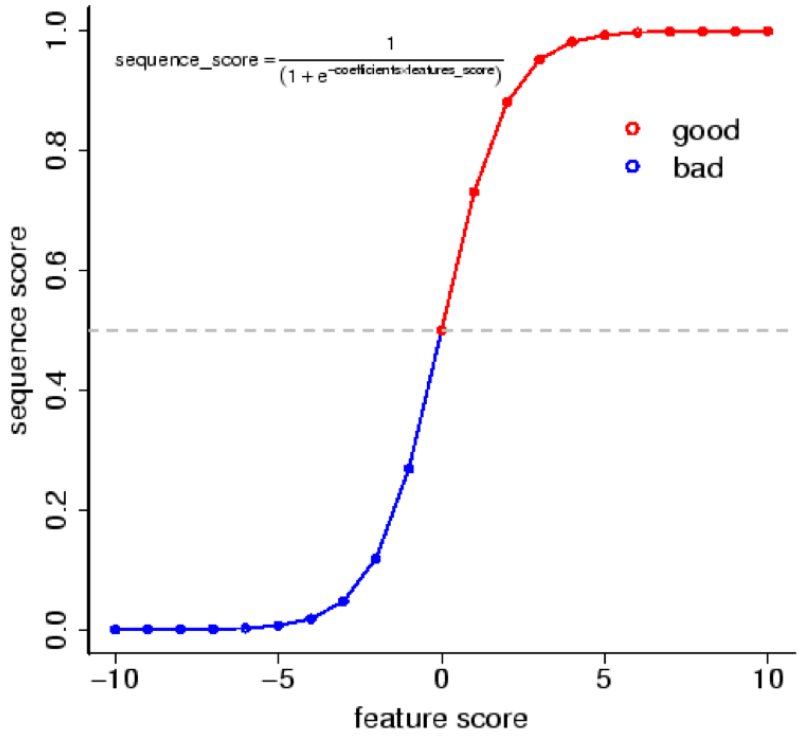


Figure 3.5 Logistic function.

The goal of the logistic regression is to determine the set of regression coefficients, i.e. β_o, β_1, β_2,..., such that *p* (logistic score) for a positive (real) sample should be as close to 1 as possible. Conversely, *p* for a negative (false) sample should be close to 0 as possible. To determine the set of coefficients, the classifier was calculated from the features of both positive and negative samples, this step was called training. Below is a description of the training procedure for the PAS classifier using human and mouse data.

1. Training datasets

Based on the EST-supported PAS, and the 10 features identified previously, the PAS sequences were encoded into feature vectors with label 1, thereby forming the positive dataset. For the negative dataset, the 1st order Markov probability was captured from real PAS sequences using empirical probabilities to generate the negative dataset with label 0.

2. Training procedure

At each round, 2,000 feature vectors were sampled from positive and negative datasets, they were passed to the generalized linear model function `glm()` provided by R in order to determine the set of regression coefficients. The same step was repeated 100 times to obtain the averaged coefficients.

3. Model validation.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.21892    0.61689 -14.944  < 2e-16 ***
pscore1      1.95052    0.09397  20.757  < 2e-16 ***
pscore3      0.18877    0.03777   4.998 5.80e-07 ***
A1           0.32878    0.06666   4.932 8.13e-07 ***
A2           2.25869    0.13761  16.414  < 2e-16 ***
A3           0.46863    0.06895   6.796 1.07e-11 ***
C1           1.05626    0.17778   5.941 2.83e-09 ***
G1           0.81248    0.23778   3.417 0.000633 ***
G2           0.74614    0.10722   6.959 3.43e-12 ***
T1           1.37205    0.21457   6.394 1.61e-10 ***
T2           0.86012    0.13423   6.408 1.47e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A

```
Correlation of Coefficients:
        (Intercept) pscore1 pscore3 A1    A2    A3    C1    G1    G2    T1
pscore1 -0.10
pscore3 -0.14        0.09
A1      -0.34       -0.51    0.00
A2      -0.35        0.22    0.06    0.07
A3       0.45        0.07   -0.46    0.05  0.11
C1      -0.50        0.09   -0.02    0.08  0.11 -0.10
G1      -0.25        0.02   -0.28    0.02  0.07 -0.02 -0.01
G2      -0.51        0.11   -0.08    0.07  0.12 -0.08  0.17  0.00
T1      -0.22        0.07    0.04   -0.01  0.11  0.06  0.07 -0.01  0.03
T2      -0.55        0.10   -0.46    0.03  0.08 -0.06  0.16  0.33  0.21 -0.01
```

B

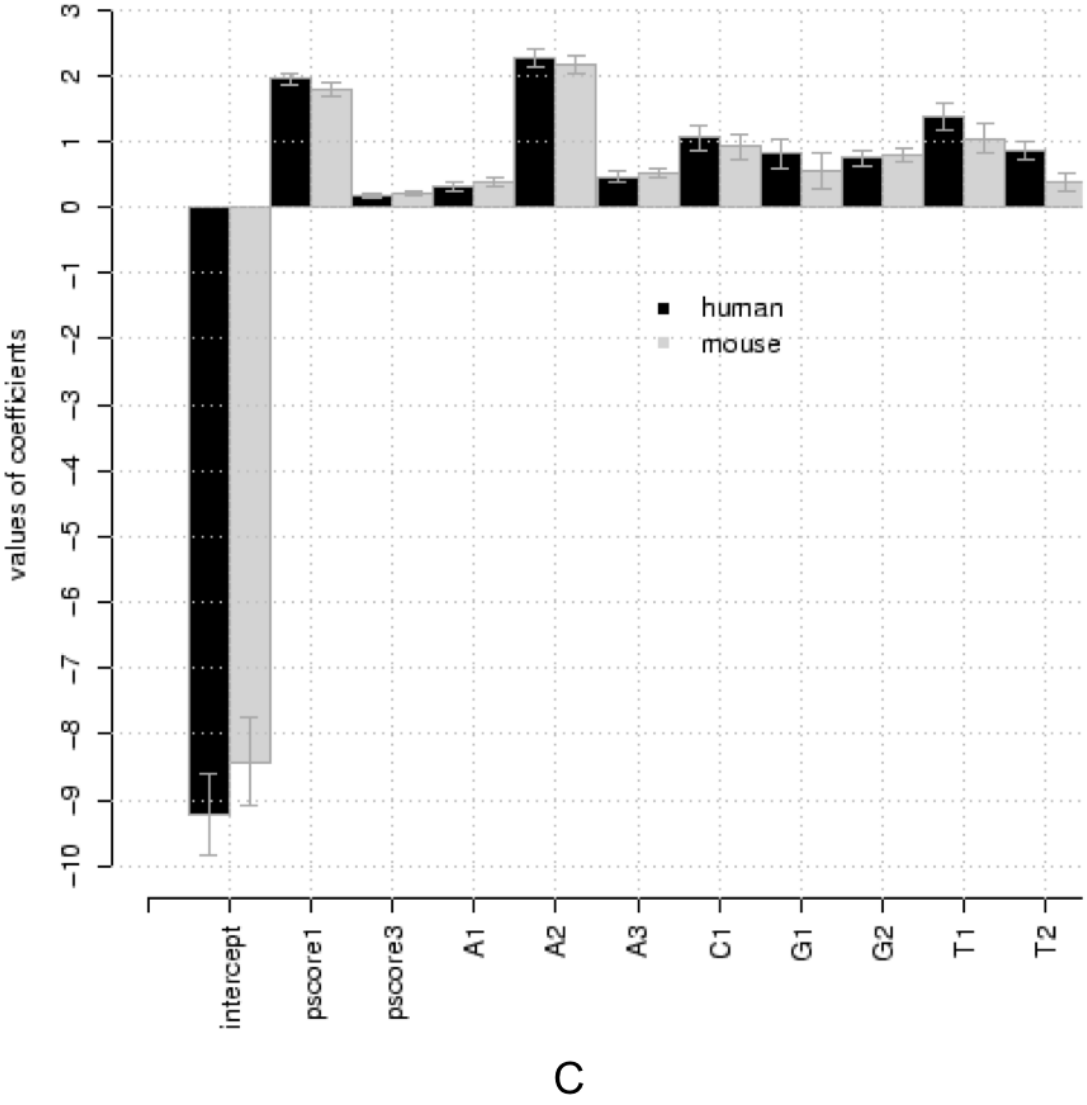


Figure 3.6 Regression model validation. A) the best coefficients and their significances are reflected in the p-values, B) correlation among the coefficients, C) the comparison of coefficients between human and mouse models.

Here it was determined whether all features suggested in the previous section are relevant in recognizing the PAS. The significance of each feature is reflected in the p-value column in Figure 3.6A. The extremely small p-values indicate that all 11 coefficients (including the intercept) are significant.

Next, it was necessary to determine whether these coefficients are redundant, that means do any of them positively or negatively correlate to each other, an issue called multi-colinearity. As shown in Figure 3.6B, their correlations were within the range (-0.51, 0.45), ruling out any significant multi-colinearity issue. Due to the fact that the polyadenylation machinery in human and mouse are highly conserved, PAS regions of human and mouse should manifest a high degree of similarity too, implying that the coefficients between human and mouse models should be similar. Such a view is confirmed in Figure 3.6C, they largely agree with each other.

4. Threshold

The logistic score returned by the logistic regression model is continuous between 0 and 1, meaning a threshold is needed to distinguish real from false PAS. Setting the threshold too high will increase the false negative rate. Conversely, setting it too low will increase the false positive rate.

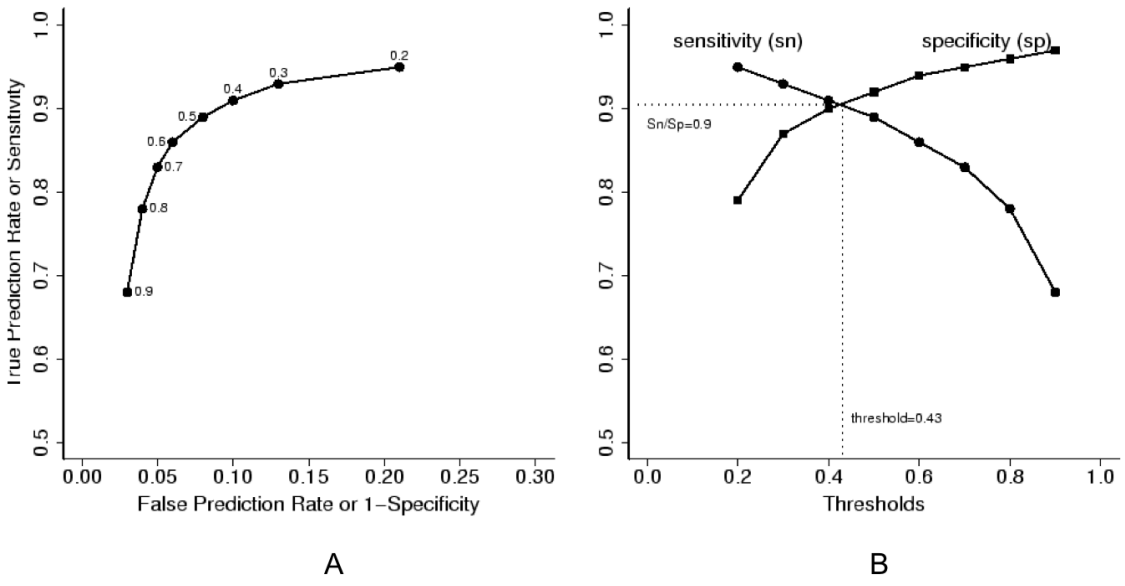


Figure 3.7 ROC of PAS classifier. A) true prediction rate versus false prediction rate for various thresholds, B) sensitivity and specificity versus threshold.

In Figure 3.7, the true and false prediction rates for various thresholds was measured and it was found that a threshold set to 0.5 was the most optimal choice. For the rest of the discussion, the default threshold is 0.5 unless stated otherwise.

5. Relative importance of features

Understanding the degree of contribution of the 10 features can help to infer their relative biological importance in defining the PAS. This was assessed by the utilization of the R function `add1()`, which computes the prediction rate by adding only one feature to the null model each time. The percentage reduction in deviance stated in Table 3.3 denotes the decrease of prediction error versus random guessing by the null model. Like the training procedure discussed before, 2,000 positive and negative sequences were randomly selected to conduct this study. The averaged results over 100 trials are tabulated below.

Feaure(s) added to null model	Human % reduction in deviance	Mouse % reduction in deviance
All 10 features	72	75
pscore1	51	55
A1	33	37
A3	18	21
pscore3	16	19
A2	15	17
T2	11	12
G1	3	3
G2	3	5
T1	3	4
C1	2	1

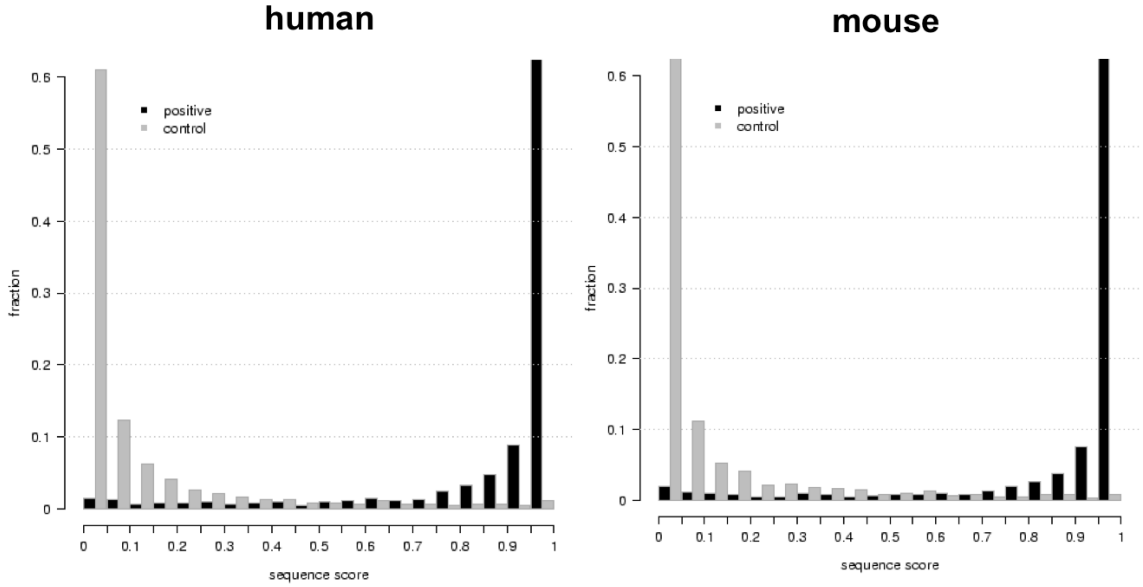
Table 3.3 Relative importance of features in human and mouse models.

As expected, the 16 hexamers (pscore1) and the A-rich region upstream (A1) play a major role in defining the PAS, nevertheless, the contribution from the other 8 features improves the PAS recognition by capturing PAS with slight variations from the norm. In addition, the finding of the importance of upstream elements is consistent with the conservation analysis discussed in the previous chapter, confirming the presence of conservation pressure to maintain definitive cis elements of polyadenylation in the upstream region instead of downstream. Furthermore, this analysis has demonstrated the advantage of logistic regression in producing interpretable parameters over other methods such as support vector machine used by polya_svm and position weight matrices used by ERPIN.

E. Results

1. Prediction performance

The PAS classifier described above was used to make prediction for all human and mouse PAS sequences, and sequences from the negative dataset. Results were compared with two other PAS classifiers, ERPIN and polya_svm.



	TP (%)	TN (%)	Sp	Sn	FDR	PC	CC	F	PPV	NPV
erpin	79.6	77.3	0.78	0.80	0.22	0.65	0.57	0.79	0.78	0.79
SVM	90.1	79.7	0.82	**0.90**	0.19	0.75	0.70	0.86	0.82	0.89
logistic	**90.9**	**92.5**	**0.92**	**0.90**	**0.08**	**0.84**	**0.83**	**0.92**	**0.92**	**0.91**

$$Sp = \frac{TP}{(TP + FP)} \qquad CC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

$$Sn = \frac{TP}{(TP + FN)} \qquad PC = \frac{TP}{(TP + FP + FN)} \qquad PPV = \frac{TP}{(TP + FP)}$$

$$FDR = \frac{FP}{(TP + FP)} \qquad F = \frac{2Sn * Sp}{(Sn + Sp)} \qquad NPV = \frac{TN}{(TN + FN)}$$

logistic

C

Figure 3.8 Predictions of human and mouse PAS sequences. A) score distribution for human, B) score distribution for mouse, C) performance parameters comparison.

As shown in Figure 3.8A-B, the classifier is able to differentiate real from false PAS sequences in human and mouse. It can achieve 92% accuracy in predicting positive data (PPV), and 91% accuracy in rejecting negative data (NPV). Formulae for PPV and NPV are provided in Figure 3.8C above. The logistic PAS classifier showed improvement in terms of PPV and NPV when compared with the other two methods as shown in the last two columns in Figure 3.8C.

Aside from this, the sensitivity and specificity of the prediction were also determined. Sensitivity calculates the proportion of real PAS detected, whereas specificity measures the proportion of predicted PAS that are real PAS indeed. These two factors always counterbalance each other, as revealed by the ROC in Figure 3.7B. Thus it is difficult to decide whether a model with higher sensitivity is better than another model with higher specificity, and vice versa. This issue was addressed by combining sensitivity and specificity into one value. Three commonly used calculations were included in this study viz. performance coefficient (PC), correlation coefficient (CC), and F-measure (F). They all share one common property that is to penalize skewed sensitivity or specificity. Their formulae are stated in Figure 3.8C. All three performance measures appeared in Figure 3.8C were the average of 100 trials where, in each trial, 2,000 real and false PAS sequences were sampled randomly, followed by prediction. Values of PC, CC and F showed that the classifier exhibited improvement in prediction as compared to other methods.

2. Prediction for other genomic sequences

It was also important to determine how well the classifier is able to distinguish PAS from other naturally occurring genomic regions such as the ORFs, 5'UTRs and intergenic regions. By using the same procedure from the previous section, the logistic method prediction was compared with the predictions from the other two methods and the results were tabulated below in Table 3.4.

Specificity	ORF	5'UTR	Intergenic (human chr1)
EPRIN	0.96	0.95	0.79
SVM	0.86	0.72	0.82
logistic	0.97	0.97	0.94

Table 3.4 Predictions of other genomic regions. NPV of different classifiers for different genomic regions are compared.

Results showed that the classifier did improve the prediction performance in other genomic regions as compared to the other two methods. EPRIN attained a similar high accuracy as the logistic method except for the intergenic region. This may be explained by the fact that ERPIN puts more emphasis on using the canonical poly(A) signal to make prediction than polya_svm and logistic, and the

intergenic region contains more canonical poly(A) signals AWTAAA than the transcribed region.

3. Score versus strength

The usefulness of the logistic classifier was assessed by examining the correlation between logistic score and strength of the PAS to determine whether stronger PAS score higher than weaker PAS. As discussed in the Introduction chapter, it was found that mutations in the PAS downstream region in F2 and FGG increase polyadenylation efficiency, which leads to elevated mRNA level, causing of acute blood clotting related diseases [Danckwardt et al 2004, 2006, Sachchithananthan et al 2005]. To assess whether logistic scores are able to reflect the outcome of such experimental studies, the dbSNP database from NCBI [Sherry et al 2001], and specific literature articles were used to identify SNPs and mutations for these two genes. The size of the F2 cDNA is 2,009 nts and one C→T transition was detected 11 nts downstream of the PAS, namely C2020T. In addition, two SNPs were documented in dbSNP viz. rs72550707 C2008T, and rs1799963 G2009A that were located right at the cleavage site. For FGG, SNP rs2066865 C1671T was located 10 nts downstream of the PAS. As the F2 gene carries multiple variations, it was necessary to calculate scores of all possible combinations. Scores of wild-type and variations are listed below:

Gene	Variation	Logistic Score
F2	Wild-type	0.985
	C2020T	0.990
	C2008T	0.985
	G2009A	0.987
	C2008T, G2009A	0.987
	C2020T, C2008T	0.990
	C2020T, G2009A	0.991
	C2020T, C2008T, G2009A	0.991
FGG	Wild-type	0.907
	C1617T	0.932

Table 3.5 Scores of F2 and FGG wild-types and mutants.

As shown above, the logistic score of mutants do increase except for C2008T in F2. At present, only C2020T in F2, and C1617T in FGG have been demonstrated to have health implications. Although, the ranking based on scores is consistent with the experimental studies, the ability to use the absolute score value to predict relative polyadenylation efficiency between various SNPs and mutations is far less clear.

Besides these two genes, an extensive analysis regarding score and strength of PAS in human and mouse was done. However, first it is useful to discuss the concept about the strength of PAS.

Figure 3.9 Schematic diagram about the strength of a PAS.

Previous work attempted to define a "strong" PAS of a gene as the site supported by >70% of its EST for that gene [Legendre et al 2003]. One limitation of such an approach is that since transcription proceeds from 5' to 3', an upstream PAS is transcribed before the downstream ones so that the upstream PAS has a longer time to be recognized by the polyadenylation machinery. As a result, the upstream PAS should have a higher chance to be chosen *ceteris paribus*. For this reason I adopted the following as an alternative definition about the strength of PAS for genes with multiple PAS, namely: if more ESTs support a 3' downstream PAS than the upstream one, the downstream one is stronger than the upstream one. For the example in Figure 3.9 above, PA3 is stronger than PA1 and PA2. However, no conclusion can be drawn between the pair PA1 and PA2. Additionally, the prediction by polya_svm was also used for comparison to

investigate the correlation between score and strength. The results are tabulated below.

	Human	Mouse
Number of genes with multiple PAS	3,439	920
Number of strong-weak pairs according to our definition	3,754	674
Number of strong sites with higher score	2,476 (66%)	487 (72%)
P-value of binomial test (P_0=0.5)	2.2e-16	2.2e-16
Number of strong-weak pairs according to Legendre's definition	1,696	279
Number of strong sites with higher score	1,197 (71%)	192 (69%)
P-value of binomial test (P_0=0.5)	2.2e-16	2.97e-10
Number of strong sites with higher score (smaller e-value) using polya_svm	2,051 (55%)	320 (47%)
P-value of binomial test (P_0=0.5)	1.45e-18	0.2037

Table 3.6 Correlation between score and strength of PAS in human and mouse.

Regardless of which definition was used, a statistically significant high proportion of strong PAS was associated with higher score value indicating that logistic score can reflect the strength of PAS.

4. Low score PAS in multiple PAS genes

As discussed in the previous section about PPV (Figure 3.8C), nearly 8% of PAS were misclassified on average by the logistic classifier. Low logistic score is often associated with the lack of poly(A) signal. However, so far only one non-canonical polyadenylation element UGUAN was reported [Venkataraman et al 2005] to facilitate polyadenylation in two genes viz. PAPOLA and PAPLOG. If it is assumed this element is unlikely to substitute for the poly(A) signal in as many as 8% of human genes, then it is likely that more than one type of PAS is present in a gene possessing a low score PAS. Hence, an analysis was done to investigate whether low score PAS are biased in multiple PAS genes. In this analysis, two thresholds were used to identify low score PAS viz. 0.2 and 0.3. The results are summarized below.

	Human	Mouse
Number of genes	12,303	7,710
Number of genes with multiple PAS	3,349 (27%)	920 (12%)
Number of low score PAS in ALL genes (<0.2)	668	379
Number of low score PAS in multiple PAS genes	429 (64%)	172 (45%)
p-value of proportion test	0.0 (P_0=0.27)	0.0 (P_0=0.12)
Number of low score PAS in ALL genes (<0.3)	935	488
Number of low score PAS in multiple PAS genes	583 (62%)	203 (41%)
p-value of proportion test	0.0 (P_0=0.3)	0.0 (P_0=0.12)

Table 3.7 Low score PAS in multiple PAS genes.

As listed in Table 3.7 above, 30% and 12% of human and mouse genes, respectively, contain multiple PAS. Altogether, 688 and 379 of low score PAS in human and mouse respectively used 0.2 as a threshold. In the absence of bias, one would expect around 27% (668x0.27=180) and 12% (379x0.12=45) of these low scoring PAS to reside in the midst of multiple PAS in human and mouse respectively. Surprisingly, 64% and 45% of low score PAS in human and mouse, respectively, were found in the midst of multiple PAS. Such a biased distribution

is statistically significant. The same biased distribution was found if the threshold for low score PAS was raised to 0.3.

These findings suggest that the polyadenylation activity of the majority of low score PAS are compensated by other stronger PAS in the same gene. Alternately, the weak PAS may be kept in the gene so that it can be activated to alter the 3' UTR in the presence of some unknown stimulating factors.

5. Score correlation between human and mouse

It is also interesting to explore whether logistic scores are conserved between human and mouse. Only genes with single PAS were considered. Homologous information, based on protein sequence, were obtained from HomologGene database in NCBI [HomoloGene 2009]. 3,636 homologous pairs satisfied our requirements.

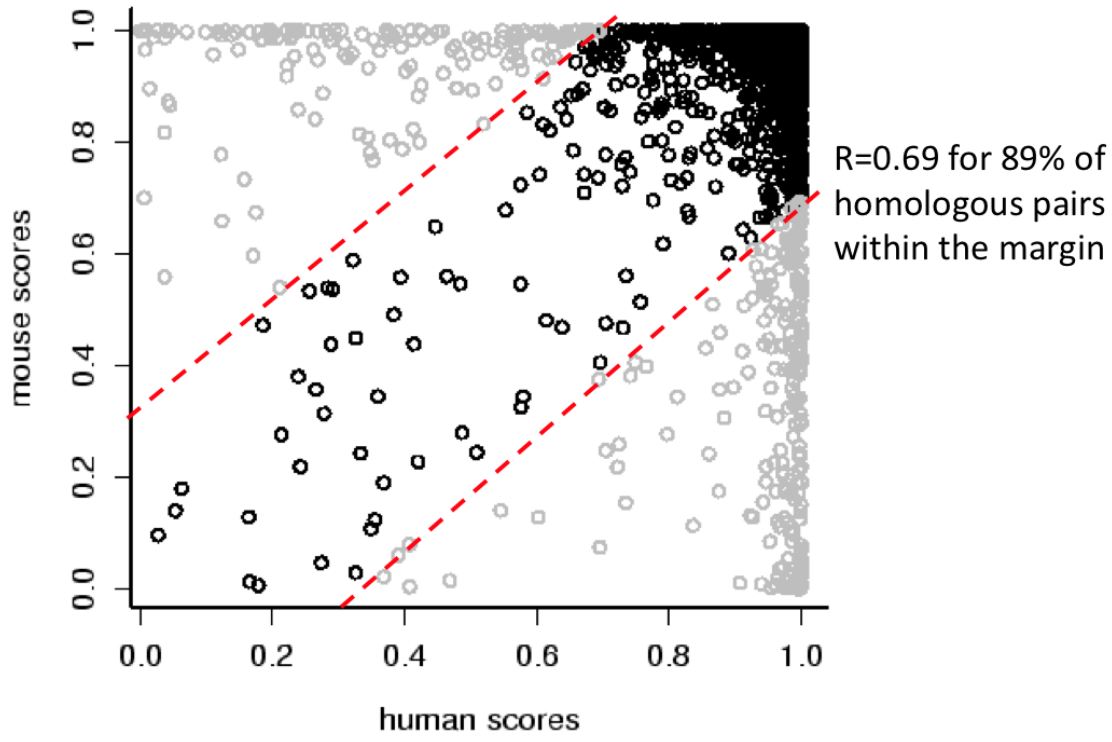


Figure 3.10 Correlation of scores between homologous genes between human and mouse.

Each dot in Figure 3.10 represents a gene. The overall correlation of score between human and mouse is only 0.19. However, the majority, 89% or 3,236 genes, are located diagonally as shown as darkened dots in Figure 3.10, and the correlation of this group is 0.69, suggesting the presence of selection pressure to conserve the core propensity of PAS though some genes exhibit great difference between homologs.

6. PAS Outliers

Previous work has identified 13 poly(A) signal hexamers [Beaudong et al 2000, Tian et al 2005]. Using the *k*mer SVD method, 16 pronounced hexamers were identified in the upstream region (Table 3.2). However, 4% of human and mouse genes do not possess any of these 16 hexamers up to 60 nts upstream from the PAS, which is double the nominal distance of the poly(A) signal from the PAS. The percentage distribution of various poly(A) signals in human and mouse is shown below.

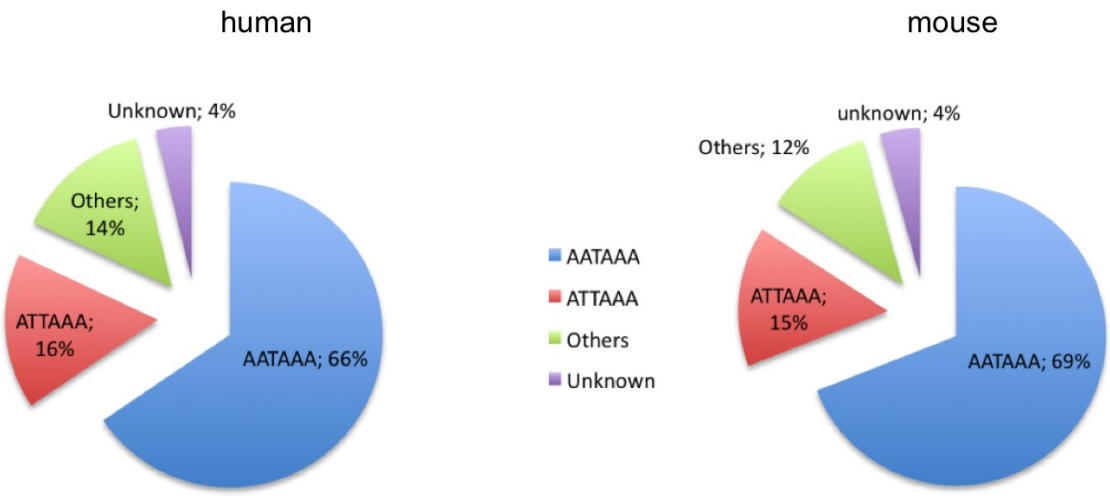


Figure 3.11 Poly(A) signals in human and mouse.

This set of non conformant polyadenylated transcripts may provide a new insight about alternative mechanisms in polyadenylation, hence further investigation of them may help to discover the shared properties of these genes. Since low score PAS is likely to be complemented by strong PAS in the same gene, the analysis was limited to selection of single PAS genes without the 16 hexamers up to 60

nts upstream from PAS. First, the level of EST supporting the PAS was assessed, with the caveat that the abundance of EST in supporting a PAS reflects its detectability rather than a true measure of its expression level. With this caveat in mind the findings are tabulated below.

	Number of single PAS genes	Number of genes without any of 16 hexamers	EST support for PAS outliers	EST support for ALL
human	6,949	121 (1.6%)	μ=11, σ=11	μ=20.64, σ=3.37
Mouse	5,150	102 (2%)	μ=5, σ=3	μ=9.13, σ=2.13

Table 3.8 EST support of PAS outliers in human and mouse.

Only a small percentage of single PAS genes do not have any of the 16 hexamers yet they were found to be polyadenylated. The EST support for all single PAS genes was estimated by averaging the repeated random sampling of single PAS genes. PAS outliers are less detectable than the mainstream as they are supported by almost half the amount of EST on average, suggesting that polyadenylation is rescued by a less optimal mechanism such as cis stimulating elements.

A search was then undertaken to identify special gene-specific sequence elements upstream and downstream of the PAS. As PAS outliers frequently

contain C-rich elements in the region up to 100 nts upstream from the PAS, they were searched for two known C-rich motifs [Yeap et al 2002, Kim et al 2007], CCCCCC and CCCUCCC with up to one substitution being allowed at any position except for U in the middle of the second motif. C-rich motifs have been reported to affect mRNA stability in erythropoietin (EPO) [Czyzyk-Krzeska et al 1999, reviewed in Waggoner et al 2003] but not polyadenylation. I speculate the suboptimal polyadenylation in mRNA maturation, due to the lack of a poly(A) signal, is compensated by increased mRNA stability in order to maintain protein level.

CCCCCC Motif	**genes with poly(A) signals**	**genes with motif**	**genes without poly(A)**	**genes with motif**
human	6,828	1,463 (21%)	121	51 (43%)
mouse	5048	971 (20%)	102	29 (28%)
CCCUCCC Motif	**genes with poly(A) signals**	**genes with motif**	**genes without poly(A)**	**genes with motif**
human	6,828	949 (14%)	121	34 (29%)
mouse	5048	691 (12%)	102	11 (10%)

Table 3.9 C-rich motif in genes without poly(A) signals.

Not only were C-rich, G-rich regions discovered up to 100 to 500 nts downstream from PAS in human but more G/C-rich were found in mouse instead.

A DNA motif finding program iTriplet [Ho et al 2009] was used to search for 10-nt long motifs with up to 2 mutations. Two G-rich motifs were found viz. GGGGCTGGAG and GGGGGGCAGG. G-rich region was reported to cause RNA polymerase II pausing [Gromak et al 2006], which may trigger transcription termination in eukaryotes. Also hnRNP H has been shown to bind a G-rich region in gene MC1R [Dalziel et al 2007]. The findings reported here are different from a previous report [Tian et al 2005], which had found G-rich regions [-100,-41] upstream and C-rich regions [+41,+100] downstream. Such a difference is probably due to the fact that only PAS outliers were included in the current analysis.

In addition, two T-rich motifs, TTGTTT and TTATCT, were identified upstream of PAS by iTriplet. It is believed that their function is to mediate stable binding of CPSF1 via hFip1 [Kaufmann et al 2004, Danckwardt et al 2007].

7. Conserved flanking region of PAS outliers

To eliminate the concern that the discovery of these PAS outliers may be solely coincidental, their conservation in remote species like human, mouse and cow was examined. Eleven PAS outlier genes were found to possess highly conserved PAS flanking regions.

Gene	Description	Gene ID	Putative poly(A) signal
STX5	syntaxin 5	Hs.6811	ATTACA
MBD6	methyl-CpG binding domain	Hs.114785	AATATT
PLEKHG3	pleckstrin homology domain	Hs.26030	AATAAC
TBC1D10B	TBC1 domain family, member	Hs.26000	AAWGAA
DLG4	discs, large homolog 4	Hs.1742	AAGGAA
PRR12	proline rich 12	Hs.57479	AACGAA
BCORL1	BCL6 co-repressor-like 1	Hs.63035	-
FGFRL1	fibroblast growth factor	Hs.53834	AWGAAA
DMWD	dystrophia myotonica, WD	Hs.1762	AATTAT
TMEM110	transmembrane protein 110	Hs.375346	AAAACA or AAACAG
TMEM30A	transmembrane protein 30A	Hs.55754	ATATTG

Table 3.10 Conserved PAS flanking regions of PAS outliers in human, mouse and cow.

Out of the 102 PAS outlier human-mouse homologous genes (Table 3.8), 11 exhibit high sequence conservation flanking the PAS. Inspection of the region 20 nts upstream from the PAS, failed to identify any conserved canonical-poly(A)-signal-like hexamers from these genes. Except for BCORL1, a list of A-rich hexamers was identified, which have not been validated experimentally. Strikingly, even the platypus genes PLEKHG3, TBC1D10B and FGFRL1 contained these A-rich hexamers, a puzzling result as one assumes it should be unfavorable for such genes to lack a canonical poly(A) signal. This implies that nature preserves a seemingly suboptimal polyadenylation for this small set of genes, and hence studying such genes experimentally may provide new insights into less understood polyadenylation compensatory factors. For example, the AU-rich database ARED 3.0 [Bakheet et al 2006] has an entry that FGFRL1 contains an AU-rich element (ARE) that is known to affect mRNA stablility. An alignment report for these 11 genes among human, mouse, cow, and in some cases platypus, can be found in Appendix E.

F. Discussion

An improved PAS classifier using logistic regression has been discussed thoroughly, suggesting that these ten features can represent the essence of an active PAS. This method was able to improve prediction in spite of using a shorter region than other methods, indicating that the core PAS elements are largely located in the [-40,+80] region. Benefiting from the tractable nature of logistic regression, it supports the view that the upstream sequence context (Table 3.2) is far more important than the downstream one in defining a PAS. However, it apparently contradicts the assertion from Chapter 2 that selection pressure may exert in the [-200,0] region. With that said, I believe the analysis has revealed two levels of information. The first is the core polyadenylation elements are largely required by all genes, and the second is gene-specific elements are positioned further upstream (up to 200nt) of the PAS.

I have also shown that logistic score corresponds to the strength of the PAS. The occurrence of low score PAS are mostly found in the midst of strong PAS, which is likely a compensatory mechanism to preserve at least one PAS per gene, and acts as alternative PAS in the presence of polyadenylation stimulating factor(s).

Although the logistic classifier was able to achieve 92% PPV for human PAS, there were still 8% or 1,366 PAS that could not be recognized by the method (Figure 3.8C). Moreover, 4% of PAS in human and mouse do not possess canonical poly(A) signals as well as the enriched poly(A) hexamers

discovered by two previous studies and the *k*mer SVD method (Table 3.2). Recall that all PAS compiled herein were supported by at least three polyadenylated ESTs. Such a perplexing observation supports the idea of the presence of additional polyadenylation elements. Furthermore, the PAS outliers carry C-rich and T-rich elements upstream, and G-rich or CG-rich region downstream that presumably function to compensate for the suboptimal nature of the PAS. To assess whether these PAS outliers regulate differently across different tissues, their expression profiles across different tissues was examined using BioGPS and random sampling [Wu et al 2009]. Overall, this examination did not find any evidence for tissue specific expression for these PAS outlier genes.

Aside from this, I identified 11 single PAS site genes without any of the 16 poly(A) signals that had highly-conserved PAS flanking regions with some conservation even among remote mammals. 10 out of 11 contained poly(A)-signal-like A-rich hexamers at around 20 nts upstream, suggesting either a novel polyadenylation factor or that CPSF1 recognition, the factor that recognizes poly(A) signals, is more flexible than previously thought. Thus further investigation is needed to explain the specificity of polyadenylation site recognition.

CHAPTER 4

iTRIPLET: A RULE-BASED NUCLEIC ACID MOTIF FINDER

A. Introduction

With the advent of high throughput sequencing techniques, large amounts of sequencing data are readily available for analysis. Natural biological signals are highly variable intrinsically making their complete identification a computationally challenging problem. Many attempts in using statistical or combinatorial approaches have been made with great success in the past. However, identifying highly degenerate and long (>20 nucleotides, nt) motifs still remains an unmet challenge as high degeneracy will diminish statistical significance of biological signals and increasing motif size will cause combinatorial explosion. Here we present a rule-based method, named iTriplet, to identify degenerate and long motifs in nucleic acid sequences.

We will adopt the sequence motif finding problem formulation originally proposed by Pevzner and Sze [Pevzner et al 2000] in this chapter. We call an oligonucleotide of length *l*, an *l*mer. A motif model is denoted by <*l*,*d*>, where *l* is the length of the motif, and *d* is the maximum number of mutations allowed with respect to the motif. An instance of a motif is termed *d*-mutant. Two *d*-mutants of the same motif must not differ by more than 2*d* differences. We call two *l*mers neighbors if their difference is ≤ 2*d*. Given *n* sequences, each of length *L* (could

be of variable length), the goal is to locate the set of *d*-mutants in each sequence from the sample where the largest difference between any pair of *d*-mutants in the set is ≤ 2*d*. In the following we will summarize two major motif finding approaches, viz. statistical and combinatorial.

The position weight matrix is often used as a statistical scoring system to identify biological signals from background. This technique implies that biological signals consist in part of conserved nucleotides that are critically important for their potency. As a result, motifs discovered by this approach tend to contain relatively invariant nucleotides at a few positions. Many transcription factor binding site prediction methods were developed based on this approach. Gibbs sampling and expectation maximization are typical techniques employed by MEME [Bailey et al 1994,1995], AlignACE [Roth et al 1998], BioProspector [Liu et al 2001], MDScan [Liu et al 2002] and MotifSampler [Thijs et al 2002]. The primary advantage of this approach is its speedy runtime and minimal memory consumption. However, statistical overrepresentation will vanish when the size of the motif to the number of mutations ratio decreases, i.e. degeneration. One improvement of this approach is to incorporate phylogenetic information in background estimation. Well-known examples of this approach include FootPrinter [Blanchette et al 2002] and PhyloGibbs [Siddharthan et al 2005]. However, such an approach is challenged by multiple substitutions occurring in distant species (homoplasy) or motif searching in a single species. Some other methods train a Markov model to capture nucleotide dependency information of known binding sites in order to make prediction for unseen cases. One extension

of the Markov model was reported in [Wang et al 2005]. The authors incorporated several features, such as gaps and polyadic sequence elements, to handle diversified transcription factor binding sites.

An alternative to a statistical approach is the combinatorial or enumerative approach [Pevzner et al 2000] where the observable biological signals are believed to be the variations of a hidden motif, and they do not exhibit conspicuous conservation at any particular position, and yet they are similar to each other. This approach is suitable for families of biological signals where the affinity of the targeting protein to the binding site relies on cooperative binding in a region rather than on a few conserved nucleotides at fixed positions. Many such examples are found in precursor RNA processing signals including the pyrimidine-rich region near 3' splice sites and the U/GU-rich region downstream of polyadenylation sites. One fundamental problem faced by the enumerative approach is the exponential growth of computing resources when the size of the motif increases. To circumvent this, existing methods such as MotifEnumerator [Sze et al 2006], MITRA [Eskin et al 2002], WINNOWER [Pevzner et al 2000], TIERESIAS [Rigoutsos et al 1998], Gemoda [Jensen et al 2006] and PMSprune [Davila et al 2007], employ various elegant pruning strategies to abandon unpromising pursuits as early as possible.

Both enumerative and statistical approaches have proven to be valuable in analyzing real biological examples and both approaches are complementary to each other. In most situations when little prior knowledge is known about the motif, we believe both approaches should be considered. Our interest is on the

discovery of motifs flanking polyadenylation sites, which are often degenerate like the downstream region or long (might due to combinatorial binding sites) therefore we have adopted the enumerative approach. We have invented a novel rule-based algorithm to identify all optimal motif candidates without the expense of exploring the entire 4^l space exhaustively. In addition, our algorithm is designed to be highly parallelizable so as to exploit today's parallel computing technology in handling massive biological data. As a proof of concept, we have evaluated our algorithm using the simulated data described in [Pevzner et al 2000]. Also we have demonstrated that our method is able to identify motifs in real promoter sequences, 5' and 3' untranslated regions (UTR), and distal enhancers from different species. Results show that our method can solve highly degenerate and/or longer motifs that overwhelm the capabilities of other methods. Furthermore, we have compared the prediction accuracy of our method with the statistical motif finding methods mentioned above and find that our method is equal to and sometimes better than these methods. Besides *in-silico* simulations, we have also verified our prediction of downstream polyadenylation motifs for three human genes using a dual Luciferase assay. Our software is developed in C++ and standard template library (STL). It has been tested on Linux platform. The software can be downloaded freely from this website http://www.rci.rutgers.edu/~gundersn/iTriplet.

B. Method

1. iTriplet Algorithm

Our rule-based enumerative algorithm is named iTriplet. It stands for **i**nter-sequence **triplet**s. A triplet consists of three neighboring *l*mers (less than 2*d* differences from each other) sampled from three different sequences. The 'inter-sequence' part of the iTriplet algorithm systematically explores tripartite combinations of *l*mers from different sequences in order to identify motif(s) that span all sequences in the sample. The span of a motif refers to the number of sequences containing its *d*-mutant. For clarity, we will explain our method by limiting to only one motif in the sample, and every sequence contains at least one occurrence *d*-mutant of the motif even though our method can deal with multiple motifs and 10-20% of contamination. We will describe our iTriplet algorithm in two parts: the 'inter-sequence' part will be discussed first, followed by the Triplet algorithm.

2. The inter-sequence part of iTriplet

If sufficient numbers of sequences are given, and the motif model is not highly degenerate, i.e. small *d* with respect to *l*, the likelihood that an *l*-sized motif can span through all sequences by chance is rare. Based on this insight, we utilize the span of a motif as the indicator to identify unusual motifs in a sample.

The inter-sequence part of iTriplet consists of two stages: initialization stage and expansion-pruning stage. It is illustrated in Figure 4.1 below:

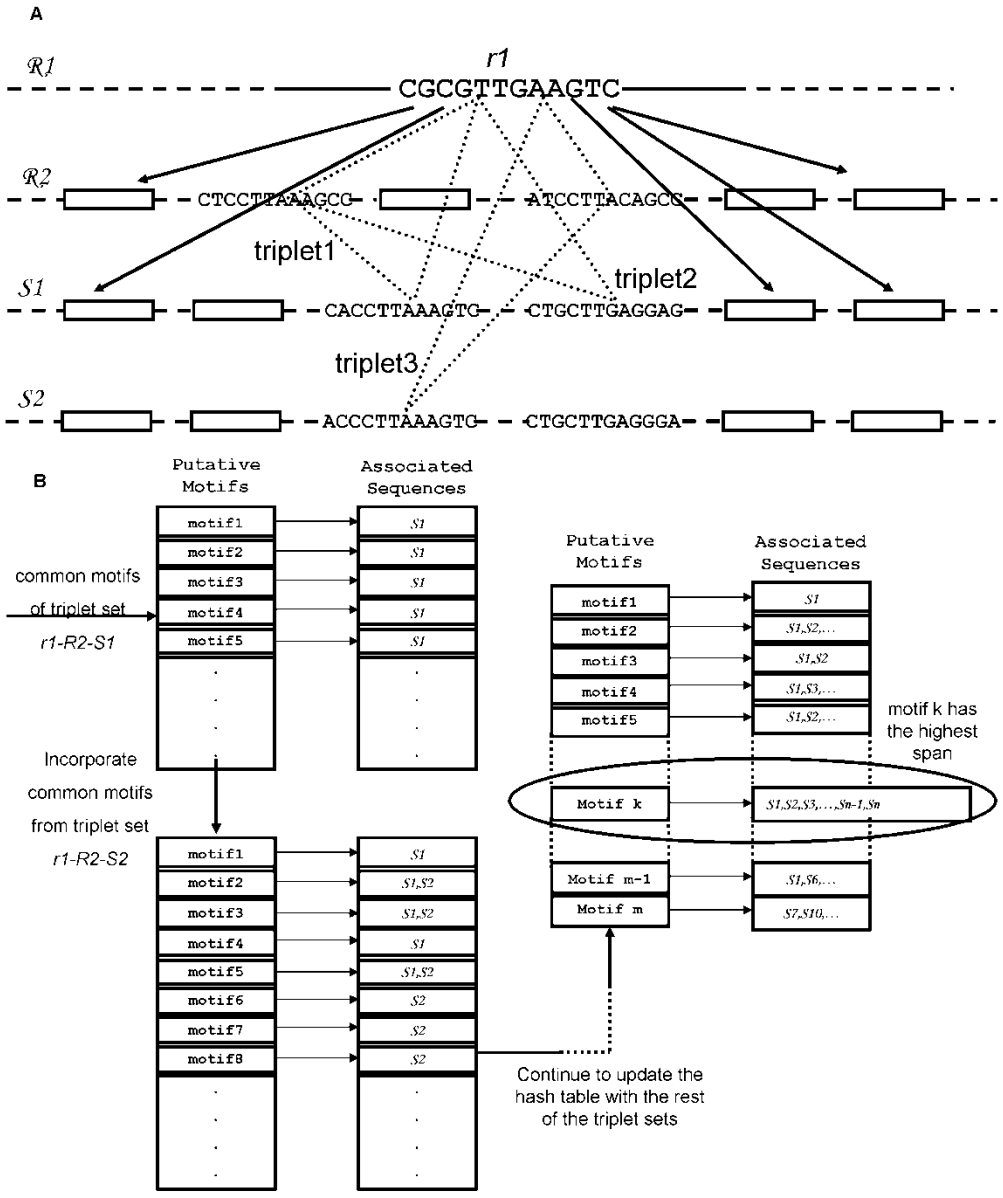


Figure 4.1 Inter-sequence algorithm. (A) For each *l*mer *r1* in *R1*, identify 2*d*-mutants in sequences R2, S1, S2, … The rectangular box represents the 2*d*-mutant of *r1*. The dotted line triangle represents a triplet. (B) Hash table to keep track of the span of the putative motif. Hash table consists of two parts viz. key and value. In this case, the key is the putative motif; value is

a list of unique sequence IDs. Putative motifs are produced by the Triplet algorithm. They are common motifs to triplets.

Here is the procedure of inter-sequence phase: given a set of *n* sequences and a motif model <*l,d*>, randomly designate two sequences from the sample as reference sequences, namely $\mathcal{R}1$ and $\mathcal{R}2$, and the rest as non reference sequences $S_1, S_2, \ldots, S_{n-2}$.

Initialization stage: Randomly select an *l*mer (*r1*) from $\mathcal{R}1$ and a non reference sequence, say S_i. Identify all possible triplets based on *r1*, *l*mers from sequences $\mathcal{R}2$ and S_i as illustrated in Figure 4.1 A. For each triplet, identify the set of motif(s), if any, common to the triplet using the Triplet algorithm (will be discussed later). Store the returned common motif(s) and its associated sequence IDs in a hash table as shown in Figure 4.1 B.

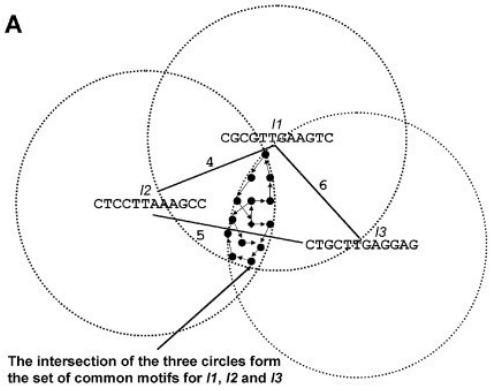
Expansion-pruning stage: Randomly select an unprocessed non-reference sequence, say S_j. Similar to initialization stage, identify all triplets based on *r1*, *l*mers from sequences $\mathcal{R}2$ and S_j. Identify the set of common motifs of all triplets using Triplet algorithm and store them in the hash table. Prune the hash table by removing all motifs that do not span all sequences processed so far. If the hash table is not empty after pruning, repeat the expansion-pruning stage with the next unprocessed non-reference sequence. If the hash table is empty after pruning, return to the initialization stage, randomly pick a different *l*mer (*r1*) from $\mathcal{R}1$, and repeat the same two-stage inter-sequence process again until all *l*mers in $\mathcal{R}1$

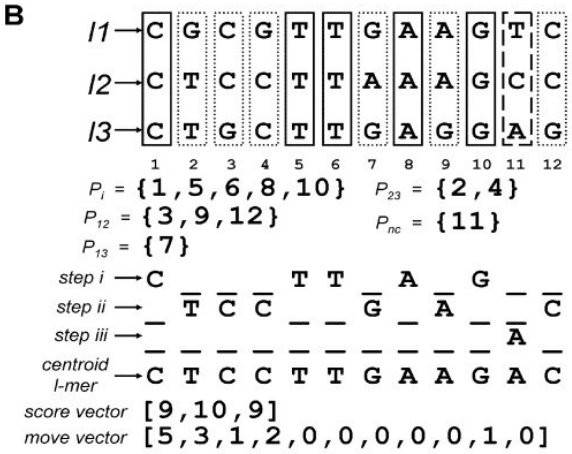
have been processed. If all non-reference sequences have been processed and the hash table is not empty, then return motif(s) in the hash table to the calling program.

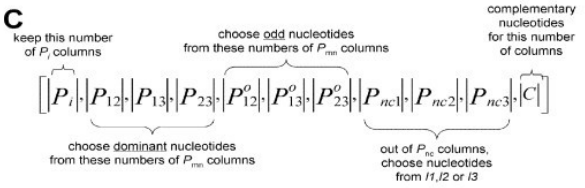
As described above, the processing of different *l*mer *r1* in $\mathcal{R}1$ are completely independent of each other. It means that they can be executed simultaneously wherein not even a single synchronization point is required. Therefore, given M processors, the algorithm can trigger up to (M-1) concurrent processes simultaneously. Theoretically, the performance gain by parallelizing this step is (M-1) times for a M-processor system where one processor is designated for overall coordination purposes. Our current parallel version of iTriplet is implemented based on this idea.

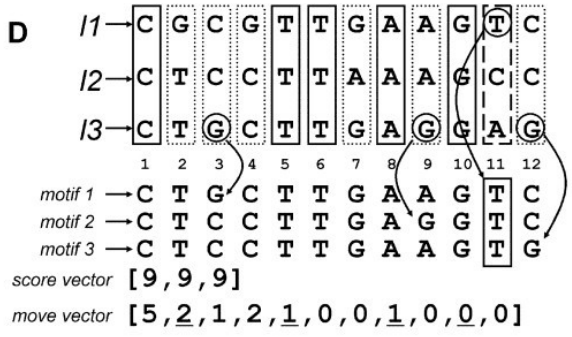
3. The Triplet part of iTriplet

The purpose of this part of the algorithm is to uncover the complete set of motifs common to all members of the triplet in a deterministic and efficient way. The clues solely come from the similarities and differences among the three *l*mers rather than the enumeration of all possible *l*mers. It is efficient because the number of motifs shared among all three *l*mers should be small. By example, the estimated probability of any three *l*mers to share at least one common motif for models <12,3> and <30,9>,is 5.47×10^{-4} and 2.97×10^{-4}, respectively.


A
I1
CGCGTTGAAGTC
4
6
I2
CTCCTTAAAGCC
5
I3
CTGCTTGAGGAG
The intersection of the three circles form
the set of common motifs for I1, I2 and I3


B
I1 → C G C G T T G A A G T C
I2 → C T C C T T A A A G C C
I3 → C T G C T T G A G G A G
1 2 3 4 5 6 7 8 9 10 11 12
P_i = {1,5,6,8,10}
P_23 = {2,4}
P_12 = {3,9,12}
P_nc = {11}
P_13 = {7}
step i → C _ _ _ T T _ A _ G _ _
step ii → _ T C C _ _ G _ A _ _ C
step iii → _ _ _ _ _ _ _ _ _ A _
centroid l-mer → C T C C T T G A A G A C
score vector [9,10,9]
move vector [5,3,1,2,0,0,0,0,0,1,0]


C
keep this number of P_i columns
choose odd nucleotides from these numbers of P_mm columns
complementary nucleotides for this number of columns
choose dominant nucleotides from these numbers of P_mm columns
out of P_nc columns, choose nucleotides from I1,I2 or I3


D
I1 → C G C G T T G A A G T C
I2 → C T C C T T A A A G C C
I3 → C T G C T T G A G G A G
1 2 3 4 5 6 7 8 9 10 11 12
motif 1 → C T G C T T G A A G T C
motif 2 → C T C C T T G A G G T C
motif 3 → C T C C T T G A A G T G
score vector [9,9,9]
move vector [5,2,1,2,1,0,0,1,0,0,0]

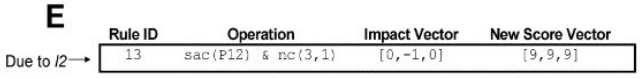

E
Rule ID
Operation
Impact Vector
New Score Vector
Due to I2 →
13
sac(P12) & nc(3,1)
[0,-1,0]
[9,9,9]

Figure 4.2 Intuition of Triplet algorithm. A triplet consists of 12mers l1, l2 and l3. l1 and l2, l1 and l3, and l2 and l3 contain 4, 6 and 5 differences respectively as labeled in the lines connecting them. Use the 12mer as the center to draw an imaginary circle. Each circle denotes the set of neighboring 12mers that are no more than 3 differences from the center 12mer. In other words, each circle represents the set of putative motifs that generate the center 12mer. Note that we do not actually generate the set of putative motifs. Centroid lmer is denoted by a diamond shape dot. The goal of the algorithm is to uncover all members of the set in the intersection (dark gray) of the three sets. (B) Centroid lmer construction. Shown are three patterns of columns viz. same nucleotide in three 12mers Pi (solid line vertical boxes in positions 1, 5, 6 and 10), all different nucleotides across three 12mers Pnc (vertical box with dashed boundary in position 11), and two out of three 12mers having the same nucleotides Pmn (dotted line vertical boxes in positions 2, 3, 4, 7, 9, and 12). The centroid lmer is constructed in stage 1 of Triplet algorithm described in the text. The number of identical positions between the centroid lmer and *l1*, *l2* and *l3*, is represented by the score vector and the selection of nucleotides encoded in move vector (C) Structure of move vector. (D) Exploratory scheme discovery from stage 2 of Triplet algorithm. Centroid lmer constructed in Figure 2B is modified by the composite operation of sac(P12) and nc(3,1) to create three extra motifs near its neighborhood. (E) Example of applying rule 13 to create a new move vector in (D).

a) Data structures of Triplet

Before we describe the algorithm, we need to define two main data structures used by this algorithm viz. move vector and score vector. The three *l*mers passed into this process are stacked up conceptually to form *l* numbers of three-nucleotide tall columns as shown in Figure 4.2 B. These columns must fall into one of the three patterns: (I) with identical nucleotides denoted by P_i; or (II) with all different nucleotides, denoted by P_{nc}; or (III) with two out of three nucleotides being the same, denoted by P_{mn} where m and n denote the indices of the two *l*mers with dominant nucleotide. We will show later that common motifs are discovered by various ways of selecting nucleotide from these three types of columns. Such selection is captured in a move vector as illustrated in Figure 4.2 C. In addition, each move vector is associated with a score vector which is defined as [*i1*,*i2*,*i3*], where *i1*, *i2* and *i3* denote the numbers of identical positions between the motif represented by the move vector and the three given *l*mers *l1*, *l2* and *l3*, respectively.

b) Three stages of Triplet

Triplet algorithm consists of three stages: 1) centroid *l*mer construction, 2) exploratory scheme discovery, and 3) motif generation. Below is the description:

Stage 1: centroid *l*mer construction. Given a triplet of three *l*mers from the calling program, identify the three column types P_i, P_{mn} and P_{nc} as discussed above. Check if the triplet satisfies this inequality: $l - d \le |P_i| + |P_{mn}| * 2/3 + |P_{nc}| * 1/3$

(derivation is in Appendix F) where $|P_{i,}|$, $|P_{mn}|$ and $|P_{nc}|$ denote the number of P_i, P_{mn} and P_{nc} patterns respectively. If the given triplet fails to satisfy this inequality, return no common motif and exit. Otherwise take these three steps to construct the initial move and score vectors: i) take the common nucleotides from columns P_i, ii) take the dominant nucleotides from P_{mn}, and iii) for columns P_{nc}, take the nucleotides from the *l*mer which is currently farthest from the work-in-progress centroid *l*mer produced by the previous two steps. Pass the newly created move and score vectors to stage 2 for further processing.

Stage 2: exploratory scheme discovery. Based on the excess score(s) (> *l-d*) in one or more of the three values in the initial score vector, formulate alternative ways to select nucleotides from P_i, P_{mn} and P_{nc} patterns through the 61 rules (will be discussed later). An execution of a rule produces a new set of move vector(s) and its associated score vector. Repeat stage 2 processing of the new move vector(s) until all newly generated score vector(s) becomes [*l-d,l-d,l-d*] i.e. no excess score. Pass all move and score vectors generated to stage 3.

Stage 3: motif generation. Generate motif by going through each value in the move vector, and select the specified number of column patterns and associated nucleotides accordingly. When all move vectors are processed, return all motifs to the calling program.

Regarding the rules mentioned in stage 2 of Triplet algorithm, they are actually made of five basic operations listed in Table 4.1 below:

Operations	Description	Examples based on Figure 4.2 D if possible
sac(P_{mn})	Instead of choosing the dominant nucleotide from P_{mn} column, choose the odd nucleotide.	sac(P_{12}), take 'G' at position 3 from *l*3 instead of 'C' from *l*1 or *l*2
compl(P_{mn})	Instead of choosing the dominant or odd nucleotide from P_{mn} column, choose nucleotides complementary to them.	Apply on the 2nd column, compl(P_{23}), take nucleotides complementary to 'G' and 'T', i.e. choose 'A' or 'C' for position 2.
nc(i,j)	Instead of taking nucleotide from *l*mer$_i$, choose from *l*mer$_j$ in a P_{nc} column.	Apply nc(3,1) to position 11. Instead of choose 'A' from *l*3, choose 'T' from *l*1 at position 11.
nc(i,0)	Instead of taking nucleotide from *l*mer$_i$, choose from the complementary nucleotide of a P_{nc} column.	Apply nc(3,0) to position 11. Instead of choose 'A' from *l*3, assign the complementary nucleotide 'G' to position 11.
sac_i(P_i)	Instead of keeping the nucleotide identical to all lmers in the triplet, take the three complementary nucleotides.	Apply sac_i(P_i) to position 1. Take 'A', 'G' or 'T' instead of 'C' at position 1.

Table 4.1 Five basic operations for triplet processing of iTriplet algorithm.

These five basic operations are the only possible alternatives to the selections which produce the centroid *l*mer. The basic operation can be applied individually or be combined with one other basic operation to act like a single operation, namely a composite operation. Basic or composite operations act on the current move vector in the light of its score vector. To facilitate searching, we pack the basic/composite operation and its impact or changes on the current score vector, namely impact vector, into a new construct called rule as shown in

Figure 4.2 E. These 61 rules are further organized into three non-mutually exclusive groups, each group has 42 rules, according to which *l*mer in the triplet possesses excess score (full list can be found in the Appendix G). The decision to select a rule is determined by the three conditions. First, it has not been chosen already. Second, the three values of the new score vector, obtained by the addition of the impact vector and the current score vector, must be ≥ *l-d*. Third, the triplet contains the column pattern(s) required by the basic and/or composite operation. Notice that every rule will reduce the total score value of the new score vector. It means that successive applications of these rules will eventually create a score vector of its minimum score values [*l-d*,*l-d*,*l-d*] and that marks the terminal state.

Regarding stage 3, one move vector may generate more than one motif. For the example in Figure 4.2 D, the new move vector due to rule 13 is [5,2,1,2,1,0,0,1,0,0,0]. The first value specifies to select the nucleotides from the five P_i column patterns which are found in positions 1, 5, 6, 8 and 10 (see Figure 4.2 B). Since there are exactly five P_i column patterns, only one way is possible. The second value of the move vector specifies to choose dominant nucleotides from two P_{12} column patterns out of three and to choose the odd nucleotide from the remaining one. It will generate three possibilities. The rest of the values in the move vector will be processed similarly.

We have given the full description of iTriplet algorithm. Regarding the correctness of the algorithm, at this stage, we have not come up with a theoretical proof yet, however we have conducted extensive testing of more than

14,000 cases including models <11,2>, <12,3>, <13,3>, <15,4>, <28,8> and <40,12>; over 2,000 cases per model. In each case, we had generated 20 sequences each of length 600 with all nucleotides occurring equally likely. In each sequence, a single *l*-size *d*-mutant was planted at a random location. After each run, we checked whether the returned motif from iTriplet was the same as planted or not. iTriplet performed correctly for all cases.

4. Time and Space Complexities of iTriplet

The inter-sequence part of iTriplet mainly iterates all combinations of triplets among sequences. Therefore, for model <*l,d*>, we estimate the time complexity of the inter-sequence part of iTriplet to be $O(nL^3pl)$ where *n*, *L* and *p* are the number of sequences, length of sequence and probability to form a triplet that shares at least one motif. As discussed before, we estimate *p* should be in the range of 10^{-4}, and *L* should normally be 10^2. Therefore, the effective time complexity of the inter-sequence part ranges from $O(nLl)$ to $O(nL^2l)$. Stage 2 of Triplet part should generate all possible score vectors as long as the score value between each *l*mer and the centroid *l*mer is at least *l-d*. In the worst case scenario, there are d^3 score vectors. The generation of actual motifs based on the move vector in step 3 should depend on the size of the motif *l*. Therefore the time complexity of Triplet is $O(d^3l)$. Hence the overall time complexity of iTriplet is $O(nL^3pl^2d^3)$. For PMSprune, the time complexity is $O(nL^2N(l,d))$, where $N(l,d)$ is $\sum_{i=0}^{d} {}^{l!}\!/_{(l-i)!i!}3^i$. After eliminating the common terms, the main difference lies in the growth of Lpl^2d^3 and $N(l,d)$ in iTriplet and PMSprune, respectively. When the motif model is small, $N(l,d)$ is smaller than Lpl^2d^3. However, when *l* increases, the

combinations of *N(l,d)* grows exponentially. iTriplet's space complexity depends on the degeneracy of the model, therefore it is *O*(*N(l,d)*) before pruning. After pruning, the space requirement will shrink.

C. *Results*

1. Simulated data

In order to examine how iTriplet method can solve more degenerate and longer motifs, we compared it with some well known enumerative methods using simulated data. The simulated sequences were generated as described in the Appendix H. Simulated datasets were constructed using a wide range of *l* and *d* parameters in order to compare the performance of different methods in dealing with various sizes of the motif and/or noisy situations. The sequential version of our method was compared with three other well-known methods that have the same focus to guarantee finding the optimal motif viz. MotifEnumerator [Sze et al 2006], RISOTTO [Pisanti et al 2006] and PMSprune [Davila et al 2007] (see Appendix I for program versions). Sequential tests were conducted on a Linux machine equipped with an Intel P4 3 GHz processor and 2 Gbytes of memory. All methods can successfully identify the planted motifs in the simulated dataset unless the runtime was longer than 6 hours. We also repeated the same set of tests for the parallel version of iTriplet on a three-node Linux cluster equipped with the same processor as a sequential test. Results are tabulated in Table 4.2 below:

Models	Neighborhood Probability	MotifEnumerator	RISOTTO	PMSprune	iTriplet	iTriplet (parallel)
11,2	0.7%	6s	2.2s	1s	2s	1s
12,3	5.4%	1m	40s	4s	33s	18s
13,3	2.4%	2m	33s	2s	6s	4s
14,4	11%	-[a]	8m	1m	3m	2m
15,4	5.6%	-	6m	16s	36s	19s
16,5	19%	-	82m	13.5m	26m	13m
18,6	28%	-	-[b]	-[b]	3h	1.5h
19,6	18%	-	-	-	27m	14m
24,8	23%	-	-	-	4h	2h
28,8	3%	-	-	-	19s	10s
30,9	5%	-	-	-	2.3m	1.5m
38,12	7%	-	-	-	1h	33m
40,12	3%	-	-	-	5m	4m

Table 4.2 Methods comparison on simulated datasets.

Neighborhood probability refers to the probability that two *l*mers differ by no more than 2*d* differences. The formula to calculate neighborhood probability is stated in the Additional file 1. Time is measured in seconds (s), minutes (m) or hours (h). (a) MotifEnumerator ran out of memory for *l* greater than 13. (b) Program took more than 6 hours to handle for the model <18,6> or longer. For the parallel version of iTriplet, reported runtime is the longest lapse time required for all nodes to finish.

The second column of Table 4.2 is the neighborhood probability of each model, which is the probability that any two *l*mers differ by no more than 2*d* by chance , a good indicator to reflect the degree of degeneracy of the model.

For short motifs (<16 nucleotides) iTriplet is comparable to the fastest (PMSprune) and is significantly faster than MotifEnumerator and RISOTTO. When motif length is longer than 16, all other methods take longer than 6 hours to process. Note that iTriplet is able to process highly degenerate <18,6> and <24,8> models which cannot be handled by these other three methods as well as other statistical based methods such as MEME, MotifSampler and BioProspector. Based on these results, we learned that the performance of all methods depends on *l* and *d*, but to a different extent. Intriguingly, the runtime of PMSprune quadrupled, though still very fast, when *l* increased from 12 to 15 even though the neighborhood probability remained relatively at the same level. A similar trend is also observed in RISOTTO but with even higher fold increment

in runtime. Such a phenomenon is not observed in our method. When neighborhood probability is doubled in models <12,3> versus <14,4>, and <14,4> versus <16,5>, the runtime of PMSprune increased 15 and 13.5 times respectively and RISOTTO increased 12 and 10 times respectively whereas iTriplet only increased 6 and 9 times, respectively. Based on these observations, we can understand that the algorithms employed by RISOTTO and PMSprune are quite sensitive to both *l* and *d* even when the neighboring probability remains at the same level. Thus RISOTTO and PMSprune take a longer time to search for the optimal motif; whereas the combined effect of *l* and *d* on performance was less severe for iTriplet. This explains why RISOTTO and PMSprune encountered difficulty in handling longer motif models. This does not exclude that iTriplet is unaffected by large *d* (high degeneracy). But one distinctive feature of our algorithm is that it can split the task into smaller subtasks which can be run independently in parallel. When comparing sequential and parallel versions of iTriplet, the parallel version averaged 1.77 times performance gain in a three-node cluster that is quite close to the theoretical gain 2.0. Testing based on the simulated data revealed that different methods have different tradeoffs in tackling the general <*l,d*> motif problem therefore further investigation is still needed to cope with various challenges of this problem.

2. Real biological sequences

Besides simulated datasets, we tested our method using multiple sets of real biological sequences. One issue with real biological sequences is the lack of prior knowledge about the size and maximum numbers of mutations permitted by

the motif. The optimal motif(s) comes from the model having the smallest neighborhood probability and produces the least number of motifs. In order to pin down the optimal motif, the algorithm must be run for a range of *l* and *d*. But we have found that the search of the optimal *l* and *d* can be done methodically by making use of the neighborhood probability of each model. In the situation when iTriplet has found too many motifs for the specified model then we can conclude that the model is too lax and so a more stringent model should be used, by increasing *l* or reducing *d* or both at the same time. Alternatively, once a satisfactory model is found, one can look for shorter models with similar neighborhood probability if the shorter alternative gives a similar result. In order to ease the effort for searching for the optimal model, iTriplet provides an autonomous mode option. Under autonomous mode, the program will explore various models using the strategy just described, and return the best models with motif length from 6 to 40 bases and maximum number of differences from 1 to 12. But the user also has the option to limit the size of motif to a specific range. Although many models are examined, only a very limited numbers of models, usually none or one, can provide the optimal motif unless the given sequences contain multiple motifs. Several reasons are that a slight change in the size and/or the maximum number of mutations will result in a substantial change in neighborhood probability which can be seen in Table 4.2. As mentioned in the Introduction section, we have included promoter and 5' UTR regions from four genes commonly chosen as test cases for motif finding algorithms [Blanchette et al 2002, Eskin et al 2002, Davila et al 2007]. In addition, we have also added a

set of 3' UTR sequences in our test in order to understand how our method performs in other regions of a gene (details in Appendix J). Table 4.3 summarizes the prediction by iTriplet for various genes and genomic regions.

Gene :	**Preproinsulin (IEB1) promoter+5' UTR**	**Remarks**
iTriplet	GTYYGGAAAYTGCAGC**YTCAGCCCC**	<25,2> model
PMSprune	CAGC**CTCAGCCCC**TT	
MITRA	C**CTCAGCCCCC**	
Published	CTCAGCCCCCAGCCATCTGCCGACCCCCCC	Transfac ID: R04457
Gene:	**DHFR (promoter+5' UTR)**	**Remarks**
iTriplet	**RWSTSGCGCSAAAC**Y	<15,3> model
PMSprune	**ATTTCG**T**G**GGC**A**	
MITRA	TGCA**ATTTC**GC**GCCA**AAC	
Published	ATTTCGCGCCAAA	Transfac ID: R01928
Gene:	**Metallothionein promoter+5' UTR**	**Remarks**
iTriplet	TTT**TGCRC**T**CG**YCCC	<15,1> model
PMSprune	CTC**TGC**A**C**A**CGG**CCC	
MITRA	**TGCGCCCGG**	
Published	TGCGCCCGG	Transfac ID: R08298
Gene:	**c-fos serum response element promoter+5' UTR**	**Remarks**
iTriplet	**CCATATTAGGAC**ATCTGCGT	<20,1> model
PMSprune	**CCA**AAT**TT**G	
MITRA	**CCATATTAGGACA**	
Published	CAGGATGTCCATATTAGGACATC	Transfac ID: R00466

3'UTR Regulatory Elements	iTriplet Prediction only	Published	Remarks
AU-rich (ARE)	**TTTTATTTATTTT**T	WWTTATTTATTWW	<14,3> model
Cytoplasmic Polyadenylation element (CPE)	**TTTTAAT**	TTTTAT and TTTTAAT	<6,1> model
Pumillio binding element (PBE)	**T**K**T**W**AATA**	TGTAAATA	<8,1> model

Table 4.3 iTriplet prediction using real biological sequences.

Motif predicted by iTriplet is presented in consensus sequence. Bold and underlined sequence represents correctly predicted nucleotide. Transfac IDs are obtained from TRANSFAC database [Wingender et al 1996]

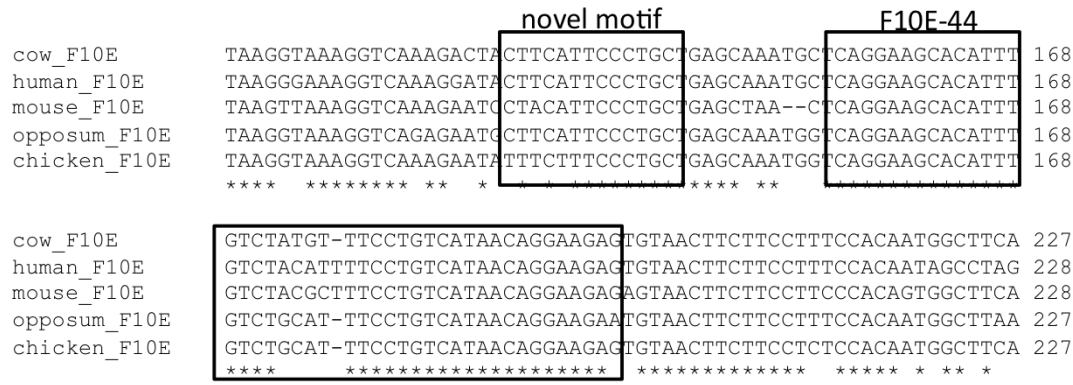
3. Distal enhancers

In addition to 5' and 3' UTR sequences, we have also applied our method to the search for distal enhancers. In a recent article [De Val et al 2008], a combinatorial regulation mechanism was reported to drive the expression of genes in the vasculogenesis pathway. The authors discovered a 44-nt conserved, and overlapping enhancer, namely FOX:ETS motif, in the MEF2C locus that binds transcription factors FoxC2 and Etv2. The binding of both, not just one, are required for vascular development in mouse. Even though the 44-nt enhancer was sufficient to cause endothelial specific expression, its effect vanished after E10.5. However, when the longer 900-nt long flanking region of FOX:ETS (called F10E in the original article) was used, its activity persisted throughout embryogenesis in "blood and lymphatic vasculature", meaning that some other unknown cis elements in F10E may participate in vascular

development. Moreover, the authors also discovered that FOX:ETS motif was not only found in the MEF2C locus, but also found in 5 other vasculogenesis genes viz. FLK1/KDR, TIE2/TEK, TAL1, NOTCH4, and CDH5/VE-CAD. At present, no motif has been identified in F10E besides FOX:ETS. As a result, it was of interest to identify any additional motif(s) shared by these six genes. Thus, a 886-nt long fragment flanking the FOX:ETS motif was extracted in each of the six genes from human. By using multiple alignment program T-COFFEE [Notredame et al 2000] to align these six 886-nt long fragments, we did not find any conserved region shared among them. Next, iTriplet was used to search for motifs using different models such as <12,3>, <13,3>, and <14,3>, and it was found that model <14,3> yielded the best motif CTCCATTGCCAGCT as shown in Figure 4.3.

A

B



C

Figure 4.3 Motif in vasculogenesis genes. A) 14-nt long consensus generated by Weblogo [Crooks et al 2004], B) location of FOX:ETS and iTriplet predicted motifs CTCCATTGCCAGCT in MEF2C, FLK1/KDR, TIE2/TEK, TAL1, NOTCH4, and CDH5/VE-CAD, C) multiple alignment of MEF2C orthologs in human, mouse, cow, opossum and chicken.

The novel motif discovered by iTriplet is not only shared among vascular development genes, it also exhibits high conservation among remote orthologs of the MEF2C locus as shown in Figure 4.3C. Hence this result is promising for further experimental studies about its biological function.

4. Multiple motifs

Multiple motifs are often identified by iTriplet for real biological sequences. Four reasons account for this: 1) the number of sequences considered is small, mostly 4 in our test therefore resulting in a higher chance to encounter random span, 2) a naturally occurring recognition site is not necessarily represented by one consensus, 3) it is possible for the biological sequence to carry more than one signal especially in the 3' UTR, and 4) the presence of low complexity repeats.

Therefore we need a scoring system to filter out random from genuine motifs. Since only a small number of sequences are given, the set of true motif instances must resemble each other more than a set of random *l*mers; otherwise no conclusion can be made. As we have discussed in the inter-sequence algorithm section, if members of the triplet are very similar to each other, the intersection will become big, i.e. high numbers of common motifs. Based on this property, we derived a straightforward scoring system based on the numbers of common motifs uncovered to support whether the finding is statistically significant. Due to this, the 5' and 3' overlapping neighbors of the true motif are often included as part of the prediction as well. Therefore in some cases of the genes listed in Table 4.3, the predicted motif is longer than the model specified.

Each prediction is a consensus of a number of common motifs. The method of constructing the consensus is similar to the frequency plot of Weblogo [Crooks et al 2004]. Nucleotides with frequency at a position greater than 30% will be included in the consensus sequence. As can be seen from Table 4.3, our predictions for promoter and 5' UTR sequences, and 3' UTR regulatory elements are largely consistent with published experimental data.

5. Sensitivity and specificity test

We also measured the prediction accuracy of iTriplet in predicting transcription factor binding sites in E. Coli. These binding sites are experimentally validated and documented in the RegulonDB database [Salgado et al 2004]. The test was conducted using the three-level testing framework described in [Hu et al 2005]. Under this testing framework, the prediction made by a method is measured at the nucleotide, binding site and motif levels. In the first and second levels, i.e. nucleotide and binding site levels, sensitivity, specificity, performance coefficient and *F*-measure are computed based on the true positive (*TP*), false positive (*FP*) and false negative (*FN*) information gathered by comparing the predicted and actual binding sites. Performance coefficient and *F*-measure were originally proposed by [Pevzner et al 2000,Tompa et al 2005] and [Hu et al 2005] respectively. Both of them have the advantage to combine sensitivity as well as specificity perspectives into a single number so as to ease interpretation. The formula for these four measurements can be found in the Appendix K. Note that at the binding site level, a prediction is considered correct when the predicted binding site overlaps with the actual binding site by at least one nucleotide.

These four measurements were calculated for each transcription factor individually. Averaged measurements of all transcription factors are used for method comparison. The Kihara group [Hu et al 2005] also suggested a third level assessment that is motif level. The rationale of this extra level test is to assess the adaptability of the method to make correct predictions for a wide range of transcription factors. The motif level measures the fraction of correct predictions out of all binding sequences and transcription factors. iTriplet was compared with the top three performers, i.e. MEME, BioProspector and MotifSampler, listed in Table 1 of [Hu et al 2005], and WEEDER [Pavesi et al 2004]. For each method, the parameter setup was adopted from [Hu et al 2005] except that no background sequence information was used for BioProspector. Motif length was set to 15, the same length used in [Hu et al 2005] except WEEDER where the maximum supported length is 12. We chose the maximum differences in the range from 3 to 5. For accuracy measurements, the top five predictions were used for the three selected methods. But in our case, we selected only the highest score consensus motif(s) instead of the top five used in [Hu et al 2005]. Although only BioProspector and MotifSampler exhibit variation in prediction even for the same input sequences, in order to maintain fair treatment, we still repeat the test ten times for all methods. Results are tabulated in Table 4.4 below:

Algorithms	Nucleotide level				Binding level				Motif level	
	nPC	*nSn*	*nSp*	*nF*	*sPC*	*sSn*	*sSp*	*sF*	*mSr*	*sSr*
iTriplet	0.195	0.292	0.322	0.286	0.319	0.489	0.418	0.422	0.853	0.591
MEME	0.180	0.551	0.214	0.296	0.258	0.733	0.280	0.397	1.000	0.817
WEEDER	0.128	0.274	0.245	0.208	0.263	0.538	0.332	0.367	0.833	0.532
BioProspector	0.102	0.372	0.129	0.179	0.212	0.704	0.224	0.328	0.986	0.670
MotifSampler	0.052	0.257	0.068	0.091	0.106	0.422	0.111	0.162	0.461	0.392

Table 4.4 Prediction Accuracy of iTriplet versus four others motif finding methods.

PC, *Sn*, *Sp* and *F* are performance coefficient, sensitivity, specificity and *F*-measure level respectively. Prefixes ‘n’ and ‘s’ represent nucleotide or binding site level measurements respectively. *mSr* and *sSr* are motif and sequence level accuracy respectively.

Table 4.4 shows the averaged measurements of iTriplet together with four other motif finding methods. iTriplet has demonstrated better prediction accuracy than the other four methods at both nucleotide as well as binding site levels except the *F*-measure is second at the nucleotide level. However, our *mSr* and *sSr* scores are ranked third mainly because these two measurements tend to favor methods with high sensitivity regardless of specificity. In the extreme situation, if a method predicts all nucleotides are part of a motif, it will score 1 for *mSr* and *sSr*. This point is further evidenced by the disproportionality of sensitivity and specificity of the other three methods except WEEDER at both nucleotide and motif levels. Therefore we think *PC* and *F*-measure are fairer measurements of prediction accuracy than *mSr* and *sSr*.

6. In vitro verification of predicted poly(A) downstream elements

To examine whether motifs predicted by iTriplet had biological activity, we chose to examine sequences important in the 3' end processing of mammalian pre-mRNA, in particular sequences found just downstream of the cleavage and polyadenylation site. Almost all eukaryotic mRNAs contain a post-transcriptionally-added poly(A) tail that is important for many aspects of mRNA

function. According to one bioinformatic study, 54% and 32% of genes in human and mouse, respectively, contain more than one polyadenylation site [Tian et al 2005]. The poly(A) tail is added at the poly(A) site (PAS) in the nucleus in a 2 two-step reaction consisting of a large cleavage complex that cleaves the pre-mRNA into two fragments followed by poly(A) tail addition to the upstream fragment [Zhao et al 1999]. Two main sequence motifs are important for cleavage/polyadenylation of mammalian mRNAs. The highly conserved and well-understood AAUAAA motif (called the poly(A) signal) is found 10-25nt upstream of the PAS. The second motif is found 10-30nts downstream of the PAS but is poorly understood due to its low conservation both in sequence and position. Although current bioinformatic approaches support the view that this motif is U/GU-rich [Salisbury et al 2006], they provide only a limited understanding of what motif(s) lies in this downstream region. First, the exact identity of this putative downstream motif for a given mammalian gene is often ambiguous and indeed it is a distinct possibility that there will be multiple motifs including auxiliary motifs. Second, in some cases where the predicted motif was examined by an extensive mutational analysis, the data supported the existence of additional motifs important for poly(A) site function [Chen et al 1998]. Thus the prediction of this downstream motif represents a type of problem suitable for analysis by iTriplet. To this end the downstream sequences of a set of genes was analyzed by iTriplet with the predicted motifs being indicated in Figure 4.4. According to a NMR structural study of the U/GU-rich binding protein CstF-64 [Perez et al 2003], we believe the binding site should not be longer than eight

nucleotides. Hence we applied a series of models ranging from 6 to 8 nucleotides long to nine genes of interest to us viz. U1A, SPR40, CDC7, DATF, LBP1, GAPDH, RAF, Mark1 and SmE. Results showed that model <8,2> yielded the best fit with the consensus TCTGATTT and this motif agrees with previous analysis performed by the Graber lab [Salisbury et al 2006] that the downstream region consists of a transition from UG-rich to U-rich in the 5' to 3' direction. MEME [Bailey et al 1994,1995] was used to process the same set of sequences with the resulting motif being BTRDGSCWSA that lacks such a transition.

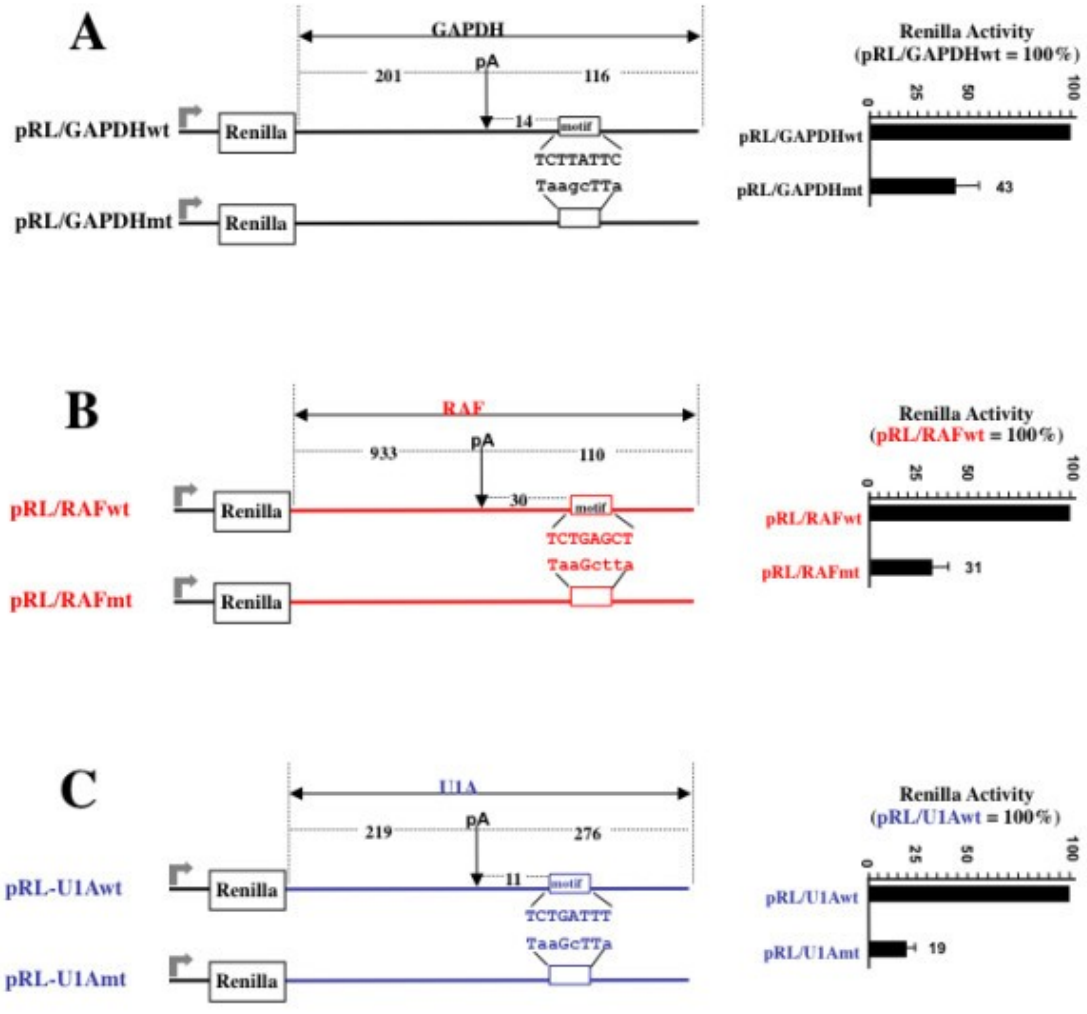


Figure 4.4 Confirmation of predicted poly(A) downstream elements by dual Luciferase reporter system. (A) pRL-GAPDHwt was made from a standard pRL-SV40 Renilla expression plasmid by replacing the SV40-derived 3'UTR and poly(A) signal sequences with the human GAPDH 3'UTR (NM_002046) and 116nt past the PAS. pRL-GAPDHmt matches pRL-GAPDHwt but having Motif A mutated as shown. Plasmids were transfected into HeLa cells and Luciferase activity measured 24 hours later. Values for Renilla Luciferase were normalized to those obtained from a co-transfected Firefly Luciferase plasmid. The pRL-GAPDHwt plasmid expresses 2.2 fold more Renilla than pRL-GAPDHmt plasmid thus Motif A is enhancing expression by 2.2 fold. (B) pRL-RAFwt (NM_002880) was made like pRL-GAPDHwt but from the human RAF gene sequences as indicated. pRL-RAFmt matches pRL-RAFwt but having Motif A mutated as shown. These plasmids were transfected and analyzed as in panel A. (C) pRL-U1Awt (NM_004596) was made like pRL-GAPDHwt but from the human U1A gene sequences as indicated. pRL-U1Amt matches pRL-U1Awt but having Motif A mutated as shown. These plasmids were transfected and analyzed as in panel A.

To test whether the TCTGATTT motifs identified by iTriplet were functional, the dual Luciferase reporter system was used where Renilla Luciferase mRNA contained the entire 3'UTR plus sequences past the PAS of the gene of interest. A co-transfected Firefly Luciferase reporter was included that serves as an internal normalization control (details can be found in Appendix

L). As diagrammed in Figure 4.4, the plasmid pRL-GAPDHwt was made from a standard pRL-SV40 Renilla expression plasmid by replacing the SV40-derived 3'UTR and downstream poly(A) signal sequences with the human GAPDH 3'UTR and poly(A) signal region (NM_002046) including 116nt past the poly(A) site. iTriplet predicted that GAPDH has a motif we call GAPDH Motif A that would potentially be important for poly(A) site activity. To determine if GAPDH Motif A is functional, we mutated it as shown to make plasmid pRL-GAPDHmt. Plasmids were transfected into HeLa cells and Luciferase activity was measured; values for Renilla Luciferase were normalized to those obtained from the co-transfected Firefly Luciferase control plasmid. The pRL-GAPDHmt plasmid expresses 43% less Renilla Luciferase than pRL-GAPDHwt, indicating Motif A enhances Renilla Luciferase expression by about 2.2-fold.

The same analysis was done in panels B and C but for the human RAF and human U1A genes, respectively. As can be seen the RAF Motif A enhances expression 3.2 fold and the U1A Motif A enhances expression by 5.1 fold. Here we have demonstrated the predictive power of iTriplet for these three genes however we do not exclude the existence of other binding sites that can also affect poly(A) activity of these genes.

D. *Conclusion*

We have presented a novel rule-based algorithm called iTriplet to solve the challenging degenerate and long motif finding problem that was unsolved before. In addition, we have confirmed our prediction for real biological signals

experimentally. The runtime of iTriplet is comparable to other well-known methods of the same design philosophy and is significantly better at analyzing longer motifs (>16 nucleotides). To our knowledge, iTriplet is the most parallelizable motif finding method in the family of guaranteed optimal motif finding algorithms developed so far. Furthermore we have shown that our method is very competitive in prediction accuracy when compared with other popular motif finding methods. Overall, our method has the superiority like other exact optimal motif finding methods to find the optimal motif in the absence of statistical overrepresentation and yet without sacrificing prediction accuracy. That said, no single method or approach is able to solve the general <*l,d*> motif problem completely in terms of guaranteed solution, speed, memory consumption and prediction accuracy. Thus, further research effort is needed to overcome various hurdles of this problem.

APPENDICES

Appendix A. Genomes, cDNAs, ESTs

Genomes

Human, chimpanzee, mouse, rat, and bovine genomes were downloaded from NCBI (ftp://ftp.ncbi.nih.gov/genomes/).

Platypus genome was downloaded from UCSC genome browser website (http://hgdownload.cse.ucsc.edu/downloads.html#platypus)

Human	March 2006 (hg18)
Chimpanzee	October 2006 (panTro2)
Mouse	July 2007 (mm9)
Rat	July 2006 (rn4)
Bovine	October 2007 (bosTau4)
Platypus	March 2007 (ornAna1)

cDNAs

All cDNAs were downloaded from RefSeq database in NCBI (ftp://ftp.ncbi.nih.gov/refseq/release/).

Human	June 2008
Chimpanzee	September 2006
Mouse	March 2008
Rat	August 2009

Bovine	June 2008

ESTs

EST sequences and their mapped genomic locations were downloaded from UCSC genome browser website (http://hgdownload.cse.ucsc.edu)

Appendix B. EST-based poly(A) sites construction

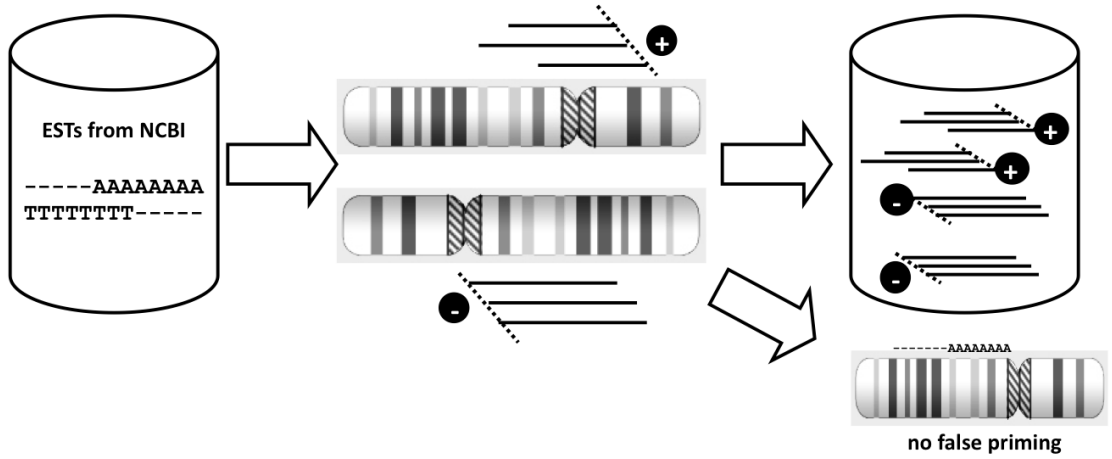


Figure B1 Workflow of EST-based poly(A) sites construction.

EST sequences were utilized to identify polyadenylation sites (PAS) in human and mouse genomes. Below steps are taken for this process:

1. Screening of EST sequences: only EST sequences ending with at least 6 A nucleotides (nts) or starting with 6 T nts were included.

2. False priming validation: poly(A/T)-ended EST sequences were mapped to the genomes. If the poly(A/T) nts is created by polyadenylation, no genomic poly(A/T) should be found at the 3'/5' of the genomes. Otherwise, the poly(A/T) of the ESTs were not really accounted by polyadenylation, so they were removed from the dataset.

3. Directionality of EST: filtered EST sequences were mapped to the genome in order to determine the direction of transcription according to the following conditions:

	Mapped to genome	Support transcript direction
Poly(A) ended EST	Plus strand	plus (5' to 3')
Poly(A) ended EST	Minus strand	minus (3' to 5')
Poly(T) started EST	Plus strand	minus (5' to 3')
Poly(T) started EST	Minus strand	plus (3' to 5')

4. PAS identification: filtered ESTs were separated into two groups by directionality, one supports plus strand transription, the other supports minus strand transcription. Within each group, ESTs were stacked up along the genome. If the polyadenylated ends of EST (at least 3) are found to terminate close to each other (±10 nts), then such a location is marked as a PAS.

In human, 899,786 out of 15 millions EST were found either have 6 or more A at the 3' end or T at the beginning. In mouse, 317,658 out of 8.5 millions EST were found.

By this procedure, 17,080 and 8,799 EST-supported PAS were found in human and mouse respectively.

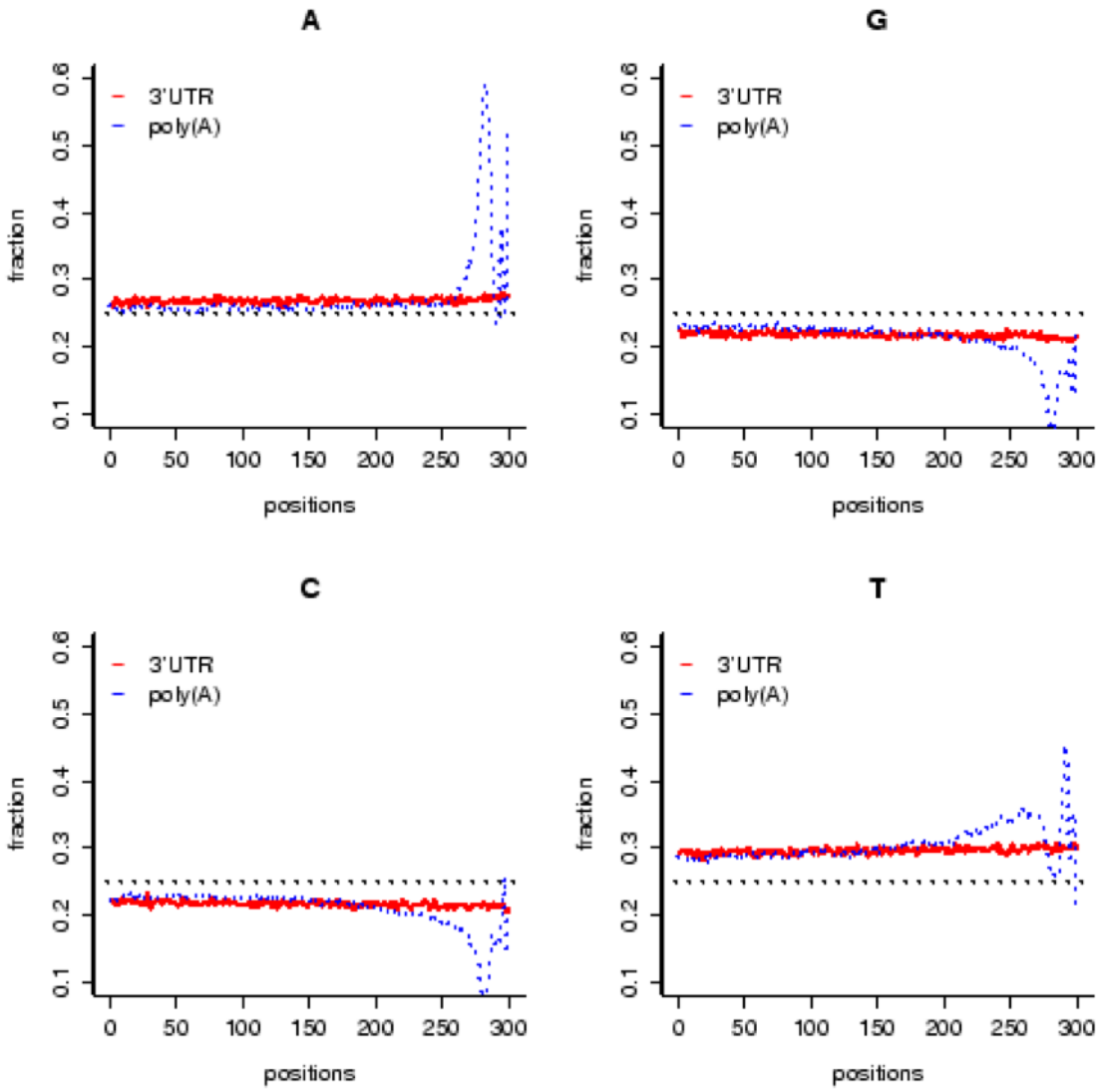
Appendix C. Pseudo PAS nucleotide composition

The most frequent poly(A) signals, i.e. AAUAAA and AUUAAA, were scanned in the human and mouse intronic regions. These sites are named pseudo PAS unless they are associated with EST-supported PAS mentioned above in section B. [-500,+500] regions of these pseudo PAS were extracted. Their nucleotide composition in human and mouse were analyzed. In human, it was found that A, C, G and T compositions were 31%, 19%, 20%, and 30% respectively. Similar composition was found in mouse as well viz. 30%, 20%, 20%, and 30% of A,C,G, and T respectively.

Appendix D. 3’ UTR nucleotide composition

17,080 and 8,799 300-nt long fragments were randomly sampled from the 3’ UTR of genes with EST-supported PAS in human and mouse, respectively. These random fragments were at least 50 nts downstream of the stop codon and 200 nts upstream from the PAS. Their nucleotide composition was directly compared with the 300-nt long upstream region of PAS position by position. Figure D.1 below illustrates the comparison separately by nucleotides:

Human


A
G
C
T
3'UTR
poly(A)
fraction
positions
0.1
0.2
0.3
0.4
0.5
0.6
0
50
100
150
200
250
300

A

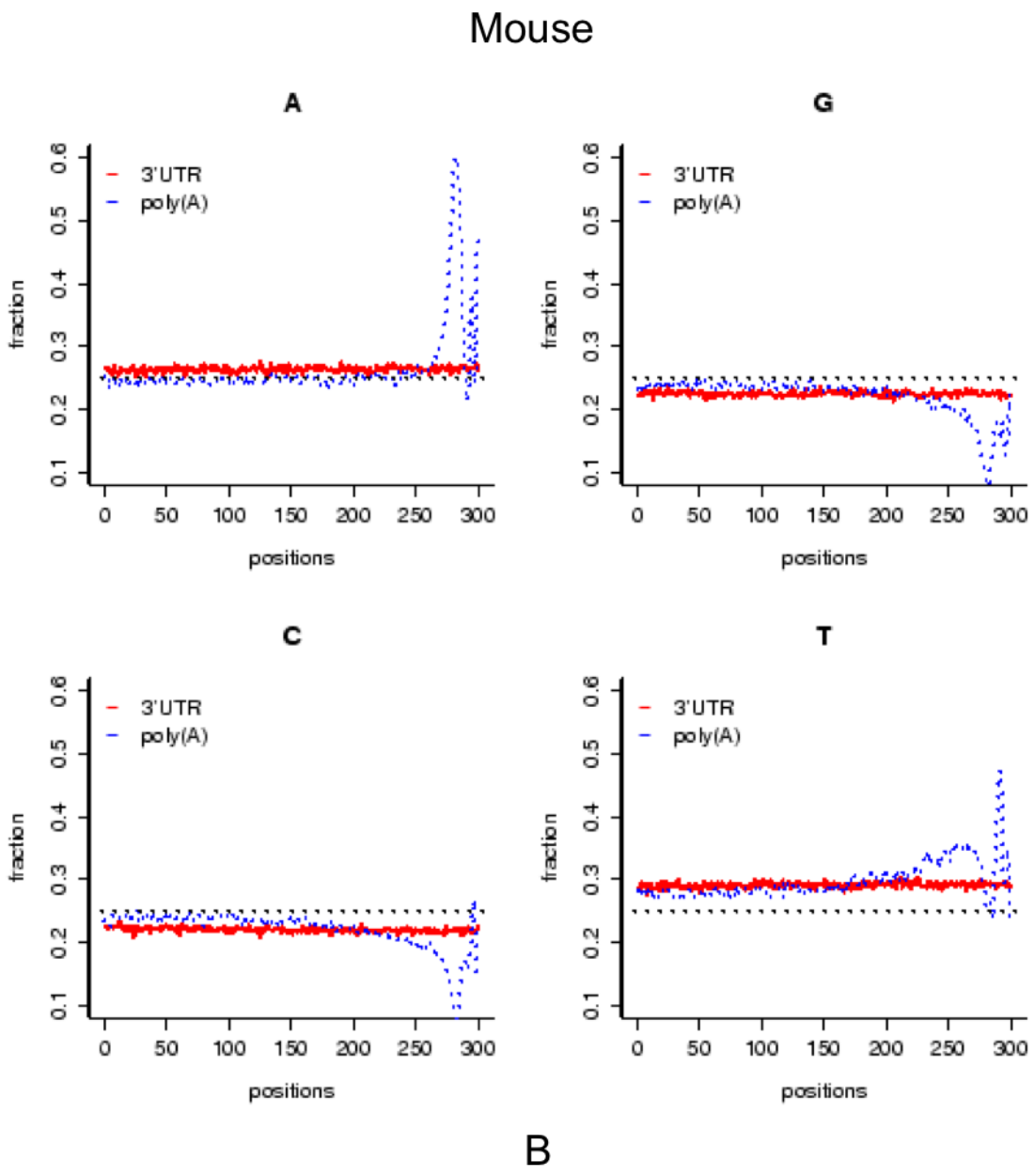


Figure D1 Nucleotide composition of A) human, and B) mouse 3' UTR.

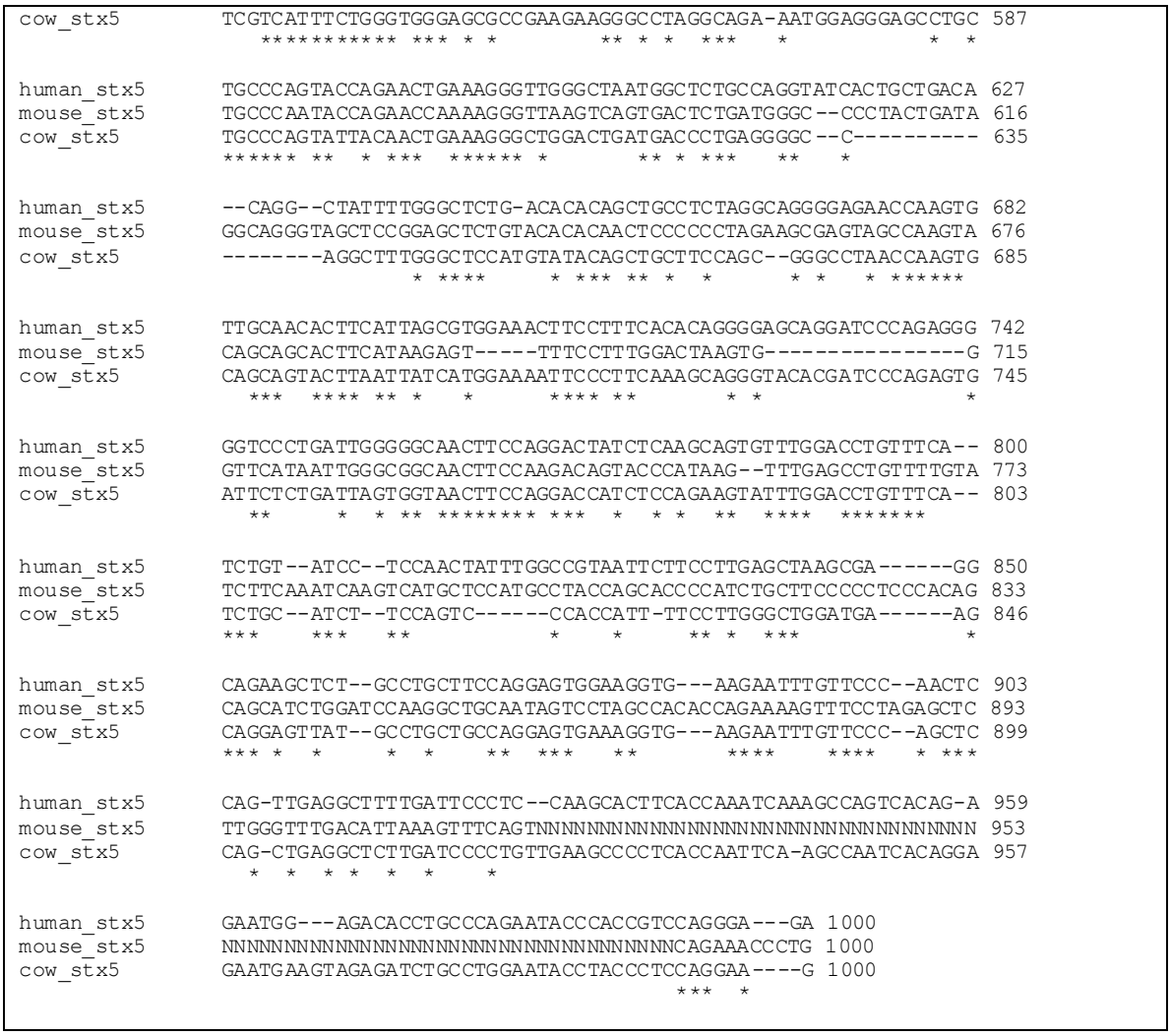
As shown above, 3' UTRs are evenly enriched in T, and then A (red).On the other hand, upstream regions of the PAS have a dramatic increase in A and T frequency at around 50 nts upstream of the PAS, which likely is the location of the poly(A) signal.

Appendix E. Conserved flanking region of PAS outliers

As discussed in chapter 3, in order to understand the conservation of PAS flanking region among genes that lack of the 16 hexamers (poly(A) signals), [-500,+500] regions were aligned among human, mouse and cow. Putative poly(A) signals are bold and underlined. They are mostly found at position 480 or 20 nts upstream of the PAS.

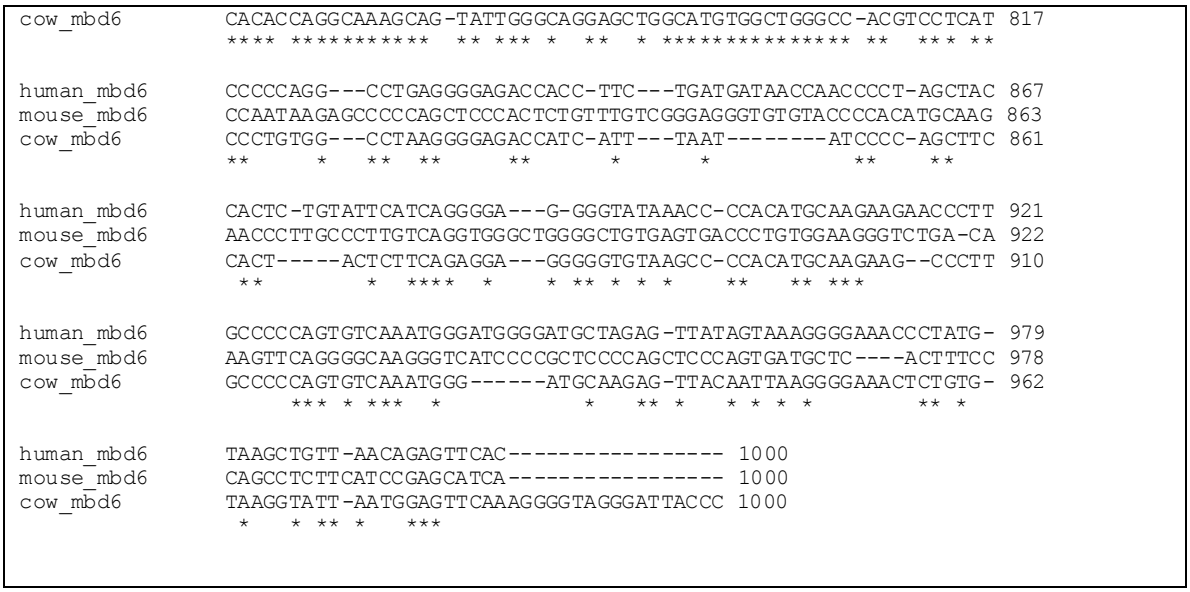
STX5:

```
human_stx5      ---CTGAGCCTGTGCAGGGTACTTGGGAGAAAGGCCCTGTTTCCCTGGAACTGCTAAGAA 57
mouse_stx5      CCACTGAGCCTGTGCAAGGTGGTTGGGAGAAAGGCC--ATTTCCTTGGAACTGCTAAGAA 58
cow_stx5        ----TGAGCCTGTGCAGGGTGCTTGGGAGAAGGGCCCCGTTTCCTTGGAACTGCTAAGAA 56
                    *********** ***  ********* ****   ***** ***************

human_stx5      TGACCACTGCCCCTGATCCCCCACCCCTTGCCTCTGGCCACCCTGTCCTCCCCCCACCAC 117
mouse_stx5      TGGCCAGTGTCCCTGATTCCCCAC-CCTTGTCTCTGGCCACTCTGTCCTATCCTCA---- 113
cow_stx5        TGACCACTGCCCTCGGTCCCCCACTCCTTGCCTCTGGCCACCCAGCCCTCCCCTCACCAC 116
                ** *** ** **  * * ****** ***** ********** * * ***  ** **

human_stx5      CCTCAGGCCTATGAAACACACAGGGTTCTAGATTTGAACTCTGCTGTGAAGTGACTGGAA 177
mouse_stx5      -----GGCCCATGAAACACACT-GGTTCTGGATTTGGACTCTGCTGTGAAGAGGCTGG-- 165
cow_stx5        CCTCCAGCCCATGAAACACACACGGTTCTGGATTTGGACTCTGCTGTGAAGTGGCTGGAA 176
                      *** ***********  ****** ****** ************** * ****

human_stx5      GGGAGCAGAGGCCAGCTG-GGGGCCAGTGGGGGAGGTTGTTTCCACTAGGAGATTTTTAT 236
mouse_stx5      --GAACAGAAGCCAGCTA-GGGGCCAGTGGGGGAGGTTGTCTCCACCTTGAGATTTTTAT 222
cow_stx5        AGGAGCAGAGGCCAACTGAGGGGCTCGTCGGGGAGATTGTCTCTACCAGGAGATTTTTAT 236
                  ** **** **** **  *****  ** ****** **** ** **   ***********

human_stx5      AAACC-CTCTCCAGCCTCTCCCAAAGGAAGCGTTGGCAGCAAAGGGAGATGATGCCCTTA 295
mouse_stx5      AAGCCTATAACCAGCCTCCCCCAAACGTAAC--TGGCAGCAAGAGGAGAAAACGCCCTTC 280
cow_stx5        AAACC-CTCCCCAGCCTGTCGCAAAGGGAACTTTGGCAGCAAGGGGAGATGATGCCTTTC 295
                ** **  *  *******  * **** * * *  *********  *****  * *** **

human_stx5      CCCACCTTCCTGTGAGTGAAGAGAGGAAG-----CAGCCCCAGGGACCAATTTTCCCAA- 349
mouse_stx5      C-----TTCTTGAGAGGCAAGAATCCTCAAG-AAGTATCCAAGGGACCAACTTCATCATC 334
cow_stx5        CCCACATTCTTGAGAGCAAGGAGAGGAAGTGAAACTGCCCCAGGGACCAACTTTCCCATC 355
                *     *** ** ***  * **                ** ********* **   **

human_stx5      ---------------TTGACCTCTTTCTTCCTCT-TTCACCATGTGAGGC-AGGGAGCCC 392
mouse_stx5      TGGGTCAGCTAGAACTTGACTGCCG-C--CTTCTCCTCACCATGTGAGGTGAGGGGCTC 391
cow_stx5        TGGGT----TAGAATTTGACCTCTT-CTTCCTCTCTTCACCACATGAGGC-AGGGGGCCC 409
                          *****  *   *  * ***  ******  *****  **** ** *

human_stx5      TGAGCCCTTCAGCTGCCTGCACAACCCCTGACATTGGCTGCTGGTGA--CTCAATCTGCC 450
mouse_stx5      TGAGCCCTACAGTTGCCTGCACAAACCT-GACTTTGGCTACTGGTGACTCTCAATATGCC 450
cow_stx5        TGAGCCCCTCGGCGGCCTGCACAACCCC-AACTTTGGCTGCAGATGACTCTCAATCTGCC 468
                *******  * *  ********** **   ** ****** * * ***  ****** ****

human_stx5      AAATGTGCTGCAGCTCGTTTTCTCCCA<u>**ATTACA**</u>GCAAGACTGTCAGCCTCACTAGCCATG 510
mouse_stx5      AAACATGCTGCAGCCTATTTCCTCCCA<u>**ATTACA**</u>GCAAGACTGTCAGCCTCACTAGTGTTG 510
cow_stx5        AAACGTGCTGCAGCCCGTTTTCTCCCA<u>**ATTACA**</u>GCAAGACTGTCAGCCTCACTAGCCATG 528
                ***  *********   *** ********************************   **

human_stx5      TCATCATTTCTGGGTGGGAGCGTCGAA---GGGCCTAGGCAGCGAGTGGAGAGAGCCCAC 567
mouse_stx5      ATGTCATTTCTGGGGGGGGGGG---GG---GGACTTCAGCAAA-AA-----GCTAACCAC 558
```

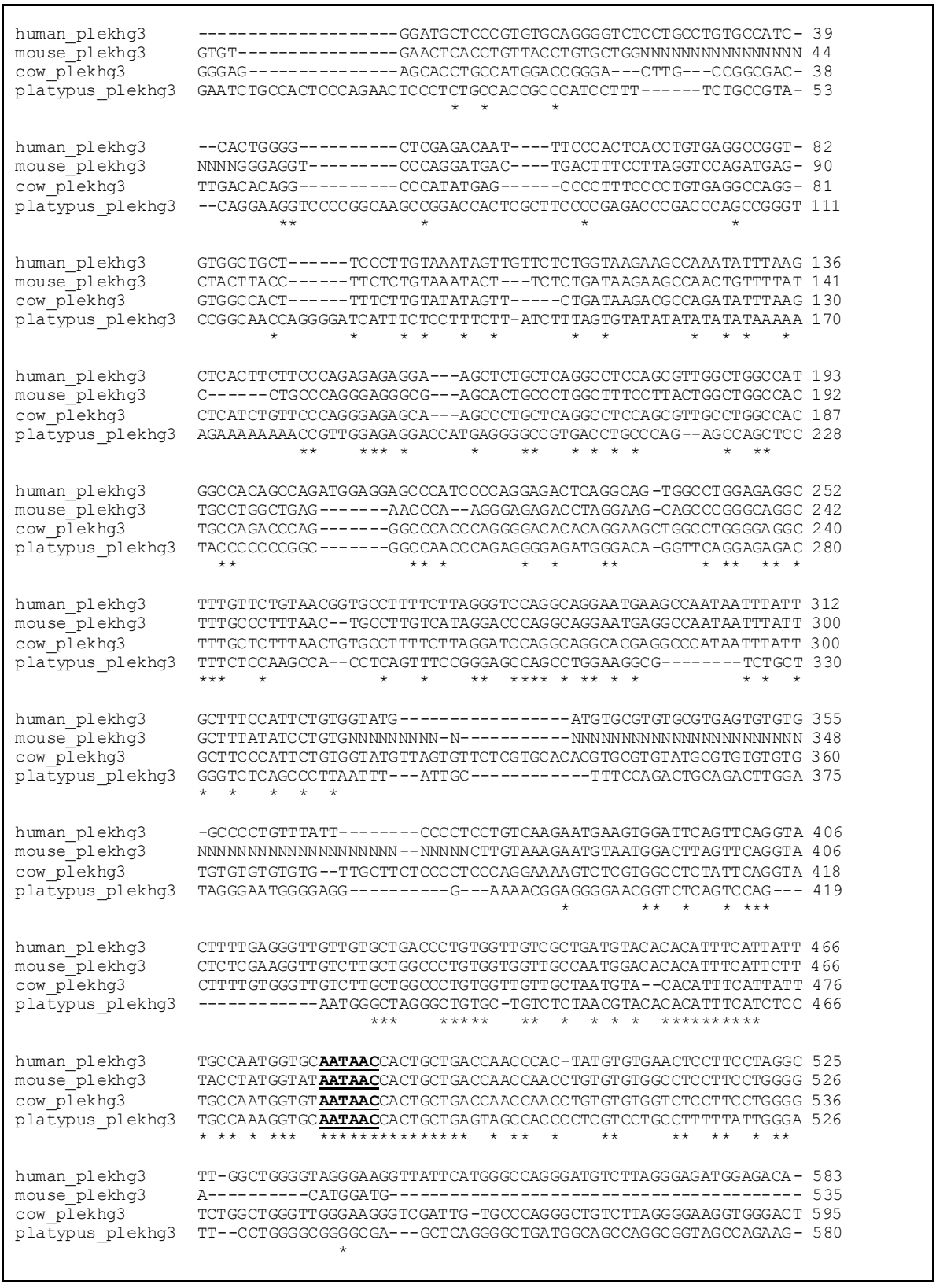

cow_stx5 TCGTCATTTCTGGGTGGGAGCGCCGAAGAAGGGCCTAGGCAGA-AATGGAGGGAGCCTGC 587
human_stx5 TGCCCAGTACCAGAACTGAAAGGGTTGGGCTAATGGCTCTGCCAGGTATCACTGCTGACA 627
mouse_stx5 TGCCCAATACCAGAACCAAAAGGGTTAAGTCAGTGACTCTGATGGGC--CCCTACTGATA 616
cow_stx5 TGCCCAGTATTACAACTGAAAGGGCTGGACTGATGACCCTGAGGGGC--C---------- 635
human_stx5 --CAGG--CTATTTTGGGCTCTG-ACACACAGCTGCCTCTAGGCAGGGGAGAACCAAGTG 682
mouse_stx5 GGCAGGGTAGCTCCGGAGCTCTGTACACACAACTCCCCCCTAGAAGCGAGTAGCCAAGTA 676
cow_stx5 --------AGGCTTTGGGCTCCATGTATACAGCTGCTTCCAGC--GGGCCTAACCAAGTG 685
human_stx5 TTGCAACACTTCATTAGCGTGGAAACTTCCTTTCACACAGGGGAGCAGGATCCCAGAGGG 742
mouse_stx5 CAGCAGCACTTCATAAGAGT-----TTTCCTTTGGACTAAGTG----------------G 715
cow_stx5 CAGCAGTACTTAATTATCATGGAAAATTCCCTTCAAAGCAGGGTACACGATCCCAGAGTG 745
human_stx5 GGTCCCTGATTGGGGGCAACTTCCAGGACTATCTCAAGCAGTGTTTGGACCTGTTTCA-- 800
mouse_stx5 GTTCATAATTGGGCGGCAACTTCCAAGACAGTACCCATAAG--TTTGAGCCTGTTTTGTA 773
cow_stx5 ATTCTCTGATTAGTGGTAACTTCCAGGACCATCTCCAGAAGTATTTGGACCTGTTTCA-- 803
human_stx5 TCTGT--ATCC--TCCAACTATTTGGCCGTAATTCTTCCTTGAGCTAAGCGA------GG 850
mouse_stx5 TCTTCAAATCAAGTCATGCTCCATGCCTACCAGCACCCCATCTGCTTCCCCCTCCCACAG 833
cow_stx5 TCTGC--ATCT--TCCAGTC------CCACCATT-TTCCTTGGGCTGGATGA------AG 846
human_stx5 CAGAAGCTCT--GCCTGCTTCCAGGAGTGGAAGGTG---AAGAATTTGTTCCC--AACTC 903
mouse_stx5 CAGCATCTGGATCCAAGGCTGCAATAGTCCTAGCCACACCAGAAAAGTTTCCTAGAGCTC 893
cow_stx5 CAGGAGTTAT--GCCTGCTGCCAGGAGTGAAAGGTG---AAGAATTTGTTCCC--AGCTC 899
human_stx5 CAG-TTGAGGCTTTTGATTCCCTC--CAAGCACTTCACCAAATCAAAGCCAGTCACAG-A 959
mouse_stx5 TTGGGTTTGACATTAAAGTTTCAGTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN 953
cow_stx5 CAG-CTGAGGCTCTTGATCCCCTGTTGAAGCCCCTCACCAATTCA-AGCCAATCACAGGA 957
human_stx5 GAATGG---AGACACCTGCCCAGAATACCCACCGTCCAGGGA---GA 1000
mouse_stx5 NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNCAGAAACCCTG 1000
cow_stx5 GAATGAAGTAGAGATCTGCCTGGAATACCTACCCTCCAGGAA----G 1000

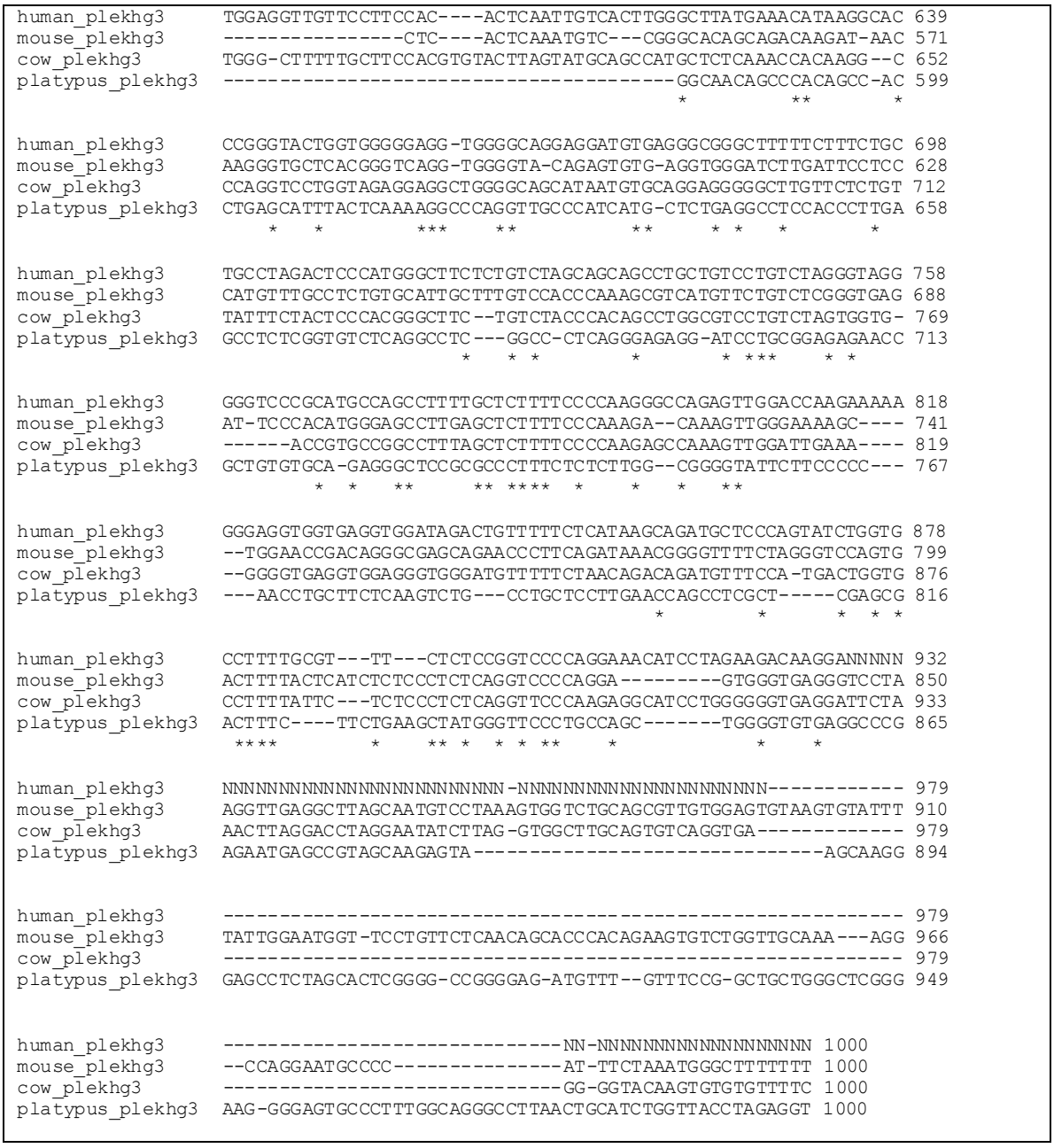
MBD6:

```
human_mbd6      ----------------GGCTGCTGCCCTCCTTCCCAGTGAAAGGTACAAAGCAATAAGC 43
mouse_mbd6      --CTGTGAGCTACTGTTGGCTGCTGCCCTCCTTCCCAGTGAAAGGTACAGAGCAATAAGC 58
cow_mbd6        ACCTGTGAGCTACTGTTGGCTGCTGCCCTCCTTCCCAATGAAAGGTACAAAGCAATAAGC 60
                                 ******************** ********** **********

human_mbd6      ATCATGCATCCTCCCCTTACCCCT-CCAACACCCCTCTGCCTCTGGCTCAGGTTGCTCAA 102
mouse_mbd6      ATCATGCATCTCCCCGATTCCC-A-CTAGCACCCCTCTGCCTCTGGCTCAGGTTGCTCAA 116
cow_mbd6        ATCATGCGTCCTCCCCTCACCCCTGCCAACACCCCTCTGCCTCTGGCTCAGGTTGCTCAA 120
                ******* **  ***    ***   * * ******************************

human_mbd6      AGCACAGATCCT-CTCTTACCCCGTCCCCAGGTTTGAAACACATAGCCTCATTTCAAGGT 161
mouse_mbd6      AGCACAGATCCCCTCTTATCCCTGTCCCCAGGACCGAAACACATAGCCTCATTTCAAAGC 176
cow_mbd6        AGCACAGATCCCCCTCTTACCCTGTCCCCAGGTTTGAAACACATAGCCTCATTTCAAGGT 180
                ********** *  *** *********   ********************** *

human_mbd6      GTAGCCAGGTTCCCCCGACTTTCCTCTGGGATATAAAAAAGGGGGTAAGGGGGCAAAGAG 221
mouse_mbd6      GTAGCCAGGTTCCCTCTGCTTTCCTCTGGGATATGGA-AAGGGGGCCAAGAGG------- 228
cow_mbd6        GTAGCCAGCTTCCCTCCACTTTCCTCTGGGATATGAA-GAGGGGGTAAGGGGGCAAAGAG 239
                ******** ***** *  *************** *  ******  * * **

human_mbd6      AGCCCTCTGGGCCTCTCCTCCCATACACACTACACTGCCCCTTCTCCCCCCATCAAAACG 281
mouse_mbd6      ---ACTTTGGATCTCTCCT-CCATACACACTACACTGGCCCTTCTCCCCA--T-CAAGCG 281
cow_mbd6        AGCCCTCTGGGTGTCTCCTCCCATACACACTACACTGCCCCTTCTCCCCC-ATCAAAACG 298
                    ** ***   ****** ***************** ***********   *  ** **

human_mbd6      CTCAGAGACGTTGTGATGATGCGACTGAGGATTATGCAACGTGGTCCAACCGGAGCGGCC 341
mouse_mbd6      CTCAGAGACGGTGTGACGATGCCACTGGGGGTTATGCAGCGTGGTCCAGCCGGAGCGGCC 341
cow_mbd6        CTCAGAGACGTTGTGACGATGCGACTGAGGATTATGCAACATGGTCCAACCGGAGCGGCC 358
                ********** ***** ***** **** ** ******* * ******* ***********

human_mbd6      AGCATGACCAGCTGTCCAGGGGCTGCCTCCTGCCTTTTCTTTTGTAAAGACAAGACCCTT 401
mouse_mbd6      AGCATGACCAGCTGTCCAGGGGCTGCCTCCTGCCTTTTTTTTTGTAAAGACAAGACCCTT 401
cow_mbd6        AGCATGACCAGCTGTCCAGGGGCTGCCTCCTGCCTTTTCTTTTGTAAAGACAAGACCCTT 418
                ************************************ ********************

human_mbd6      GGGAGTTTTAATTCTGTTTTGTACTTGCCCTGTGGGGCCTCCACTGCTTTTCTATGGGAG 461
mouse_mbd6      GGGAGTTTTCATTCTGTTTTGTACTTGCCCTGTGGGGCCTCCACTGCTTTTCTATGGGAG 461
cow_mbd6        GGGAGTTTTAATTCTGTTTTGTACTTGCCCTGTGGGGCCTCCACTGCTTTTCTATGGGAG 478
                ********* ************************************************

human_mbd6      ACACTCTTAATTTAACAGATGAGAATATTTTGAAACTCTGGCTCTGGCTCTGTACTCATT 521
mouse_mbd6      ACACTCTTAATTTAACAGATGAGAATATTTTGAAACTCTGGTTCTGACTCTGTACTCATT 521
cow_mbd6        ACACTCTTAATTTAACAGATGAGAATATTTTGAAACTCTGGCTCTGACTCTGTCCTCATT 538
                ************************************* **** ****** ******

human_mbd6      TTTT-ATTTAGTTCTTTGGTAAGAACAGGTTACAATTTAAATCCATCTCTTGTAGTATAG 580
mouse_mbd6      TTTTTATTTAGCTCTTCGGTAAGAACAGATTACAGCTGAAAT-AGTC--TCAT--AATAG 576
cow_mbd6        TTTT-CCTTAGTTCTTTTGTAAGAACAGATTATAATTTAAATCCTTC---TGTGGTAGAG 594
                ****   **** ****  ********** *** *  * ****   **     *   * **

human_mbd6      AGTGGCTTAGATTGCCTGTTATGACGAATGAATATCTATATCCTAGTGCTGCTTCCTCCC 640
mouse_mbd6      TGTGGCTTAGACTGTT----A--ACG-TTATGCACG--TCTCTCAGTGCTGCCTCTCCGA 627
cow_mbd6        AGT-GCTTAGACTGTC-------ACG-CTGTGTATGTCTCTCTCAGTGCTGCCTCTTCCC 645
                 ** ******* **         ***  *    *    * **  ******** **  *

human_mbd6      ---CAGGAAACACAGCAGAGGCCACACAGAGTACAACAGCATTTAATGGTCAGAAACAGT 697
mouse_mbd6      AATCCAGGAACACAGCAGAGACTC--CACGGTACAACAGCATTTAATGGTCAGAGACAGT 685
cow_mbd6        ---CAGGAAA--CAGCAGGGGCCAGACAGAGTACAACAGCATTTAATGGTCAGAAACAGT 700
                   *  * **  ****** * *    **  ********************** *****

human_mbd6      TGTACAGTATTACAGTCAGCCACAGAAGCTGTGTTGGGGGACAAGACCCAAT-CCTTCCC 756
mouse_mbd6      TGTACAGTGTTAGACTCGGCCAGAGACAGAC-GTTAGAGGACGGGATCCAGTCCCTCTCC 744
cow_mbd6        TGTACAGTATTACAGTCAGCCACAGAAGCTGTGCTGGAGGACAGGACCCAAT-CCTCCCC 759
                ******** *** * ** **** ***      * * * ****  ** *** * ***  **

human_mbd6      CACACCAGGCAAAGCAG-TATTGGACATGAGTTGGCATGTGGCTGGGCCCACGTCCTTAT 815
mouse_mbd6      CACAGCAGGCAAAGCACTTACTGGACGGGAACT-GCATGTGGCTGGGCCCACAACCTCAT 803
```

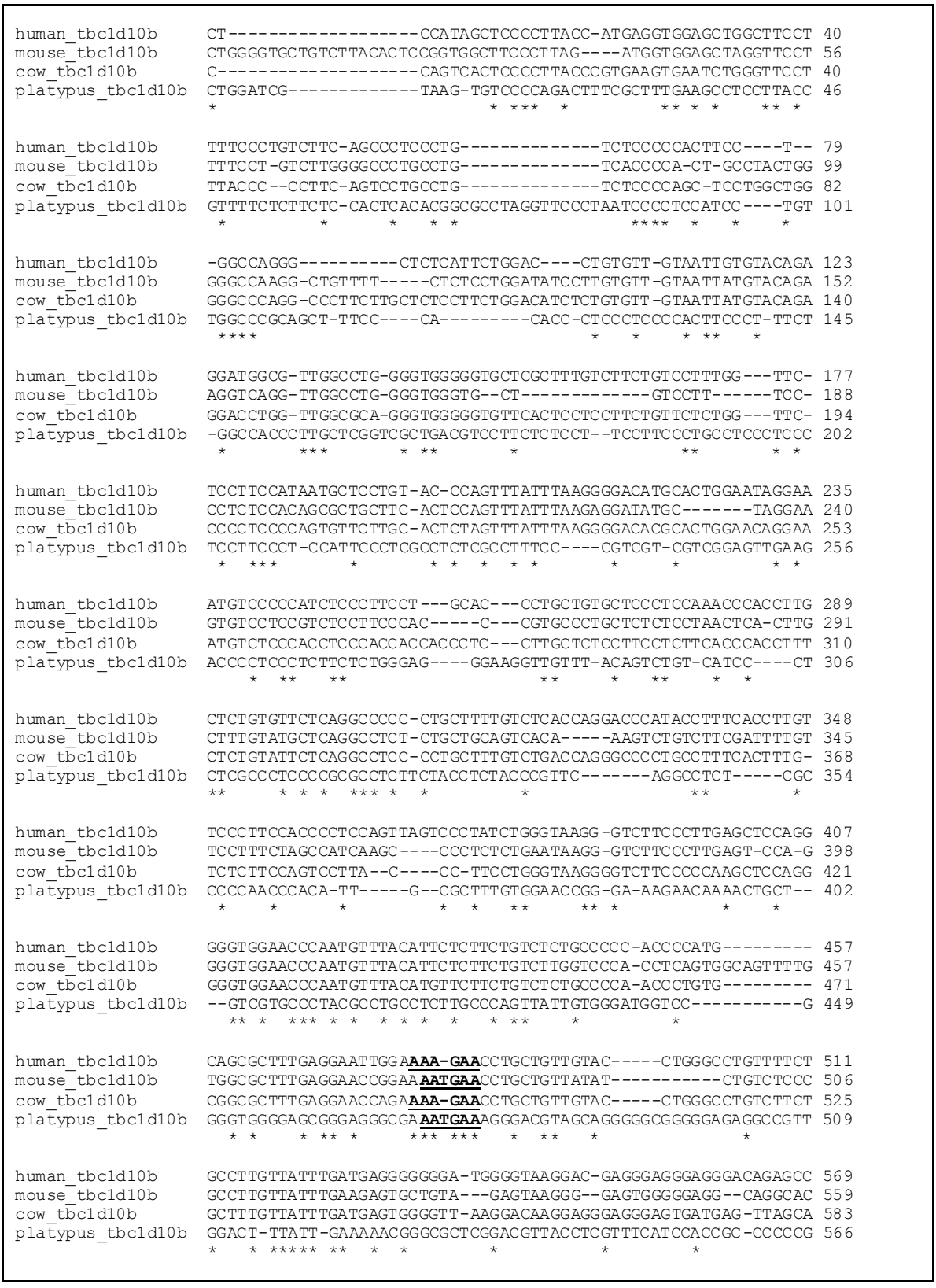

cow_mbd6 CACACCAGGCAAAGCAG-TATTGGGCAGGAGCTGGCATGTGGCTGGGCC-ACGTCCTCAT 817
**** ********** ** *** * ** * *************** ** *** **
human_mbd6 CCCCCAGG---CCTGAGGGGAGACCACC-TTC---TGATGATAACCAACCCCT-AGCTAC 867
mouse_mbd6 CCAATAAGAGCCCCCAGCTCCCACTCTGTTTGTCGGGAGGGTGTGTACCCCACATGCAAG 863
cow_mbd6 CCCTGTGG---CCTAAGGGGAGACCATC-ATT---TAAT--------ATCCCC-AGCTTC 861
** * ** ** ** * * ** **
human_mbd6 CACTC-TGTATTCATCAGGGGA---G-GGGTATAAACC-CCACATGCAAGAAGAACCCTT 921
mouse_mbd6 AACCCTTGCCCTTGTCAGGTGGGCTGGGGCTGTGAGTGACCCTGTGGAAGGGTCTGA-CA 922
cow_mbd6 CACT-----ACTCTTCAGAGGA---GGGGGTGTAAGCC-CCACATGCAAGAAG--CCCTT 910
** * **** * * ** * * * ** ** ***
human_mbd6 GCCCCCAGTGTCAAATGGGATGGGGATGCTAGAG-TTATAGTAAAGGGGAAACCCTATG- 979
mouse_mbd6 AAGTTCAGGGGCAAGGGTCATCCCCGCTCCCCAGCTCCCAGTGATGCTC----ACTTTCC 978
cow_mbd6 GCCCCCAGTGTCAAATGGG------ATGCAAGAG-TTACAATTAAGGGGAAACTCTGTG- 962
*** * *** * * ** * * * * * ** *
human_mbd6 TAAGCTGTT-AACAGAGTTCAC---------- 1000
mouse_mbd6 CAGCCTCTTCATCCGAGCATCA---------- 1000
cow_mbd6 TAAGGTATT-AATGGAGTTCAAAGGGGTAGGGATTACCC 1000
* * ** * ***

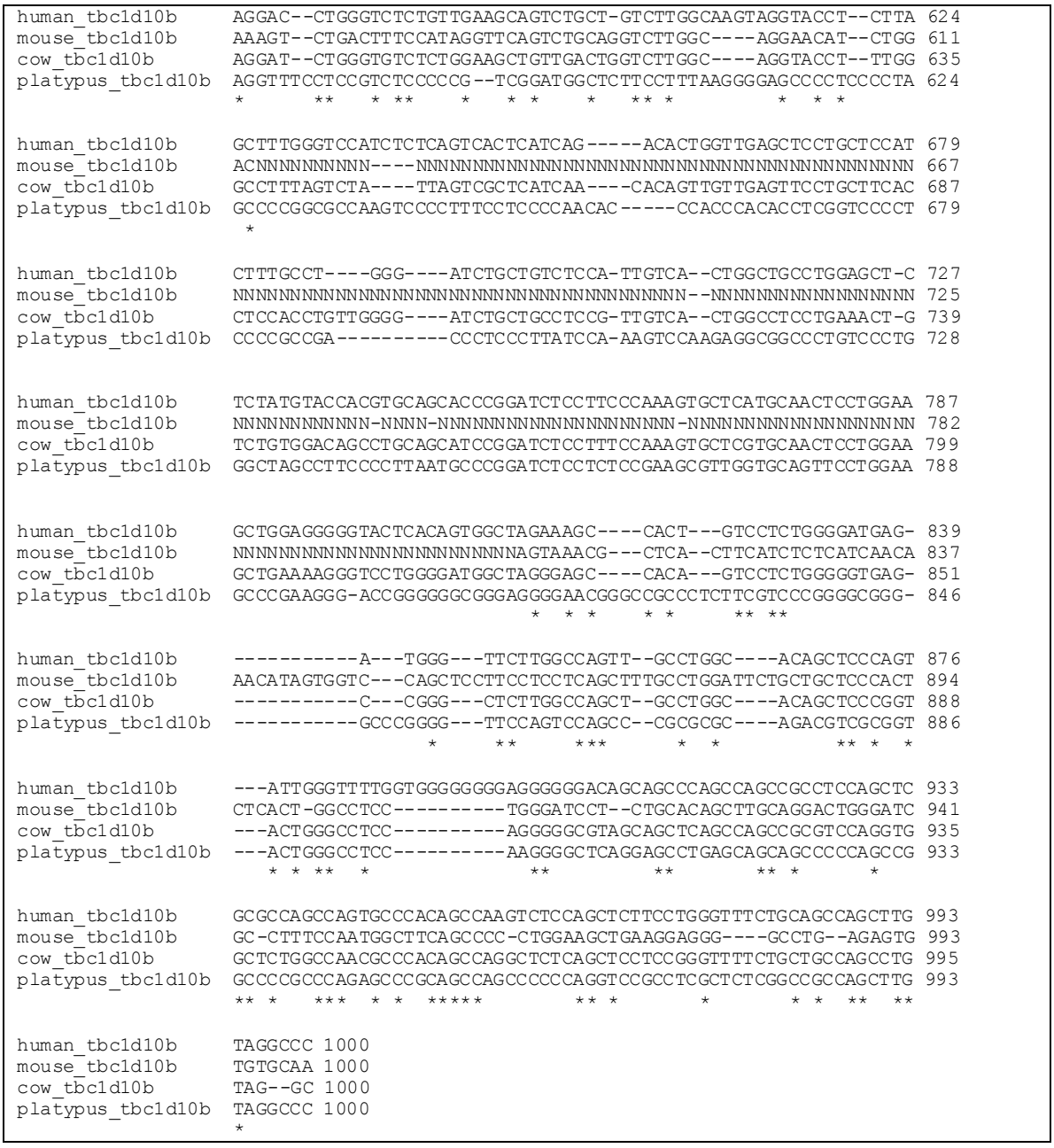
PLEKHG3:


human_plekhg3 ----------GGATGCTCCCGTGTGCAGGGGTCTCCTGCCTGTGCCATC- 39
mouse_plekhg3 GTGT----------GAACTCACCTGTTACCTGTGCTGGNNNNNNNNNNNNNNN 44
cow_plekhg3 GGGAG----------AGCACCTGCCATGGACCGGGA---CTTG---CCGGCGAC- 38
platypus_plekhg3 GAATCTGCCACTCCCAGAACTCCCTCTGCCACCGCCCATCCTTT------TCTGCCGTA- 53
human_plekhg3 --CACTGGGG----------CTCGAGACAAT----TTCCCACTCACCTGTGAGGCCGGT- 82
mouse_plekhg3 NNNNGGGAGGT---------CCCAGGATGAC----TGACTTTCCTTAGGTCCAGATGAG- 90
cow_plekhg3 TTGACACAGG----------CCCATATGAG------CCCCTTTCCCCTGTGAGGCCAGG- 81
platypus_plekhg3 --CAGGAAGGTCCCCGGCAAGCCGGACCACTCGCTTCCCCGAGACCCGACCCAGCCGGGT 111
human_plekhg3 GTGGCTGCT------TCCCTTGTAAATAGTTGTTCTCTGGTAAGAAGCCAAATATTTAAG 136
mouse_plekhg3 CTACTTACC------TTCTCTGTAAATACT---TCTCTGATAAGAAGCCAACTGTTTTAT 141
cow_plekhg3 GTGGCCACT------TTTCTTGTATATAGTT-----CTGATAAGACGCCAGATATTTAAG 130
platypus_plekhg3 CCGGCAACCAGGGGATCATTTCTCCTTTCTT-ATCTTTAGTGTATATATATATATAAAAA 170
human_plekhg3 CTCACTTCTTCCCAGAGAGAGGA---AGCTCTGCTCAGGCCTCCAGCGTTGGCTGGCCAT 193
mouse_plekhg3 C------CTGCCCAGGGAGGGCG---AGCACTGCCCTGGCTTTCCTTACTGGCTGGCCAC 192
cow_plekhg3 CTCATCTGTTCCCAGGGAGAGCA---AGCCCTGCTCAGGCCTCCAGCGTTGCCTGGCCAC 187
platypus_plekhg3 AGAAAAAAAACCGTTGGAGAGGACCATGAGGGGCCGTGACCTGCCCAG--AGCCAGCTCC 228
human_plekhg3 GGCCACAGCCAGATGGAGGAGCCCATCCCCAGGAGACTCAGGCAG-TGGCCTGGAGAGGC 252
mouse_plekhg3 TGCCTGGCTGAG-------AACCCA--AGGGAGAGACCTAGGAAG-CAGCCCGGGCAGGC 242
cow_plekhg3 TGCCAGACCCAG-------GGCCCACCCAGGGGACACACAGGAAGCTGGCCTGGGGAGGC 240
platypus_plekhg3 TACCCCCCCGGC-------GGCCAACCCAGAGGGGAGATGGGACA-GGTTCAGGAGAGAC 280
human_plekhg3 TTTGTTCTGTAACGGTGCCTTTTCTTAGGGTCCAGGCAGGAATGAAGCCAATAATTTATT 312
mouse_plekhg3 TTTGCCCTTTAAC--TGCCTTGTCATAGGACCCAGGCAGGAATGAGGCCAATAATTTATT 300
cow_plekhg3 TTTGCTCTTTAACTGTGCCTTTTCTTAGGATCCAGGCAGGCACGAGGCCCATAATTTATT 300
platypus_plekhg3 TTTCTCCAAGCCA--CCTCAGTTTCCGGGAGCCAGCCTGGAAGGCG--------TCTGCT 330
human_plekhg3 GCTTTCCATTCTGTGGTATG----------ATGTGCGTGTGCGTGAGTGTGTG 355
mouse_plekhg3 GCTTTATATCCTGTGNNNNNNNNN-N----------NNNNNNNNNNNNNNNNNNNNN 348
cow_plekhg3 GCTTCCCATTCTGTGGTATGTTAGTGTTCTCGTGCACACGTGCGTGTATGCGTGTGTGTG 360
platypus_plekhg3 GGGTCTCAGCCCTTAATTT---ATTGC----------TTTCCAGACTGCAGACTTGGA 375
human_plekhg3 -GCCCCTGTTTATT--------CCCCTCCTGTCAAGAATGAAGTGGATTCAGTTCAGGTA 406
mouse_plekhg3 NNNNNNNNNNNNNNNNNNN--NNNNNCTTGTAAAGAATGTAATGGACTTAGTTCAGGTA 406
cow_plekhg3 TGTGTGTGTGTG--TTGCTTCTCCCCTCCCAGGAAAAGTCTCGTGGCCTCTATTCAGGTA 418
platypus_plekhg3 TAGGGAATGGGGAGG----------G---AAAACGGAGGGGAACGGTCTCAGTCCAG--- 419
human_plekhg3 CTTTTGAGGGTTGTTGTGCTGACCCTGTGGTTGTCGCTGATGTACACACATTTCATTATT 466
mouse_plekhg3 CTCTCGAAGGTTGTCTTGCTGGCCCTGTGGTGGTTGCCAATGGACACACATTTCATTCTT 466
cow_plekhg3 CTTTTGTGGGTTGTCTTGCTGGCCCTGTGGTTGTTGCTAATGTA--CACATTTCATTATT 476
platypus_plekhg3 ------------AATGGGCTAGGGCTGTGC-TGTCTCTAACGTACACACATTTCATCTCC 466
human_plekhg3 TGCCAATGGTGCAATAACCACTGCTGACCAACCCAC-TATGTGTGAACTCCTTCCTAGGC 525
mouse_plekhg3 TACCTATGGTATAATAACCACTGCTGACCAACCAACCTGTGTGTGGCCTCCTTCCTGGGG 526
cow_plekhg3 TGCCAATGGTGTAATAACCACTGCTGACCAACCAACCTGTGTGTGGTCTCCTTCCTGGGG 536
platypus_plekhg3 TGCCAAAGGTGCAATAACCACTGCTGAGTAGCCACCCCTCGTCCTGCCTTTTTATTGGGA 526
human_plekhg3 TT-GGCTGGGGTAGGGAAGGTTATTCATGGGCCAGGGATGTCTTAGGGAGATGGAGACA- 583
mouse_plekhg3 A----------CATGGATG---------- 535
cow_plekhg3 TCTGGCTGGGTTGGGAAGGGTCGATTG-TGCCCAGGGCTGTCTTAGGGGAAGGTGGGACT 595
platypus_plekhg3 TT--CCTGGGGCGGGGCGA---GCTCAGGGGCTGATGGCAGCCAGGCGGTAGCCAGAAG- 580

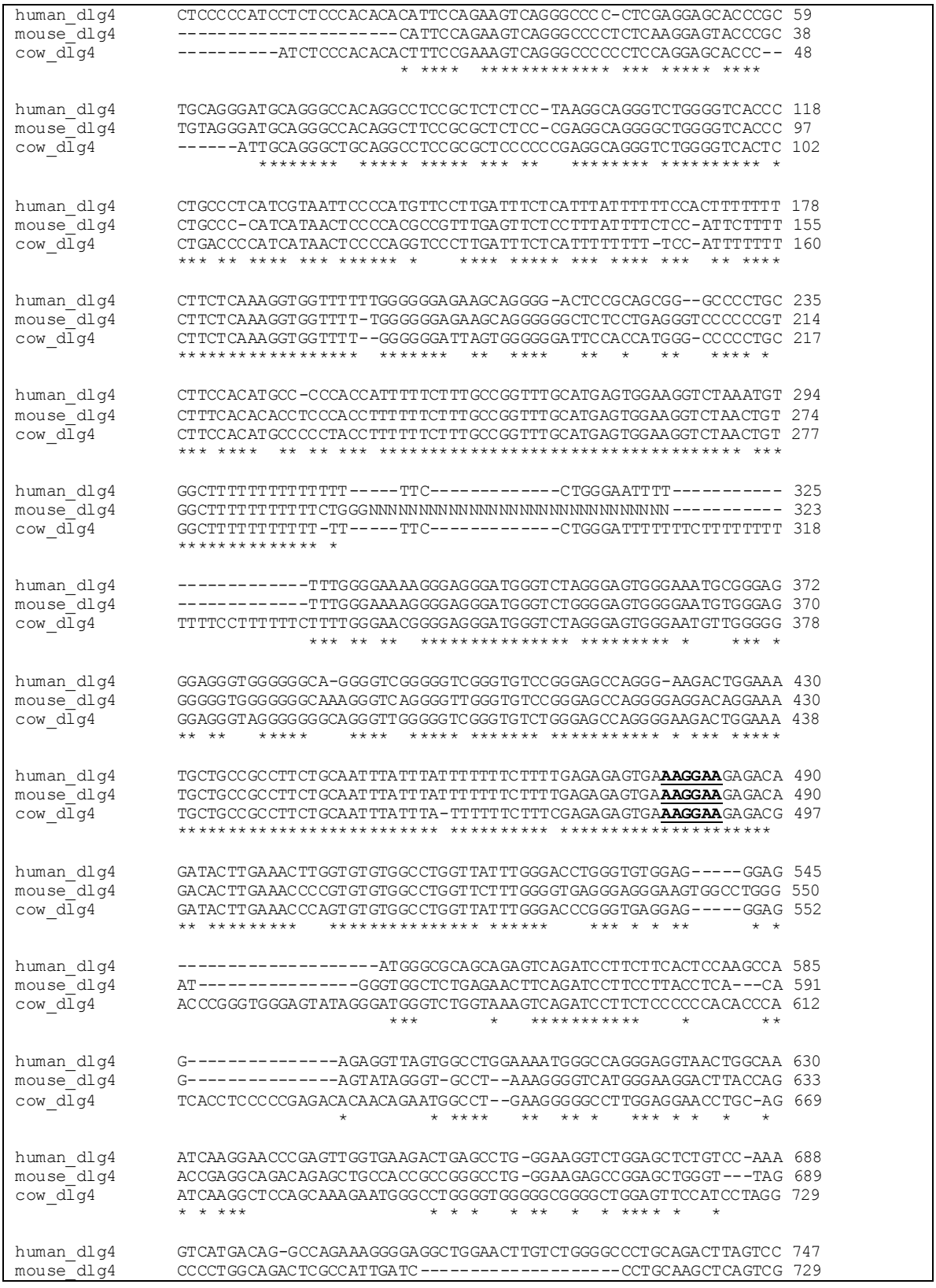

human_plekhg3 TGGAGGTTGTTCCTTCCAC----ACTCAATTGTCACTTGGGCTTATGAAACATAAGGCAC 639
mouse_plekhg3 ----------------CTC----ACTCAAATGTC---CGGGCACAGCAGACAAGAT-AAC 571
cow_plekhg3 TGGG-CTTTTTGCTTCCACGTGTACTTAGTATGCAGCCATGCTCTCAAACCACAAGG--C 652
platypus_plekhg3 ----------GGCAACAGCCCACAGCC-AC 599
human_plekhg3 CCGGGTACTGGTGGGGGAGG-TGGGGCAGGAGGATGTGAGGGCGGGCTTTTTCTTTCTGC 698
mouse_plekhg3 AAGGGTGCTCACGGGTCAGG-TGGGGTA-CAGAGTGTG-AGGTGGGATCTTGATTCCTCC 628
cow_plekhg3 CCAGGTCCTGGTAGAGGAGGCTGGGGCAGCATAATGTGCAGGAGGGGGCTTGTTCTCTGT 712
platypus_plekhg3 CTGAGCATTTACTCAAAAGGCCCAGGTTGCCCATCATG-CTCTGAGGCCTCCACCCTTGA 658
human_plekhg3 TGCCTAGACTCCCATGGGCTTCTCTGTCTAGCAGCAGCCTGCTGTCCTGTCTAGGGTAGG 758
mouse_plekhg3 CATGTTTGCCTCTGTGCATTGCTTTGTCCACCCAAAGCGTCATGTTCTGTCTCGGGTGAG 688
cow_plekhg3 TATTTCTACTCCCACGGGCTTC--TGTCTACCCACAGCCTGGCGTCCTGTCTAGTGGTG- 769
platypus_plekhg3 GCCTCTCGGTGTCTCAGGCCTC---GGCC-CTCAGGGAGAGG-ATCCTGCGGAGAGAACC 713
human_plekhg3 GGGTCCCGCATGCCAGCCTTTTGCTCTTTTCCCCAAGGGCCAGAGTTGGACCAAGAAAAA 818
mouse_plekhg3 AT-TCCCACATGGGAGCCTTGAGCTCTTTTCCCAAAGA--CAAAGTTGGGAAAAGC---- 741
cow_plekhg3 ------ACCGTGCCGGCCTTTAGCTCTTTTCCCCAAGAGCCAAAGTTGGATTGAAA---- 819
platypus_plekhg3 GCTGTGTGCA-GAGGGCTCCGCGCCCTTTCTCTCTTGG--CGGGGTATTCTTCCCCC--- 767
human_plekhg3 GGGAGGTGGTGAGGTGGATAGACTGTTTTTCTCATAAGCAGATGCTCCCAGTATCTGGTG 878
mouse_plekhg3 --TGGAACCGACAGGGCGAGCAGAACCCTTCAGATAAACGGGGTTTTCTAGGGTCCAGTG 799
cow_plekhg3 --GGGGTGAGGTGGAGGGTGGGATGTTTTTCTAACAGACAGATGTTTCCA-TGACTGGTG 876
platypus_plekhg3 ---AACCTGCTTCTCAAGTCTG---CCTGCTCCTTGAACCAGCCTCGCT-----CGAGCG 816
human_plekhg3 CCTTTTGCGT---TT---CTCTCCGGTCCCCAGGAAACATCCTAGAAGACAAGGANNNNN 932
mouse_plekhg3 ACTTTTACTCATCTCTCCCTCTCAGGTCCCCAGGA---------GTGGGTGAGGGTCCTA 850
cow_plekhg3 CCTTTTATTC---TCTCCCTCTCAGGTTCCCAAGAGGCATCCTGGGGGGTGAGGATTCTA 933
platypus_plekhg3 ACTTTC----TTCTGAAGCTATGGGTTCCCTGCCAGC-------TGGGGTGTGAGGCCCG 865
human_plekhg3 NNNNNNNNNNNNNNNNNNNNNNN-NNNNNNNNNNNNNNNNNNNNN---------- 979
mouse_plekhg3 AGGTTGAGGCTTAGCAATGTCCTAAAGTGGTCTGCAGCGTTGTGGAGTGTAAGTGTATTT 910
cow_plekhg3 AACTTAGGACCTAGGAATATCTTAG-GTGGCTTGCAGTGTCAGGTGA---------- 979
platypus_plekhg3 AGAATGAGCCGTAGCAAGAGTA----------AGCAAGG 894
human_plekhg3 ---------- 979
mouse_plekhg3 TATTGGAATGGT-TCCTGTTCTCAACAGCACCCACAGAAGTGTCTGGTTGCAAA---AGG 966
cow_plekhg3 ---------- 979
platypus_plekhg3 GAGCCTCTAGCACTCGGGG-CCGGGGAG-ATGTTT--GTTTCCG-GCTGCTGGGCTCGGG 949
human_plekhg3 ----------NN-NNNNNNNNNNNNNNNNN 1000
mouse_plekhg3 --CCAGGAATGCCCC----------AT-TTCTAAATGGGCTTTTTTT 1000
cow_plekhg3 ----------GG-GGTACAAGTGTGTGTTTTC 1000
platypus_plekhg3 AAG-GGGAGTGCCCTTTGGCAGGGCCTTAACTGCATCTGGTTACCTAGAGGT 1000

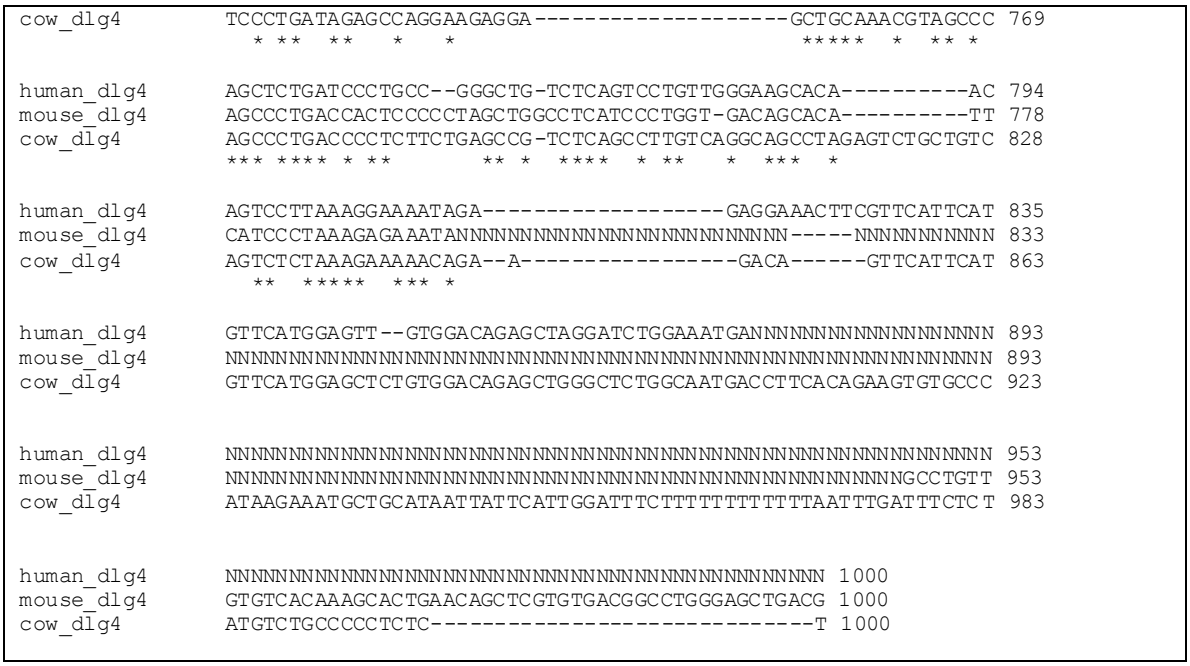
TBC1D10B:


human_tbc1d10b CT----------CCATAGCTCCCCTTACC-ATGAGGTGGAGCTGGCTTCCT 40
mouse_tbc1d10b CTGGGGTGCTGTCTTACACTCCGGTGGCTTCCCTTAG----ATGGTGGAGCTAGGTTCCT 56
cow_tbc1d10b C----------CAGTCACTCCCCTTACCCGTGAAGTGAATCTGGGTTCCT 40
platypus_tbc1d10b CTGGATCG----------TAAG-TGTCCCCAGACTTTCGCTTTGAAGCCTCCTTACC 46
human_tbc1d10b TTTCCCTGTCTTC-AGCCCTCCCTG----------TCTCCCCCACTTCC----T-- 79
mouse_tbc1d10b TTTCCT-GTCTTGGGGCCCTGCCTG----------TCACCCCA-CT-GCCTACTGG 99
cow_tbc1d10b TTACCC--CCTTC-AGTCCTGCCTG----------TCTCCCCAGC-TCCTGGCTGG 82
platypus_tbc1d10b GTTTTCTCTTCTC-CACTCACACGGCGCCTAGGTTCCCTAATCCCCTCCATCC----TGT 101
human_tbc1d10b -GGCCAGGG----------CTCTCATTCTGGAC----CTGTGTT-GTAATTGTGTACAGA 123
mouse_tbc1d10b GGGCCAAGG-CTGTTTT-----CTCTCCTGGATATCCTTGTGTT-GTAATTATGTACAGA 152
cow_tbc1d10b GGGCCCAGG-CCCTTCTTGCTCTCCTTCTGGACATCTCTGTGTT-GTAATTATGTACAGA 140
platypus_tbc1d10b TGGCCCGCAGCT-TTCC----CA---------CACC-CTCCCTCCCCACTTCCCT-TTCT 145
human_tbc1d10b GGATGGCG-TTGGCCTG-GGGTGGGGGTGCTCGCTTTGTCTTCTGTCCTTTGG---TTC- 177
mouse_tbc1d10b AGGTCAGG-TTGGCCTG-GGGTGGGTG--CT----------GTCCTT------TCC- 188
cow_tbc1d10b GGACCTGG-TTGGCGCA-GGGTGGGGGTGTTCACTCCTCCTTCTGTTCTCTGG---TTC- 194
platypus_tbc1d10b -GGCCACCCTTGCTCGGTCGCTGACGTCCTTCTCTCCT--TCCTTCCCTGCCTCCCTCCC 202
human_tbc1d10b TCCTTCCATAATGCTCCTGT-AC-CCAGTTTATTTAAGGGGACATGCACTGGAATAGGAA 235
mouse_tbc1d10b CCTCTCCACAGCGCTGCTTC-ACTCCAGTTTATTTAAGAGGATATGC-------TAGGAA 240
cow_tbc1d10b CCCCTCCCCAGTGTTCTTGC-ACTCTAGTTTATTTAAGGGGACACGCACTGGAACAGGAA 253
platypus_tbc1d10b TCCTTCCCT-CCATTCCCTCGCCTCTCGCCTTTCC----CGTCGT-CGTCGGAGTTGAAG 256
human_tbc1d10b ATGTCCCCCATCTCCCTTCCT---GCAC---CCTGCTGTGCTCCCTCCAAACCCACCTTG 289
mouse_tbc1d10b GTGTCCTCCGTCTCCTTCCCAC-----C---CGTGCCCTGCTCTCTCCTAACTCA-CTTG 291
cow_tbc1d10b ATGTCTCCCACCTCCCACCACCACCCTC---CTTGCTCTCCTTCCTCTTCACCCACCTTT 310
platypus_tbc1d10b ACCCCTCCCTCTTCTCTGGGAG----GGAAGGTTGTTT-ACAGTCTGT-CATCC----CT 306
human_tbc1d10b CTCTGTGTTCTCAGGCCCCC-CTGCTTTTGTCTCACCAGGACCCATACCTTTCACCTTGT 348
mouse_tbc1d10b CTTTGTATGCTCAGGCCTCT-CTGCTGCAGTCACA-----AAGTCTGTCTTCGATTTTGT 345
cow_tbc1d10b CTCTGTATTCTCAGGCCTCC-CCTGCTTTGTCTGACCAGGGCCCCTGCCTTTCACTTTG- 368
platypus_tbc1d10b CTCGCCCTCCCCGCGCCTCTTCTACCTCTACCCGTTC-------AGGCCTCT-----CGC 354
human_tbc1d10b TCCCTTCCACCCCTCCAGTTAGTCCCTATCTGGGTAAGG-GTCTTCCCTTGAGCTCCAGG 407
mouse_tbc1d10b TCCTTTCTAGCCATCAAGC----CCCTCTCTGAATAAGG-GTCTTCCCTTGAGT-CCA-G 398
cow_tbc1d10b TCTCTTCCAGTCCTTA--C----CC-TTCCTGGGTAAGGGGTCTTCCCCCAAGCTCCAGG 421
platypus_tbc1d10b CCCCAACCCACA-TT-----G--CGCTTTGTGGAACCGG-GA-AAGAACAAAACTGCT-- 402
human_tbc1d10b GGGTGGAACCCAATGTTTACATTCTCTTCTGTCTCTGCCCCC-ACCCCATG--------- 457
mouse_tbc1d10b GGGTGGAACCCAATGTTTACATTCTCTTCTGTCTTGGTCCCA-CCTCAGTGGCAGTTTTG 457
cow_tbc1d10b GGGTGGAACCCAATGTTTACATGTTCTTCTGTCTCTGCCCCA-ACCCTGTG--------- 471
platypus_tbc1d10b --GTCGTGCCCTACGCCTGCCTCTTGCCCAGTTATTGTGGGATGGTCC----------G 449
human_tbc1d10b CAGCGCTTTGAGGAATTGGAAAA-GAACCTGCTGTTGTAC-----CTGGGCCTGTTTTCT 511
mouse_tbc1d10b TGGCGCTTTGAGGAACCGGAAAATGAACCTGCTGTTATAT----------CTGTCTCCC 506
cow_tbc1d10b CGGCGCTTTGAGGAACCAGAAAA-GAACCTGCTGTTGTAC-----CTGGGCCTGTCTTCT 525
platypus_tbc1d10b GGGTGGGGAGCGGGAGGGCGAAATGAAAGGGACGTAGCAGGGGGCGGGGGAGAGGCCGTT 509
human_tbc1d10b GCCTTGTTATTTGATGAGGGGGGGA-TGGGGTAAGGAC-GAGGGAGGGAGGGACAGAGCC 569
mouse_tbc1d10b GCCTTGTTATTTGAAGAGTGCTGTA---GAGTAAGGG--GAGTGGGGGAGG--CAGGCAC 559
cow_tbc1d10b GCTTTGTTATTTGATGAGTGGGGTT-AAGGACAAGGAGGGAGGGAGTGATGAG-TTAGCA 583
platypus_tbc1d10b GGACT-TTATT-GAAAAACGGGCGCTCGGACGTTACCTCGTTTCATCCACCGC-CCCCCG 566

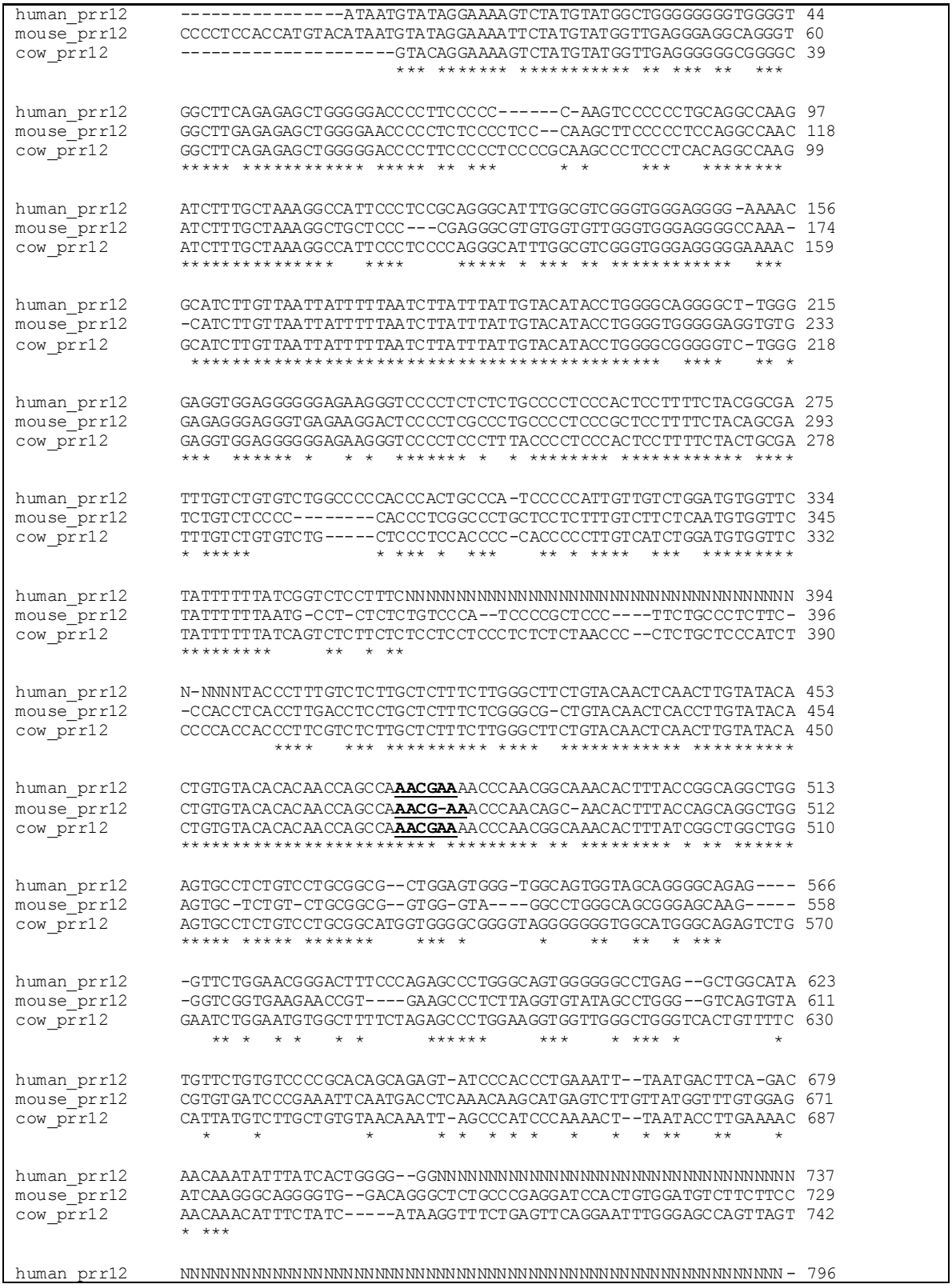

human_tbc1d10b AGGAC--CTGGGTCTCTGTTGAAGCAGTCTGCT-GTCTTGGCAAGTAGGTACCT--CTTA 624
mouse_tbc1d10b AAAGT--CTGACTTTCCATAGGTTCAGTCTGCAGGTCTTGGC----AGGAACAT--CTGG 611
cow_tbc1d10b AGGAT--CTGGGTGTCTCTGGAAGCTGTTGACTGGTCTTGGC----AGGTACCT--TTGG 635
platypus_tbc1d10b AGGTTTCCTCCGTCTCCCCCG--TCGGATGGCTCTTCCTTTAAGGGGAGCCCCTCCCCTA 624
human_tbc1d10b GCTTTGGGTCCATCTCTCAGTCACTCATCAG-----ACACTGGTTGAGCTCCTGCTCCAT 679
mouse_tbc1d10b ACNNNNNNNNNN----NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN 667
cow_tbc1d10b GCCTTTAGTCTA----TTAGTCGCTCATCAA----CACAGTTGTTGAGTTCCTGCTTCAC 687
platypus_tbc1d10b GCCCCGGCGCCAAGTCCCCTTTCCTCCCCAACAC-----CCACCCACACCTCGGTCCCCT 679
human_tbc1d10b CTTTGCCT----GGG----ATCTGCTGTCTCCA-TTGTCA--CTGGCTGCCTGGAGCT-C 727
mouse_tbc1d10b NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN--NNNNNNNNNNNNNNNNNN 725
cow_tbc1d10b CTCCACCTGTTGGGG----ATCTGCTGCCTCCG-TTGTCA--CTGGCCTCCTGAAACT-G 739
platypus_tbc1d10b CCCCGCCGA----------CCCTCCCTTATCCA-AAGTCCAAGAGGCGGCCCTGTCCCTG 728
human_tbc1d10b TCTATGTACCACGTGCAGCACCCGGATCTCCTTCCCAAAGTGCTCATGCAACTCCTGGAA 787
mouse_tbc1d10b NNNNNNNNNNNN-NNNN-NNNNNNNNNNNNNNNNNNNNN-NNNNNNNNNNNNNNNNNNN 782
cow_tbc1d10b TCTGTGGACAGCCTGCAGCATCCGGATCTCCTTTCCAAAGTGCTCGTGCAACTCCTGGAA 799
platypus_tbc1d10b GGCTAGCCTTCCCCTTAATGCCCGGATCTCCTCTCCGAAGCGTTGGTGCAGTTCCTGGAA 788
human_tbc1d10b GCTGGAGGGGGTACTCACAGTGGCTAGAAAGC----CACT---GTCCTCTGGGGATGAG- 839
mouse_tbc1d10b NNNNNNNNNNNNNNNNNNNNNNNNNAGTAAACG---CTCA--CTTCATCTCTCATCAACA 837
cow_tbc1d10b GCTGAAAAGGGTCCTGGGGATGGCTAGGGAGC----CACA---GTCCTCTGGGGGTGAG- 851
platypus_tbc1d10b GCCCGAAGGG-ACCGGGGGGCGGGAGGGGAACGGGCCGCCCTCTTCGTCCCGGGGCGGG- 846
human_tbc1d10b -----------A---TGGG---TTCTTGGCCAGTT--GCCTGGC----ACAGCTCCCAGT 876
mouse_tbc1d10b AACATAGTGGTC---CAGCTCCTTCCTCCTCAGCTTTGCCTGGATTCTGCTGCTCCCACT 894
cow_tbc1d10b -----------C---CGGG---CTCTTGGCCAGCT--GCCTGGC----ACAGCTCCCGGT 888
platypus_tbc1d10b -----------GCCCGGGG---TTCCAGTCCAGCC--CGCGCGC----AGACGTCGCGGT 886
human_tbc1d10b ---ATTGGGTTTTGGTGGGGGGGGAGGGGGGACAGCAGCCCAGCCAGCCGCCTCCAGCTC 933
mouse_tbc1d10b CTCACT-GGCCTCC----------TGGGATCCT--CTGCACAGCTTGCAGGACTGGGATC 941
cow_tbc1d10b ---ACTGGGCCTCC----------AGGGGGCGTAGCAGCTCAGCCAGCCGCGTCCAGGTG 935
platypus_tbc1d10b ---ACTGGGCCTCC----------AAGGGGCTCAGGAGCCTGAGCAGCAGCCCCCAGCCG 933
human_tbc1d10b GCGCCAGCCAGTGCCCACAGCCAAGTCTCCAGCTCTTCCTGGGTTTCTGCAGCCAGCTTG 993
mouse_tbc1d10b GC-CTTTCCAATGGCTTCAGCCCC-CTGGAAGCTGAAGGAGGG----GCCTG--AGAGTG 993
cow_tbc1d10b GCTCTGGCCAACGCCCACAGCCAGGCTCTCAGCTCCTCCGGGTTTTCTGCTGCCAGCCTG 995
platypus_tbc1d10b GCCCCGCCCAGAGCCCGCAGCCAGCCCCCCAGGTCCGCCTCGCTCTCGGCCGCCAGCTTG 993
human_tbc1d10b TAGGCCC 1000
mouse_tbc1d10b TGTGCAA 1000
cow_tbc1d10b TAG--GC 1000
platypus_tbc1d10b TAGGCCC 1000

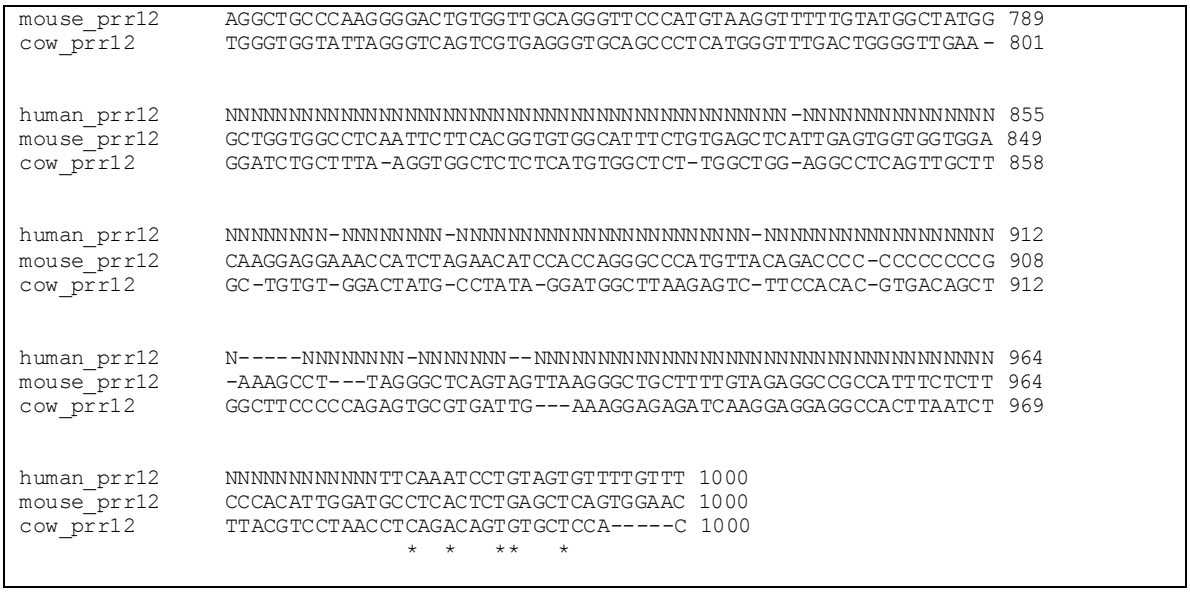
DLG4:


human_dlg4 CTCCCCCATCCTCTCCCACACACATTCCAGAAGTCAGGGCCCC-CTCGAGGAGCACCCGC 59
mouse_dlg4 ----------CATTCCAGAAGTCAGGGCCCCTCTCAAGGAGTACCCGC 38
cow_dlg4 ----------ATCTCCCACACACTTTCCGAAAGTCAGGGCCCCCCTCCAGGAGCACCC-- 48
human_dlg4 TGCAGGGATGCAGGGCCACAGGCCTCCGCTCTCTCC-TAAGGCAGGGTCTGGGGTCACCC 118
mouse_dlg4 TGTAGGGATGCAGGGCCACAGGCTTCCGCGCTCTCC-CGAGGCAGGGGCTGGGGTCACCC 97
cow_dlg4 ------ATTGCAGGGCTGCAGGCCTCCGCGCTCCCCCCGAGGCAGGGTCTGGGGTCACTC 102
human_dlg4 CTGCCCTCATCGTAATTCCCCATGTTCCTTGATTTCTCATTTATTTTTTCCACTTTTTTT 178
mouse_dlg4 CTGCCC-CATCATAACTCCCCACGCCGTTTGAGTTCTCCTTTATTTTCTCC-ATTCTTTT 155
cow_dlg4 CTGACCCCATCATAACTCCCCAGGTCCCTTGATTTCTCATTTTTTTT-TCC-ATTTTTTT 160
human_dlg4 CTTCTCAAAGGTGGTTTTTTGGGGGGAGAAGCAGGGG-ACTCCGCAGCGG--GCCCCTGC 235
mouse_dlg4 CTTCTCAAAGGTGGTTTT-TGGGGGGAGAAGCAGGGGGGCTCTCCTGAGGGTCCCCCCGT 214
cow_dlg4 CTTCTCAAAGGTGGTTTT--GGGGGGATTAGTGGGGGGATTCCACCATGGG-CCCCCTGC 217
human_dlg4 CTTCCACATGCC-CCCACCATTTTTCTTTGCCGGTTTGCATGAGTGGAAGGTCTAAATGT 294
mouse_dlg4 CTTTCACACACCTCCCACCTTTTTTCTTTGCCGGTTTGCATGAGTGGAAGGTCTAACTGT 274
cow_dlg4 CTTCCACATGCCCCCTACCTTTTTTCTTTGCCGGTTTGCATGAGTGGAAGGTCTAACTGT 277
human_dlg4 GGCTTTTTTTTTTTTTT-----TTC----------CTGGGAATTTT---------- 325
mouse_dlg4 GGCTTTTTTTTTTTCTGGGNNNNNNNNNNNNNNNNNNNNNNNNNNNNN---------- 323
cow_dlg4 GGCTTTTTTTTTT-TT-----TTC----------CTGGGATTTTTTTCTTTTTTTT 318
human_dlg4 ----------TTTGGGGAAAAGGGAGGGATGGGTCTAGGGAGTGGGAAATGCGGGAG 372
mouse_dlg4 ----------TTTGGGAAAAGGGGAGGGATGGGTCTGGGGAGTGGGGAATGTGGGAG 370
cow_dlg4 TTTTCCTTTTTTCTTTTGGGAACGGGGAGGGATGGGTCTAGGGAGTGGGAATGTTGGGGG 378
human_dlg4 GGAGGGTGGGGGGGCA-GGGGTCGGGGGTCGGGTGTCCGGGAGCCAGGG-AAGACTGGAAA 430
mouse_dlg4 GGGGGTGGGGGGGCAAAGGGTCAGGGGTTGGGTGTCCGGGAGCCAGGGGAGGACAGGAAA 430
cow_dlg4 GGAGGGTAGGGGGGGCAGGGTTGGGGGTCGGGTGTCTGGGAGCCAGGGGAAGACTGGAAA 438
human_dlg4 TGCTGCCGCCTTCTGCAATTTATTTATTTTTTTCTTTTGAGAGAGTGA**AAGGAA**GAGACA 490
mouse_dlg4 TGCTGCCGCCTTCTGCAATTTATTTATTTTTTTCTTTTGAGAGAGTGA**AAGGAA**GAGACA 490
cow_dlg4 TGCTGCCGCCTTCTGCAATTTATTTA-TTTTTTCTTTCGAGAGAGTGA**AAGGAA**GAGACG 497
human_dlg4 GATACTTGAAACTTGGTGTGTGGCCTGGTTATTTGGGACCTGGGTGTGGAG-----GGAG 545
mouse_dlg4 GACACTTGAAACCCCGTGTGTGGCCTGGTTCTTTGGGGTGAGGGAGGGAAGTGGCCTGGG 550
cow_dlg4 GATACTTGAAACCCAGTGTGTGGCCTGGTTATTTGGGACCCGGGTGAGGAG-----GGAG 552
human_dlg4 -------------------ATGGGCGCAGCAGAGTCAGATCCTTCTTCACTCCAAGCCA 585
mouse_dlg4 AT----------------GGGTGGCTCTGAGAACTTCAGATCCTTCCTTACCTCA---CA 591
cow_dlg4 ACCCGGGTGGGAGTATAGGGATGGGTCTGGTAAAGTCAGATCCTTCTCCCCCCACACCCA 612
human_dlg4 G---------------AGAGGTTAGTGGCCTGGAAAATGGGCCAGGGAGGTAACTGGCAA 630
mouse_dlg4 G----------------AGTATAGGGT-GCCT--AAAGGGGTCATGGGAAGGACTTACCAG 633
cow_dlg4 TCACCTCCCCCGAGACACAACAGAATGGCCT--GAAGGGGGCCTTGGAGGAACCTGC-AG 669
human_dlg4 ATCAAGGAACCCGAGTTGGTGAAGACTGAGCCTG-GGAAGGTCTGGAGCTCTGTCC-AAA 688
mouse_dlg4 ACCGAGGCAGACAGAGCTGCCACCGCCGGGCCTG-GGAAGAGCCGGAGCTGGGT---TAG 689
cow_dlg4 ATCAAGGCTCCAGCAAAGAATGGGCCTGGGGTGGGGGCGGGGCTGGAGTTCCATCCTAGG 729
human_dlg4 GTCATGACAG-GCCAGAAAGGGGAGGCTGGAACTTGTCTGGGGCCCTGCAGACTTAGTCC 747
mouse_dlg4 CCCCTGGCAGACTCGCCATTGATC--------------------CCTGCAAGCTCAGTCG 729

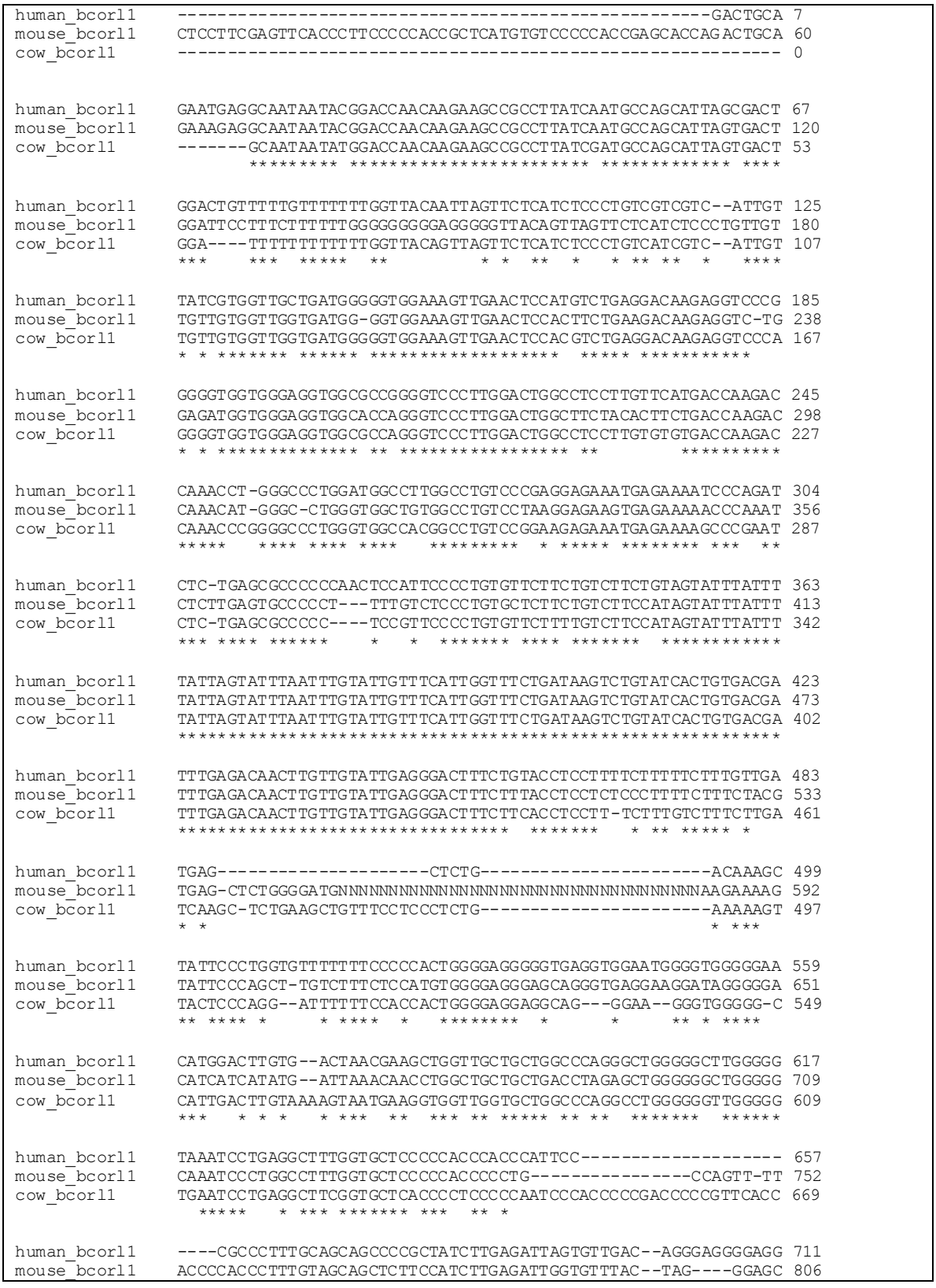

cow_dlg4 TCCCTGATAGAGCCAGGAAGAGGA----------GCTGCAAACGTAGCCC 769
human_dlg4 AGCTCTGATCCCTGCC--GGGCTG-TCTCAGTCCTGTTGGGAAGCACA----------AC 794
mouse_dlg4 AGCCCTGACCACTCCCCCTAGCTGGCCTCATCCCTGGT-GACAGCACA----------TT 778
cow_dlg4 AGCCCTGACCCCTCTTCTGAGCCG-TCTCAGCCTTGTCAGGCAGCCTAGAGTCTGCTGTC 828
human_dlg4 AGTCCTTAAAGGAAAATAGA----------GAGGAAACTTCGTTCATTCAT 835
mouse_dlg4 CATCCCTAAAGAGAAATANNNNNNNNNNNNNNNNNNNNNNN-----NNNNNNNNNN 833
cow_dlg4 AGTCTCTAAAGAAAAACAGA--A----------GACA------GTTCATTCAT 863
human_dlg4 GTTCATGGAGTT--GTGGACAGAGCTAGGATCTGGAAATGANNNNNNNNNNNNNNNNNN 893
mouse_dlg4 NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN 893
cow_dlg4 GTTCATGGAGCTCTGTGGACAGAGCTGGGCTCTGGCAATGACCTTCACAGAAGTGTGCCC 923
human_dlg4 NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN 953
mouse_dlg4 NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNGCCTGTT 953
cow_dlg4 ATAAGAAATGCTGCATAATTATTCATTGGATTTCTTTTTTTTTTTAATTTGATTTCTCT 983
human_dlg4 NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN 1000
mouse_dlg4 GTGTCACAAAGCACTGAACAGCTCGTGTGACGGCCTGGGAGCTGACG 1000
cow_dlg4 ATGTCTGCCCCCTCTC----------T 1000

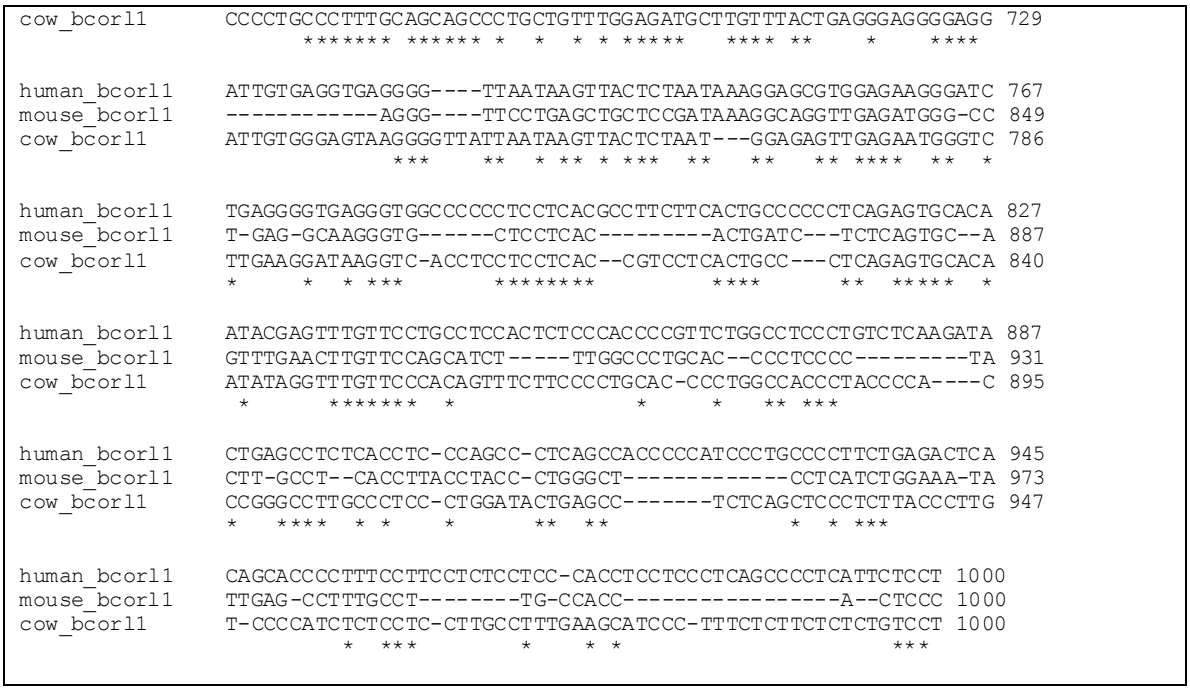
PRR12:


human_prr12 ----------ATAATGTATAGGAAAAGTCTATGTATGGCTGGGGGGGGTGGGGT 44
mouse_prr12 CCCCTCCACCATGTACATAATGTATAGGAAAATTCTATGTATGGTTGAGGGAGGCAGGGT 60
cow_prr12 ----------GTACAGGAAAAGTCTATGTATGGTTGAGGGGGGCGGGGC 39
human_prr12 GGCTTCAGAGAGCTGGGGGACCCCTTCCCCC------C-AAGTCCCCCCTGCAGGCCAAG 97
mouse_prr12 GGCTTGAGAGAGCTGGGGAACCCCCTCTCCCCTCC--CAAGCTTCCCCCTCCAGGCCAAC 118
cow_prr12 GGCTTCAGAGAGCTGGGGGACCCCTTCCCCCTCCCCGCAAGCCCTCCCTCACAGGCCAAG 99
human_prr12 ATCTTTGCTAAAGGCCATTCCCTCCGCAGGGCATTTGGCGTCGGGTGGGAGGGG-AAAAC 156
mouse_prr12 ATCTTTGCTAAAGGCTGCTCCC---CGAGGGCGTGTGGTGTTGGGTGGGAGGGGCCAAA- 174
cow_prr12 ATCTTTGCTAAAGGCCATTCCCTCCCCAGGGCATTTGGCGTCGGGTGGGAGGGGGAAAAC 159
human_prr12 GCATCTTGTTAATTATTTTTAATCTTATTTATTGTACATACCTGGGGCAGGGGCT-TGGG 215
mouse_prr12 -CATCTTGTTAATTATTTTTAATCTTATTTATTGTACATACCTGGGGTGGGGGAGGTGTG 233
cow_prr12 GCATCTTGTTAATTATTTTTAATCTTATTTATTGTACATACCTGGGGCGGGGGTC-TGGG 218
human_prr12 GAGGTGGAGGGGGGAGAAGGGTCCCCTCTCTCTGCCCCTCCCACTCCTTTTCTACGGCGA 275
mouse_prr12 GAGAGGGAGGGTGAGAAGGACTCCCCTCGCCCTGCCCCTCCCGCTCCTTTTCTACAGCGA 293
cow_prr12 GAGGTGGAGGGGGGAGAAGGGTCCCCTCCCTTTACCCCTCCCACTCCTTTTCTACTGCGA 278
human_prr12 TTTGTCTGTGTCTGGCCCCCACCCACTGCCCA-TCCCCCATTGTTGTCTGGATGTGGTTC 334
mouse_prr12 TCTGTCTCCCC--------CACCCTCGGCCCTGCTCCTCTTTGTCTTCTCAATGTGGTTC 345
cow_prr12 TTTGTCTGTGTCTG-----CTCCCTCCACCCC-CACCCCCTTGTCATCTGGATGTGGTTC 332
human_prr12 TATTTTTTATCGGTCTCCTTTCNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN 394
mouse_prr12 TATTTTTTAATG-CCT-CTCTCTGTCCCA--TCCCCGCTCCC----TTCTGCCCTCTTC- 396
cow_prr12 TATTTTTTATCAGTCTCTTCTCTCCTCCTCCCTCTCTCTAACCC--CTCTGCTCCCATCT 390
human_prr12 N-NNNNTACCCTTTGTCTCTTGCTCTTTCTTGGGCTTCTGTACAACTCAACTTGTATACA 453
mouse_prr12 -CCACCTCACCTTGACCTCCTGCTCTTTCTCGGGCG-CTGTACAACTCACCTTGTATACA 454
cow_prr12 CCCCACCACCCTTCGTCTCTTGCTCTTTCTTGGGCTTCTGTACAACTCAACTTGTATACA 450
human_prr12 CTGTGTACACACAACCAGCCA**AACGAA**AACCCAACGGCAAACACTTTACCGGCAGGCTGG 513
mouse_prr12 CTGTGTACACACAACCAGCCA**AACG-AA**ACCCAACAGC-AACACTTTACCAGCAGGCTGG 512
cow_prr12 CTGTGTACACACAACCAGCCA**AACGAA**AACCCAACGGCAAACACTTTATCGGCTGGCTGG 510
human_prr12 AGTGCCTCTGTCCTGCGGCG--CTGGAGTGGG-TGGCAGTGGTAGCAGGGGCAGAG---- 566
mouse_prr12 AGTGC-TCTGT-CTGCGGCG--GTGG-GTA----GGCCTGGGCAGCGGGAGCAAG----- 558
cow_prr12 AGTGCCTCTGTCCTGCGGCATGGTGGGGCGGGGTAGGGGGGGTGGCATGGGCAGAGTCTG 570
human_prr12 -GTTCTGGAACGGGACTTTCCCAGAGCCCTGGGCAGTGGGGGGCCTGAG--GCTGGCATA 623
mouse_prr12 -GGTCGGTGAAGAACCGT----GAAGCCCTCTTAGGTGTATAGCCTGGG--GTCAGTGTA 611
cow_prr12 GAATCTGGAATGTGGCTTTTCTAGAGCCCTGGAAGGTGGTTGGGCTGGGTCACTGTTTTC 630
human_prr12 TGTTCTGTGTCCCCGCACAGCAGAGT-ATCCCACCCTGAAATT--TAATGACTTCA-GAC 679
mouse_prr12 CGTGTGATCCCGAAATTCAATGACCTCAAACAAGCATGAGTCTTGTTATGGTTTGTGGAG 671
cow_prr12 CATTATGTCTTGCTGTGTAACAAATT-AGCCCATCCCAAAACT--TAATACCTTGAAAAC 687
human_prr12 AACAAATATTTATCACTGGGG--GGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN 737
mouse_prr12 ATCAAGGGCAGGGGTG--GACAGGGCTCTGCCCGAGGATCCACTGTGGATGTCTTCTTCC 729
cow_prr12 AACAAACATTTCTATC-----ATAAGGTTTCTGAGTTCAGGAATTTGGGAGCCAGTTAGT 742
human_prr12 NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN- 796

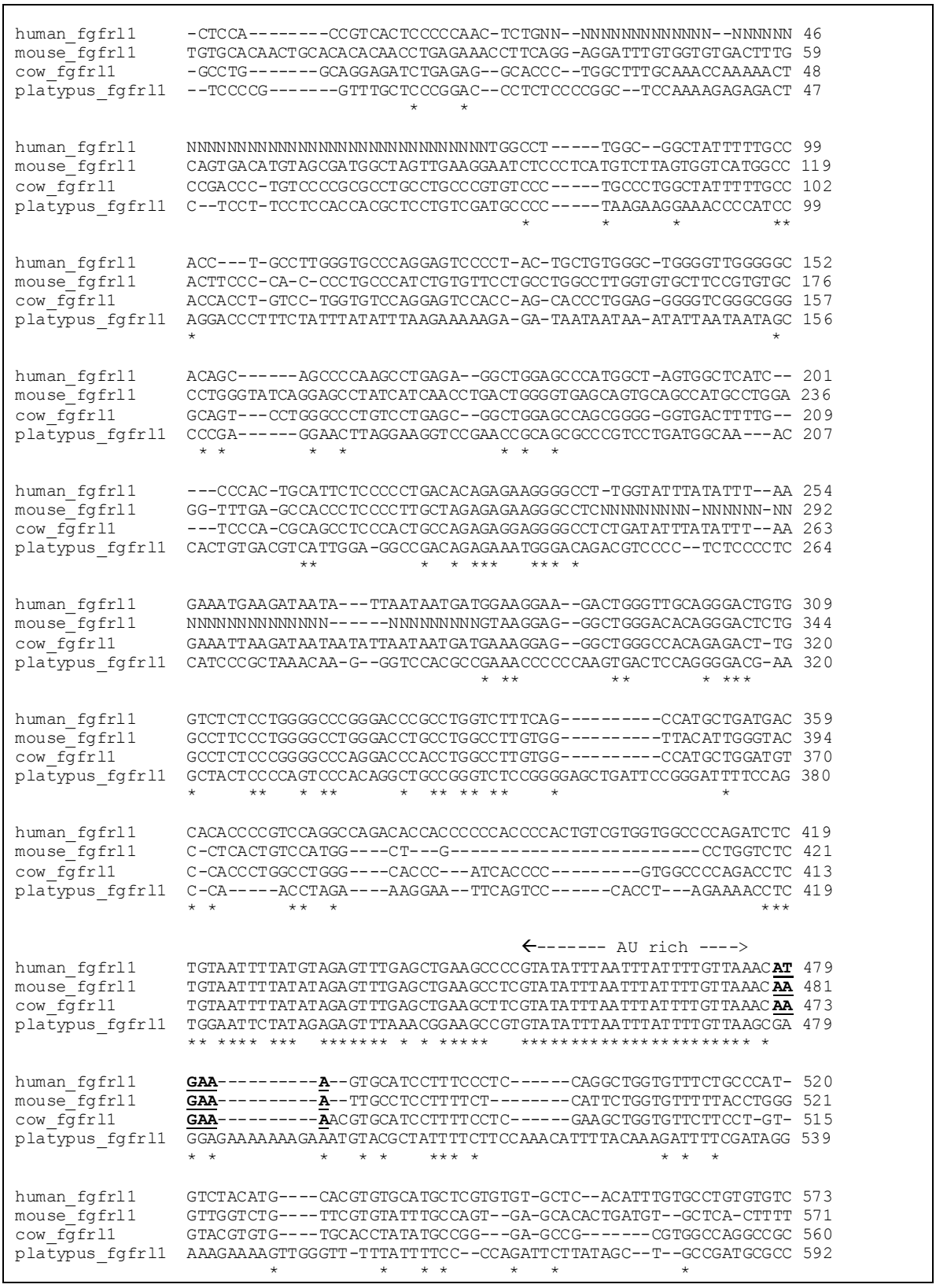

mouse_prr12 AGGCTGCCCAAGGGGACTGTGGTTGCAGGGTTCCCATGTAAGGTTTTTGTATGGCTATGG 789
cow_prr12 TGGGTGGTATTAGGGTCAGTCGTGAGGGTGCAGCCCTCATGGGTTTGACTGGGGTTGAA- 801
human_prr12 NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN-NNNNNNNNNNNNNNN 855
mouse_prr12 GCTGGTGGCCTCAATTCTTCACGGTGTGGCATTTCTGTGAGCTCATTGAGTGGTGGTGGA 849
cow_prr12 GGATCTGCTTTA-AGGTGGCTCTCTCATGTGGCTCT-TGGCTGG-AGGCCTCAGTTGCTT 858
human_prr12 NNNNNNNN-NNNNNNNN-NNNNNNNNNNNNNNNNNNNNNNN-NNNNNNNNNNNNNNNNN 912
mouse_prr12 CAAGGAGGAAACCATCTAGAACATCCACCAGGGCCCATGTTACAGACCCC-CCCCCCCCG 908
cow_prr12 GC-TGTGT-GGACTATG-CCTATA-GGATGGCTTAAGAGTC-TTCCACAC-GTGACAGCT 912
human_prr12 N-----NNNNNNNN-NNNNNNN--NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN 964
mouse_prr12 -AAAGCCT---TAGGGCTCAGTAGTTAAGGGCTGCTTTTGTAGAGGCCGCCATTTCTCTT 964
cow_prr12 GGCTTCCCCCAGAGTGCGTGATTG---AAAGGAGAGATCAAGGAGGAGGCCACTTAATCT 969
human_prr12 NNNNNNNNNNNNTTCAAATCCTGTAGTGTTTTGTTT 1000
mouse_prr12 CCCACATTGGATGCCTCACTCTGAGCTCAGTGGAAC 1000
cow_prr12 TTACGTCCTAACCTCAGACAGTGTGCTCCA-----C 1000
* * ** *

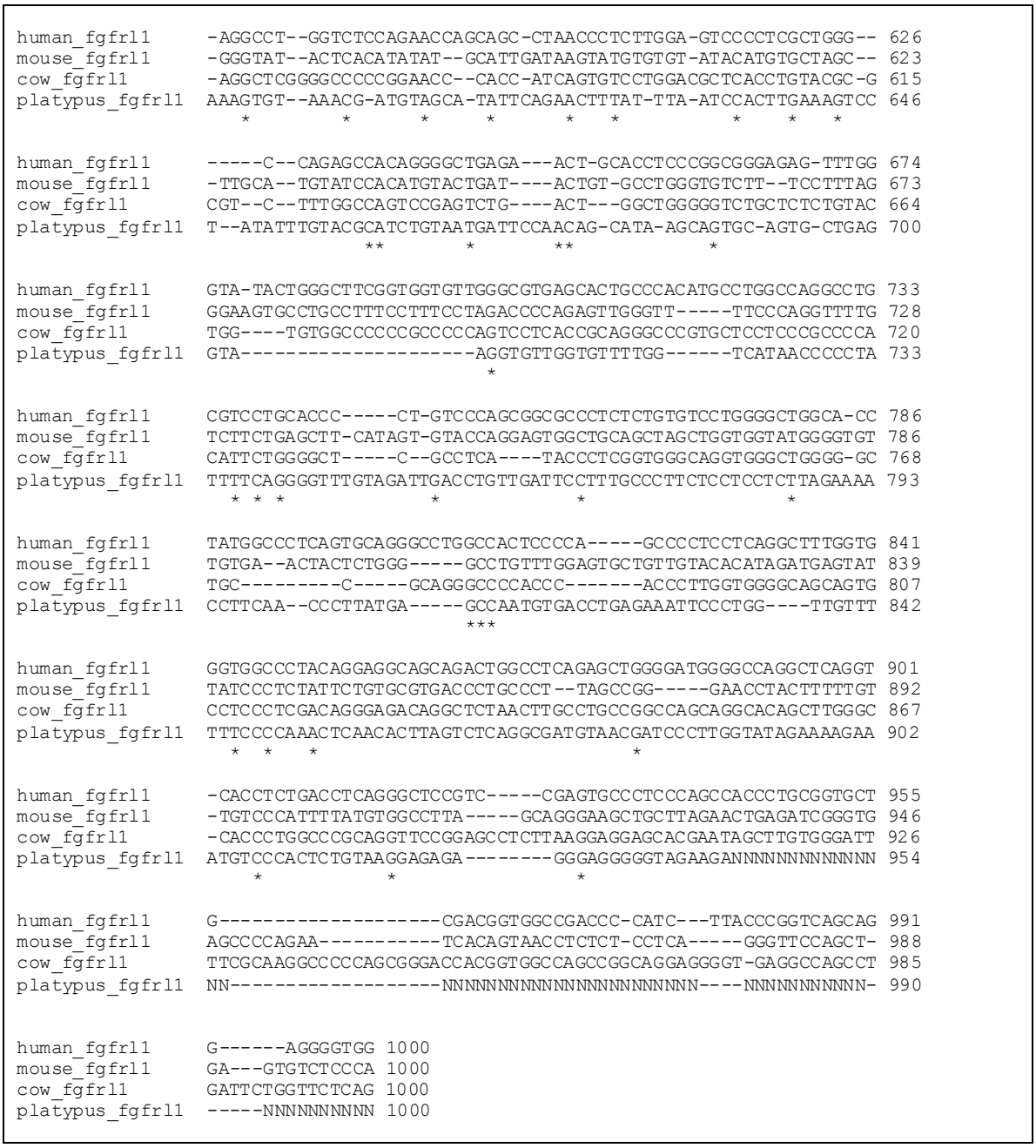
BCORL1:


human_bcorl1 ----------GACTGCA 7
mouse_bcorl1 CTCCTTCGAGTTCACCCTTCCCCCACCGCTCATGTGTCCCCCACCGAGCACCAGACTGCA 60
cow_bcorl1 ---------- 0
human_bcorl1 GAATGAGGCAATAATACGGACCAACAAGAAGCCGCCTTATCAATGCCAGCATTAGCGACT 67
mouse_bcorl1 GAAAGAGGCAATAATACGGACCAACAAGAAGCCGCCTTATCAATGCCAGCATTAGTGACT 120
cow_bcorl1 -------GCAATAATATGGACCAACAAGAAGCCGCCTTATCGATGCCAGCATTAGTGACT 53
human_bcorl1 GGACTGTTTTTGTTTTTTTGGTTACAATTAGTTCTCATCTCCCTGTCGTCGTC--ATTGT 125
mouse_bcorl1 GGATTCCTTTCTTTTTTGGGGGGGGGAGGGGGTTACAGTTAGTTCTCATCTCCCTGTTGT 180
cow_bcorl1 GGA----TTTTTTTTTTTTGGTTACAGTTAGTTCTCATCTCCCTGTCATCGTC--ATTGT 107
human_bcorl1 TATCGTGGTTGCTGATGGGGGTGGAAAGTTGAACTCCATGTCTGAGGACAAGAGGTCCCG 185
mouse_bcorl1 TGTTGTGGTTGGTGATGG-GGTGGAAAGTTGAACTCCACTTCTGAAGACAAGAGGTC-TG 238
cow_bcorl1 TGTTGTGGTTGGTGATGGGGGTGGAAAGTTGAACTCCACGTCTGAGGACAAGAGGTCCCA 167
human_bcorl1 GGGGTGGTGGGAGGTGGCGCCGGGGTCCCTTGGACTGGCCTCCTTGTTCATGACCAAGAC 245
mouse_bcorl1 GAGATGGTGGGAGGTGGCACCAGGGTCCCTTGGACTGGCTTCTACACTTCTGACCAAGAC 298
cow_bcorl1 GGGGTGGTGGGAGGTGGCGCCAGGGTCCCTTGGACTGGCCTCCTTGTGTGTGACCAAGAC 227
human_bcorl1 CAAACCT-GGGCCCTGGATGGCCTTGGCCTGTCCCGAGGAGAAATGAGAAAATCCCAGAT 304
mouse_bcorl1 CAAACAT-GGGC-CTGGGTGGCTGTGGCCTGTCCTAAGGAGAAGTGAGAAAAACCCAAAT 356
cow_bcorl1 CAAACCCGGGGCCCTGGGTGGCCACGGCCTGTCCGGAAGAGAAATGAGAAAAGCCCGAAT 287
human_bcorl1 CTC-TGAGCGCCCCCCAACTCCATTCCCCTGTGTTCTTCTGTCTTCTGTAGTATTTATTT 363
mouse_bcorl1 CTCTTGAGTGCCCCCT---TTTGTCTCCCTGTGCTCTTCTGTCTTCCATAGTATTTATTT 413
cow_bcorl1 CTC-TGAGCGCCCCC----TCCGTTCCCCTGTGTTCTTTTGTCTTCCATAGTATTTATTT 342
human_bcorl1 TATTAGTATTTAATTTGTATTGTTTCATTGGTTTCTGATAAGTCTGTATCACTGTGACGA 423
mouse_bcorl1 TATTAGTATTTAATTTGTATTGTTTCATTGGTTTCTGATAAGTCTGTATCACTGTGACGA 473
cow_bcorl1 TATTAGTATTTAATTTGTATTGTTTCATTGGTTTCTGATAAGTCTGTATCACTGTGACGA 402
human_bcorl1 TTTGAGACAACTTGTTGTATTGAGGGACTTTCTGTACCTCCTTTTCTTTTTCTTTGTTGA 483
mouse_bcorl1 TTTGAGACAACTTGTTGTATTGAGGGACTTTCTTTACCTCCTCTCCCTTTTCTTTCTACG 533
cow_bcorl1 TTTGAGACAACTTGTTGTATTGAGGGACTTTCTTCACCTCCTT-TCTTTGTCTTTCTTGA 461
human_bcorl1 TGAG----------CTCTG----------ACAAAGC 499
mouse_bcorl1 TGAG-CTCTGGGGATGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAAGAAAAG 592
cow_bcorl1 TCAAGC-TCTGAAGCTGTTTCCTCCCTCTG----------AAAAAGT 497
human_bcorl1 TATTCCCTGGTGTTTTTTTCCCCCACTGGGGAGGGGGTGAGGTGGAATGGGGTGGGGGAA 559
mouse_bcorl1 TATTCCCAGCT-TGTCTTTCTCCATGTGGGGAGGGAGCAGGGTGAGGAAGGATAGGGGGA 651
cow_bcorl1 TACTCCCAGG--ATTTTTTCCACCACTGGGGAGGAGGCAG---GGAA--GGGTGGGGG-C 549
human_bcorl1 CATGGACTTGTG--ACTAACGAAGCTGGTTGCTGCTGGCCCAGGGCTGGGGGCTTGGGGG 617
mouse_bcorl1 CATCATCATATG--ATTAAACAACCTGGCTGCTGCTGACCTAGAGCTGGGGGGCTGGGGG 709
cow_bcorl1 CATTGACTTGTAAAAGTAATGAAGGTGGTTGGTGCTGGCCCAGGCCTGGGGGGTTGGGGG 609
human_bcorl1 TAAATCCTGAGGCTTTGGTGCTCCCCCACCCACCCATTCC---------- 657
mouse_bcorl1 CAAATCCCTGGCCTTTGGTGCTCCCCCACCCCCTG----------CCAGTT-TT 752
cow_bcorl1 TGAATCCTGAGGCTTCGGTGCTCACCCCTCCCCCAATCCCACCCCCGACCCCCGTTCACC 669
human_bcorl1 ----CGCCCTTTGCAGCAGCCCCGCTATCTTGAGATTAGTGTTGAC--AGGGAGGGGAGG 711
mouse_bcorl1 ACCCCACCCTTTGTAGCAGCTCTTCCATCTTGAGATTGGTGTTTAC--TAG----GGAGC 806

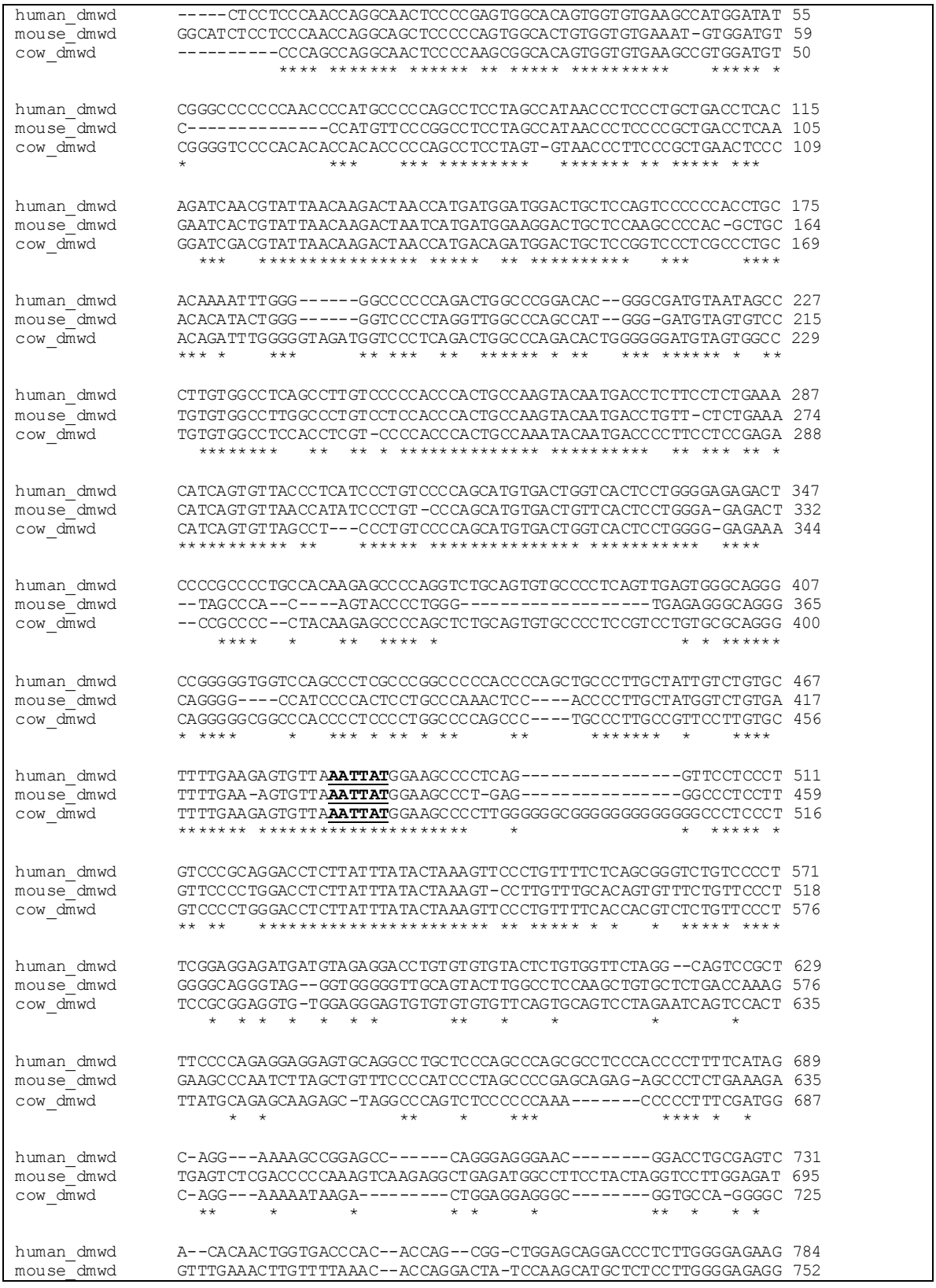

cow_bcorl1 CCCCTGCCCTTTGCAGCAGCCCTGCTGTTTGGAGATGCTTGTTTACTGAGGGAGGGGAGG 729
human_bcorl1 ATTGTGAGGTGAGGGG----TTAATAAGTTACTCTAATAAAGGAGCGTGGAGAAGGGATC 767
mouse_bcorl1 ----------AGGG----TTCCTGAGCTGCTCCGATAAAGGCAGGTTGAGATGGG-CC 849
cow_bcorl1 ATTGTGGGAGTAAGGGGTTATTAATAAGTTACTCTAAT---GGAGAGTTGAGAATGGGTC 786
human_bcorl1 TGAGGGGTGAGGGTGGCCCCCCTCCTCACGCCTTCTTCACTGCCCCCCTCAGAGTGCACA 827
mouse_bcorl1 T-GAG-GCAAGGGTG------CTCCTCAC---------ACTGATC---TCTCAGTGC--A 887
cow_bcorl1 TTGAAGGATAAGGTC-ACCTCCTCCTCAC--CGTCCTCACTGCC---CTCAGAGTGCACA 840
human_bcorl1 ATACGAGTTTGTTCCTGCCTCCACTCTCCCACCCCGTTCTGGCCTCCCTGTCTCAAGATA 887
mouse_bcorl1 GTTTGAACTTGTTCCAGCATCT-----TTGGCCCTGCAC--CCCTCCCC---------TA 931
cow_bcorl1 ATATAGGTTTGTTCCCACAGTTTCTTCCCCTGCAC-CCCTGGCCACCCTACCCCA----C 895
human_bcorl1 CTGAGCCTCTCACCTC-CCAGCC-CTCAGCCACCCCCATCCCTGCCCCTTCTGAGACTCA 945
mouse_bcorl1 CTT-GCCT--CACCTTACCTACC-CTGGGCT-------------CCTCATCTGGAAA-TA 973
cow_bcorl1 CCGGGCCTTGCCCTCC-CTGGATACTGAGCC-------TCTCAGCTCCCTCTTACCCTTG 947
human_bcorl1 CAGCACCCCTTTCCTTCCTCTCCTCC-CACCTCCTCCCTCAGCCCCTCATTCTCCT 1000
mouse_bcorl1 TTGAG-CCTTTGCCT--------TG-CCACC----------------A--CTCCC 1000
cow_bcorl1 T-CCCCATCTCTCCTC-CTTGCCTTTGAAGCATCCC-TTTCTCTTCTCTCTGTCCT 1000

FGFRL1:


human_fgfrl1 -CTCCA--------CCGTCACTCCCCCAAC-TCTGNN--NNNNNNNNNNNN--NNNNNN 46
mouse_fgfrl1 TGTGCACAACTGCACACACAACCTGAGAAACCTTCAGG-AGGATTTGTGGTGTGACTTTG 59
cow_fgfrl1 -GCCTG-------GCAGGAGATCTGAGAG--GCACCC--TGGCTTTGCAAACCAAAAACT 48
platypus_fgfrl1 --TCCCCG-------GTTTGCTCCCGGAC--CCTCTCCCCGGC--TCCAAAAGAGAGACT 47
human_fgfrl1 NNNNNNNNNNNNNNNNNNNNNNNNNNNTGGCCT-----TGGC--GGCTATTTTTGCC 99
mouse_fgfrl1 CAGTGACATGTAGCGATGGCTAGTTGAAGGAATCTCCCTCATGTCTTAGTGGTCATGGCC 119
cow_fgfrl1 CCGACCC-TGTCCCCGCGCCTGCCTGCCCGTGTCCC-----TGCCCTGGCTATTTTTGCC 102
platypus_fgfrl1 C--TCCT-TCCTCCACCACGCTCCTGTCGATGCCCC-----TAAGAAGGAAACCCCATCC 99
human_fgfrl1 ACC---T-GCCTTGGGTGCCCAGGAGTCCCCT-AC-TGCTGTGGGC-TGGGGTTGGGGGC 152
mouse_fgfrl1 ACTTCCC-CA-C-CCCTGCCCATCTGTGTTCCTGCCTGGCCTTGGTGTGCTTCCGTGTGC 176
cow_fgfrl1 ACCACCT-GTCC-TGGTGTCCAGGAGTCCACC-AG-CACCCTGGAG-GGGGTCGGGCGGG 157
platypus_fgfrl1 AGGACCCTTTCTATTTATATTTAAGAAAAAGA-GA-TAATAATAA-ATATTAATAATAGC 156
human_fgfrl1 ACAGC------AGCCCCAAGCCTGAGA--GGCTGGAGCCCATGGCT-AGTGGCTCATC-- 201
mouse_fgfrl1 CCTGGGTATCAGGAGCCTATCATCAACCTGACTGGGGTGAGCAGTGCAGCCATGCCTGGA 236
cow_fgfrl1 GCAGT---CCTGGGCCCTGTCCTGAGC--GGCTGGAGCCAGCGGGG-GGTGACTTTTG-- 209
platypus_fgfrl1 CCCGA------GGAACTTAGGAAGGTCCGAACCGCAGCGCCCGTCCTGATGGCAA---AC 207
human_fgfrl1 ---CCCAC-TGCATTCTCCCCCTGACACAGAGAAGGGGCCT-TGGTATTTATATTT--AA 254
mouse_fgfrl1 GG-TTTGA-GCCACCCTCCCCTTGCTAGAGAGAAGGGCCTCNNNNNNNNN-NNNNNN-NN 292
cow_fgfrl1 ---TCCCA-CGCAGCCTCCCACTGCCAGAGAGGAGGGGCCTCTGATATTTATATTT--AA 263
platypus_fgfrl1 CACTGTGACGTCATTGGA-GGCCGACAGAGAAATGGGACAGACGTCCCC--TCTCCCCTC 264
human_fgfrl1 GAAATGAAGATAATA---TTAATAATGATGGAAGGAA--GACTGGGTTGCAGGGACTGTG 309
mouse_fgfrl1 NNNNNNNNNNNNNN------NNNNNNNNNGTAAGGAG--GGCTGGGACACAGGGACTCTG 344
cow_fgfrl1 GAAATTAAGATAATAATATTAATAATGATGAAAGGAG--GGCTGGGCCACAGAGACT-TG 320
platypus_fgfrl1 CATCCCGCTAAACAA-G--GGTCCACGCCGAAACCCCCCAAGTGACTCCAGGGGACG-AA 320
human_fgfrl1 GTCTCTCCTGGGGCCCGGGACCCGCCTGGTCTTTCAG----------CCATGCTGATGAC 359
mouse_fgfrl1 GCCTTCCCTGGGGCCTGGGACCTGCCTGGCCTTGTGG----------TTACATTGGGTAC 394
cow_fgfrl1 GCCTCTCCCGGGGCCCAGGACCCACCTGGCCTTGTGG----------CCATGCTGGATGT 370
platypus_fgfrl1 GCTACTCCCCAGTCCCACAGGCTGCCGGGTCTCCGGGGAGCTGATTCCGGGATTTTCCAG 380
human_fgfrl1 CACACCCCGTCCAGGCCAGACACCACCCCCCACCCCACTGTCGTGGTGGCCCCAGATCTC 419
mouse_fgfrl1 C-CTCACTGTCCATGG----CT---G------------------------CCTGGTCTC 421
cow_fgfrl1 C-CACCCTGGCCTGGG----CACCC---ATCACCCC---------GTGGCCCCAGACCTC 413
platypus_fgfrl1 C-CA-----ACCTAGA----AAGGAA--TTCAGTCC------CACCT---AGAAAACCTC 419
← AU rich ---->
human_fgfrl1 TGTAATTTTATGTAGAGTTTGAGCTGAAGCCCCGTATATTTAATTTATTTTGTTAAACAT 479
mouse_fgfrl1 TGTAATTTTATATAGAGTTTGAGCTGAAGCCTCGTATATTTAATTTATTTTGTTAAACAA 481
cow_fgfrl1 TGTAATTTTATATAGAGTTTGAGCTGAAGCTTCGTATATTTAATTTATTTTGTTAAACAA 473
platypus_fgfrl1 TGGAATTCTATAGAGAGTTTAAACGGAAGCCGTGTATATTTAATTTATTTTGTTAAGCGA 479
human_fgfrl1 GAA----------A--GTGCATCCTTTCCCTC------CAGGCTGGTGTTTCTGCCCAT- 520
mouse_fgfrl1 GAA----------A--TTGCCTCCTTTTCT--------CATTCTGGTGTTTTTACCTGGG 521
cow_fgfrl1 GAA----------AACGTGCATCCTTTTCCTC------GAAGCTGGTGTTCTTCCT-GT- 515
platypus_fgfrl1 GGAGAAAAAAAGAAATGTACGCTATTTTCTTCCAAACATTTTACAAAGATTTTCGATAGG 539
human_fgfrl1 GTCTACATG----CACGTGTGCATGCTCGTGTGT-GCTC--ACATTTGTGCCTGTGTGTC 573
mouse_fgfrl1 GTTGGTCTG----TTCGTGTATTTGCCAGT--GA-GCACACTGATGT--GCTCA-CTTTT 571
cow_fgfrl1 GTACGTGTG----TGCACCTATATGCCGG---GA-GCCG-------CGTGGCCAGGCCGC 560
platypus_fgfrl1 AAAGAAAAGTTGGGTT-TTTATTTTCC--CCAGATTCTTATAGC--T--GCCGATGCGCC 592

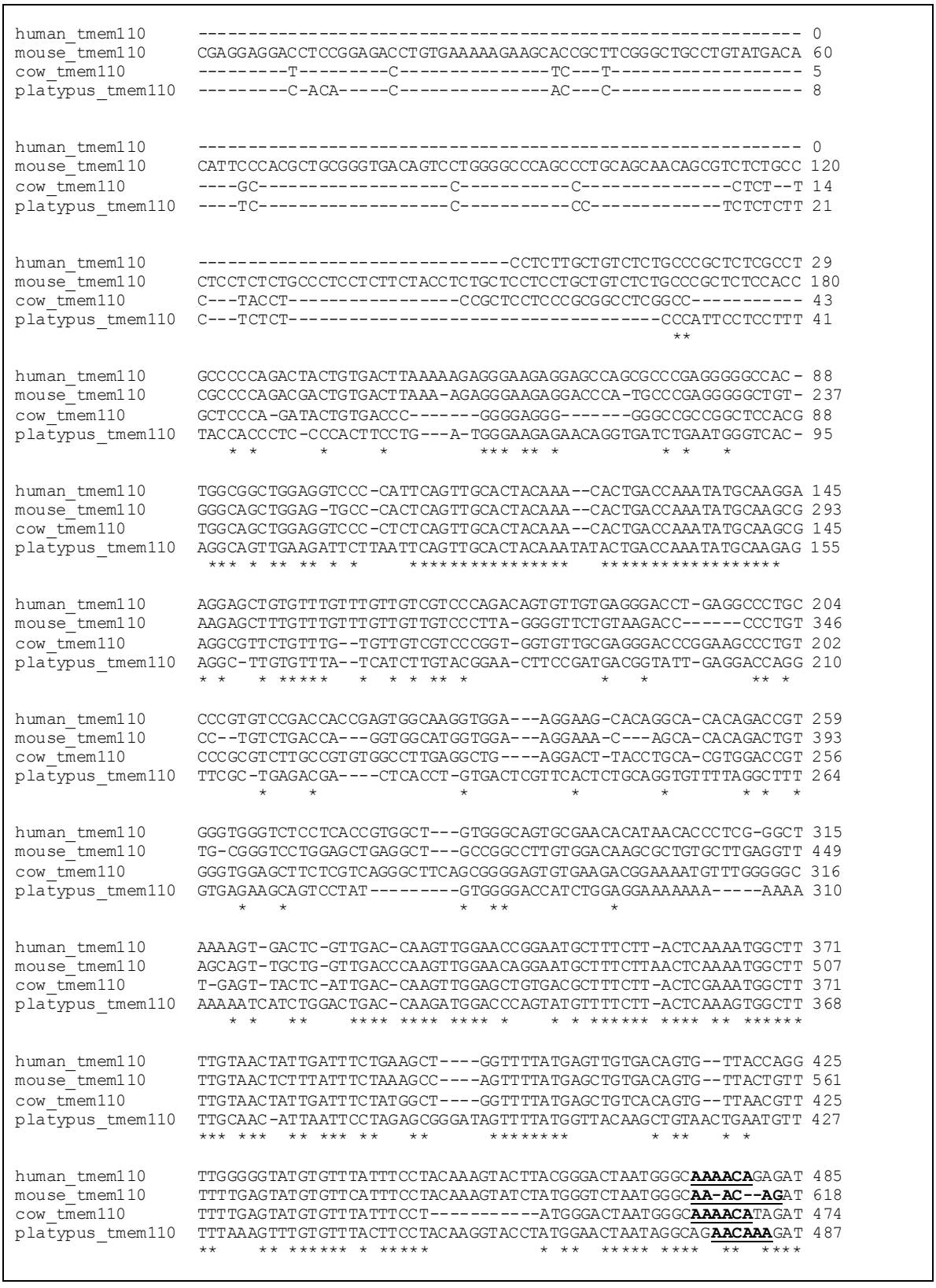

human_fgfrl1 -AGGCCT--GGTCTCCAGAACCAGCAGC-CTAACCCTCTTGGA-GTCCCCTCGCTGGG-- 626
mouse_fgfrl1 -GGGTAT--ACTCACATATAT--GCATTGATAAGTATGTGTGT-ATACATGTGCTAGC-- 623
cow_fgfrl1 -AGGCTCGGGGCCCCCGGAACC--CACC-ATCAGTGTCCTGGACGCTCACCTGTACGC-G 615
platypus_fgfrl1 AAAGTGT--AAACG-ATGTAGCA-TATTCAGAACTTTAT-TTA-ATCCACTTGAAAGTCC 646
human_fgfrl1 -----C--CAGAGCCACAGGGGCTGAGA---ACT-GCACCTCCCGGCGGGAGAG-TTTGG 674
mouse_fgfrl1 -TTGCA--TGTATCCACATGTACTGAT----ACTGT-GCCTGGGTGTCTT--TCCTTTAG 673
cow_fgfrl1 CGT--C--TTTGGCCAGTCCGAGTCTG----ACT---GGCTGGGGGTCTGCTCTCTGTAC 664
platypus_fgfrl1 T--ATATTTGTACGCATCTGTAATGATTCCAACAG-CATA-AGCAGTGC-AGTG-CTGAG 700
human_fgfrl1 GTA-TACTGGGCTTCGGTGGTGTTGGGCGTGAGCACTGCCCACATGCCTGGCCAGGCCTG 733
mouse_fgfrl1 GGAAGTGCCTGCCTTTCCTTTCCTAGACCCCAGAGTTGGGTT-----TTCCCAGGTTTTG 728
cow_fgfrl1 TGG----TGTGGCCCCCCGCCCCCAGTCCTCACCGCAGGGCCCGTGCTCCTCCCGCCCCA 720
platypus_fgfrl1 GTA----------AGGTGTTGGTGTTTTGG------TCATAACCCCCTA 733
human_fgfrl1 CGTCCTGCACCC-----CT-GTCCCAGCGGCGCCCTCTCTGTGTCCTGGGGCTGGCA-CC 786
mouse_fgfrl1 TCTTCTGAGCTT-CATAGT-GTACCAGGAGTGGCTGCAGCTAGCTGGTGGTATGGGGTGT 786
cow_fgfrl1 CATTCTGGGGCT-----C--GCCTCA----TACCCTCGGTGGGCAGGTGGGCTGGGG-GC 768
platypus_fgfrl1 TTTTCAGGGGTTTGTAGATTGACCTGTTGATTCCTTTGCCCTTCTCCTCCTCTTAGAAAA 793
human_fgfrl1 TATGGCCCTCAGTGCAGGGCCTGGCCACTCCCCA-----GCCCCTCCTCAGGCTTTGGTG 841
mouse_fgfrl1 TGTGA--ACTACTCTGGG-----GCCTGTTTGGAGTGCTGTTGTACACATAGATGAGTAT 839
cow_fgfrl1 TGC---------C-----GCAGGGCCCCACCC-------ACCCTTGGTGGGGCAGCAGTG 807
platypus_fgfrl1 CCTTCAA--CCCTTATGA-----GCCAATGTGACCTGAGAAATTCCCTGG----TTGTTT 842
human_fgfrl1 GGTGGCCCTACAGGAGGCAGCAGACTGGCCTCAGAGCTGGGGATGGGGCCAGGCTCAGGT 901
mouse_fgfrl1 TATCCCTCTATTCTGTGCGTGACCCTGCCCT--TAGCCGG-----GAACCTACTTTTTGT 892
cow_fgfrl1 CCTCCCTCGACAGGGAGACAGGCTCTAACTTGCCTGCCGGCCAGCAGGCACAGCTTGGGC 867
platypus_fgfrl1 TTTCCCCAAACTCAACACTTAGTCTCAGGCGATGTAACGATCCCTTGGTATAGAAAAGAA 902
human_fgfrl1 -CACCTCTGACCTCAGGGCTCCGTC-----CGAGTGCCCTCCCAGCCACCCTGCGGTGCT 955
mouse_fgfrl1 -TGTCCCATTTTATGTGGCCTTA-----GCAGGGAAGCTGCTTAGAACTGAGATCGGGTG 946
cow_fgfrl1 -CACCCTGGCCCGCAGGTTCCGGAGCCTCTTAAGGAGGAGCACGAATAGCTTGTGGGATT 926
platypus_fgfrl1 ATGTCCCACTCTGTAAGGAGAGA--------GGGAGGGGGTAGAAGANNNNNNNNNNNN 954
human_fgfrl1 G----------CGACGGTGGCCGACCC-CATC---TTACCCGGTCAGCAG 991
mouse_fgfrl1 AGCCCCAGAA----------TCACAGTAACCTCTCT-CCTCA-----GGGTTCCAGCT- 988
cow_fgfrl1 TTCGCAAGGCCCCCAGCGGGACCACGGTGGCCAGCCGGCAGGAGGGGT-GAGGCCAGCCT 985
platypus_fgfrl1 NN----------NNNNNNNNNNNNNNNNNNNNN----NNNNNNNNNNN- 990
human_fgfrl1 G------AGGGGTGG 1000
mouse_fgfrl1 GA---GTGTCTCCCA 1000
cow_fgfrl1 GATTCTGGTTCTCAG 1000
platypus_fgfrl1 -----NNNNNNNNNN 1000

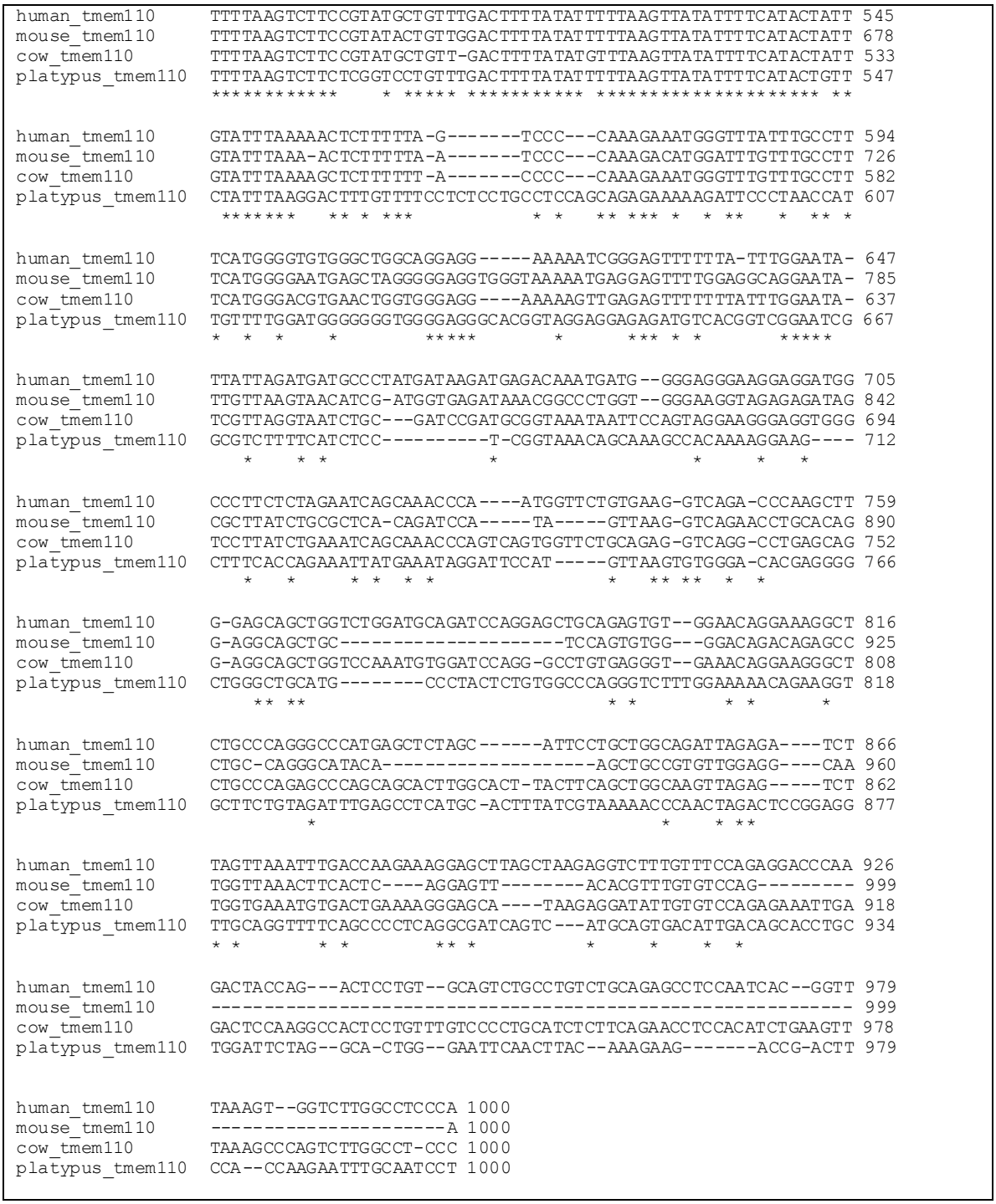
DMWD:


human_dmwd -----CTCCTCCCAACCAGGCAACTCCCCGAGTGGCACAGTGGTGTGAAGCCATGGATAT 55
mouse_dmwd GGCATCTCCTCCCAACCAGGCAGCTCCCCCAGTGGCACTGTGGTGTGAAAT-GTGGATGT 59
cow_dmwd ----------CCCAGCCAGGCAACTCCCCAAGCGGCACAGTGGTGTGAAGCCGTGGATGT 50
human_dmwd CGGGCCCCCCCAACCCCATGCCCCCAGCCTCCTAGCCATAACCCTCCCTGCTGACCTCAC 115
mouse_dmwd C--------------CCATGTTCCCGGCCTCCTAGCCATAACCCTCCCCGCTGACCTCAA 105
cow_dmwd CGGGGTCCCCACACACCACACCCCCAGCCTCCTAGT-GTAACCCTTCCCGCTGAACTCCC 109
human_dmwd AGATCAACGTATTAACAAGACTAACCATGATGGATGGACTGCTCCAGTCCCCCCACCTGC 175
mouse_dmwd GAATCACTGTATTAACAAGACTAATCATGATGGAAGGACTGCTCCAAGCCCCAC-GCTGC 164
cow_dmwd GGATCGACGTATTAACAAGACTAACCATGACAGATGGACTGCTCCGGTCCCTCGCCCTGC 169
human_dmwd ACAAAATTTGGG------GGCCCCCCAGACTGGCCCGGACAC--GGGCGATGTAATAGCC 227
mouse_dmwd ACACATACTGGG------GGTCCCCTAGGTTGGCCCAGCCAT--GGG-GATGTAGTGTCC 215
cow_dmwd ACAGATTTGGGGGTAGATGGTCCCTCAGACTGGCCCAGACACTGGGGGGATGTAGTGGCC 229
human_dmwd CTTGTGGCCTCAGCCTTGTCCCCCACCCACTGCCAAGTACAATGACCTCTTCCTCTGAAA 287
mouse_dmwd TGTGTGGCCTTGGCCCTGTCCTCCACCCACTGCCAAGTACAATGACCTGTT-CTCTGAAA 274
cow_dmwd TGTGTGGCCTCCACCTCGT-CCCCACCCACTGCCAAATACAATGACCCCTTCCTCCGAGA 288
human_dmwd CATCAGTGTTACCCTCATCCCTGTCCCCAGCATGTGACTGGTCACTCCTGGGGAGAGACT 347
mouse_dmwd CATCAGTGTTAACCATATCCCTGT-CCCAGCATGTGACTGTTCACTCCTGGGA-GAGACT 332
cow_dmwd CATCAGTGTTAGCCT---CCCTGTCCCCAGCATGTGACTGGTCACTCCTGGGG-GAGAAA 344
human_dmwd CCCCGCCCCTGCCACAAGAGCCCCAGGTCTGCAGTGTGCCCCTCAGTTGAGTGGGCAGGG 407
mouse_dmwd --TAGCCCA--C----AGTACCCCTGGG----------TGAGAGGGCAGGG 365
cow_dmwd --CCGCCCC--CTACAAGAGCCCCAGCTCTGCAGTGTGCCCCTCCGTCCTGTGCGCAGGG 400
human_dmwd CCGGGGGTGGTCCAGCCCTCGCCCGGCCCCCACCCCAGCTGCCCTTGCTATTGTCTGTGC 467
mouse_dmwd CAGGGG----CCATCCCCACTCCTGCCCAAACTCC----ACCCCTTGCTATGGTCTGTGA 417
cow_dmwd CAGGGGGCGGCCCACCCCTCCCCTGGCCCCAGCCC----TGCCCTTGCCGTTCCTTGTGC 456
human_dmwd TTTTGAAGAGTGTTA**AATTAT**GGAAGCCCCTCAG----------GTTCCTCCCT 511
mouse_dmwd TTTTGAA-AGTGTTA**AATTAT**GGAAGCCCT-GAG----------GGCCCTCCTT 459
cow_dmwd TTTTGAAGAGTGTTA**AATTAT**GGAAGCCCCTTGGGGGGCGGGGGGGGGGGGCCCTCCCT 516
human_dmwd GTCCCGCAGGACCTCTTATTTATACTAAAGTTCCCTGTTTTCTCAGCGGGTCTGTCCCCT 571
mouse_dmwd GTTCCCCTGGACCTCTTATTTATACTAAAGT-CCTTGTTTGCACAGTGTTTCTGTTCCCT 518
cow_dmwd GTCCCCTGGGACCTCTTATTTATACTAAAGTTCCCTGTTTTCACCACGTCTCTGTTCCCT 576
human_dmwd TCGGAGGAGATGATGTAGAGGACCTGTGTGTGTACTCTGTGGTTCTAGG--CAGTCCGCT 629
mouse_dmwd GGGGCAGGGTAG--GGTGGGGGTTGCAGTACTTGGCCTCCAAGCTGTGCTCTGACCAAAG 576
cow_dmwd TCCGCGGAGGTG-TGGAGGGAGTGTGTGTGTGTTCAGTGCAGTCCTAGAATCAGTCCACT 635
human_dmwd TTCCCCAGAGGAGGAGTGCAGGCCTGCTCCCAGCCCAGCGCCTCCCACCCCTTTTCATAG 689
mouse_dmwd GAAGCCCAATCTTAGCTGTTTCCCCATCCCTAGCCCCGAGCAGAG-AGCCCTCTGAAAGA 635
cow_dmwd TTATGCAGAGCAAGAGC-TAGGCCCAGTCTCCCCCCAAA-------CCCCCTTTCGATGG 687
human_dmwd C-AGG---AAAAGCCGGAGCC------CAGGGAGGGAAC--------GGACCTGCGAGTC 731
mouse_dmwd TGAGTCTCGACCCCCAAAGTCAAGAGGCTGAGATGGCCTTCCTACTAGGTCCTTGGAGAT 695
cow_dmwd C-AGG---AAAAATAAGA---------CTGGAGGAGGGC--------GGTGCCA-GGGGC 725
human_dmwd A--CACAACTGGTGACCCAC--ACCAG--CGG-CTGGAGCAGGACCCTCTTGGGGAGAAG 784
mouse_dmwd GTTTGAAACTTGTTTTAAAC--ACCAGGACTA-TCCAAGCATGCTCTCCTTGGGGAGAGG 752

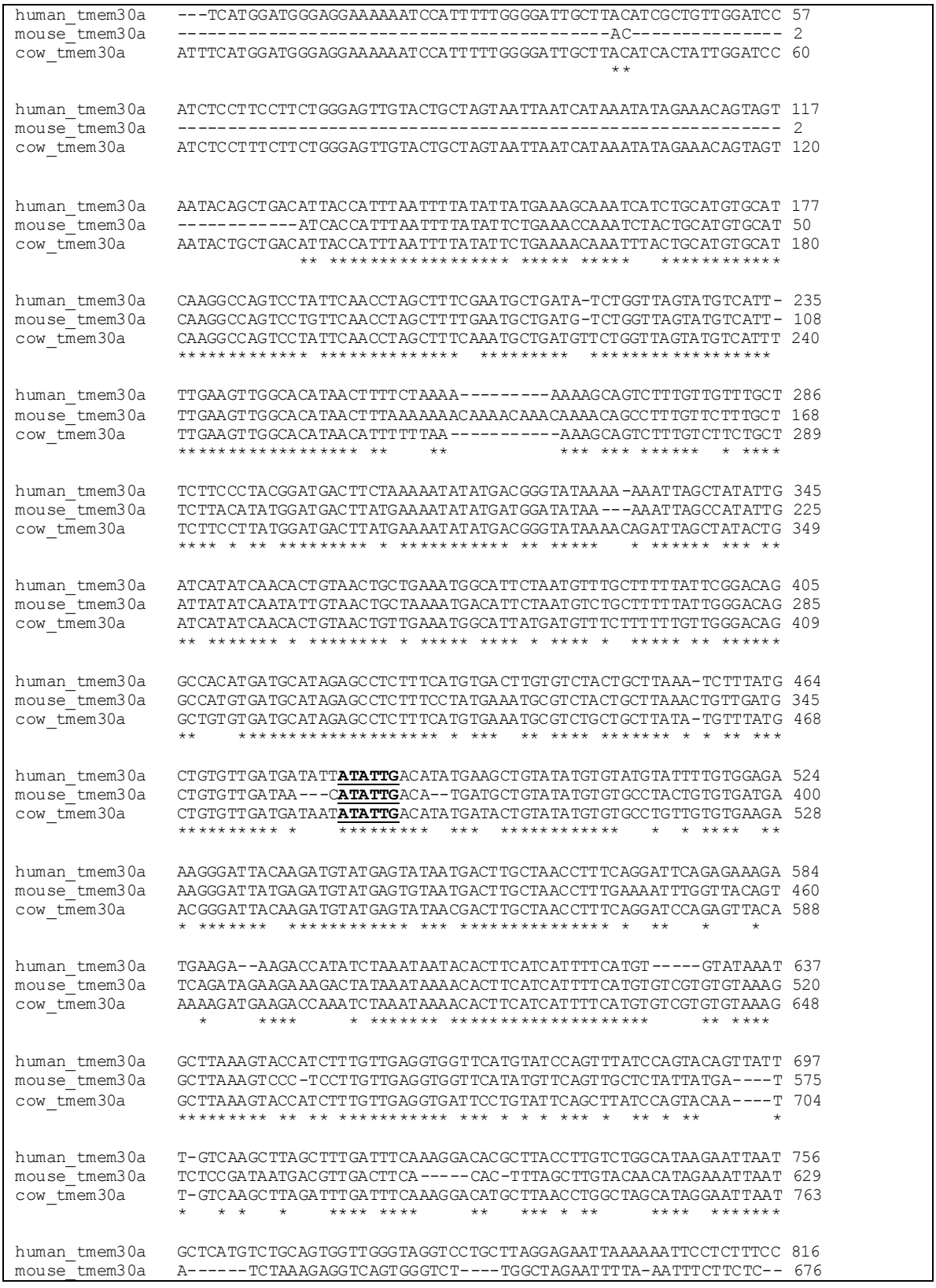
```
cow_dmwd        A--CAAAGCTGGTGGGTCAGTTGCTGGC-CGGCCGGGGGCAGGGCCCTTCAGATGGGGAG 782
                      * ** **     *    *  *  *        *** *  *     *  * *  *

human_dmwd      AGCATCCTG------CCCGCAGCCAGGG-CCCCTCATCAA-AGTCCTCGGTGTTTTTT-- 834
mouse_dmwd      AGGATGCTGGAATTGACTGCACTCCCTGCCTCCTCTGAACATGCCTTTGCAGTCTGCTGC 812
cow_dmwd        GA--TCCCT------CACACTGTC-----CCCACCCCCAA-AC-TCTCAGAGGGAAGG-- 825
                    * *           *   *     * *  *   *       *    *

human_dmwd      ----AAATTA-TCAGAACTGCC-C----AGGACCACGTT--TCCCAG---GCCCTGCC-C 878
mouse_dmwd      CCCTGGCCCATTTATGACTGGC-CATCTAGTGCCAGCTGGAGGTCATGATTTCCTCCC-C 870
cow_dmwd        ----GAGCCG-CCAGTACCCCTTC----AAGGTC---TT--CAGTGG---GCGTTTTTAA 868
                           *  **     *    *    *   *               *

human_dmwd      AGCTGGGACTCCTCGGTCCTTGCC---TCCTAGTTTCTCAGGCCT---GGCC--CTCTCA 930
mouse_dmwd      AGA--GAACT-GGCCACCCTAGAAAGAAGCTAACTTGTC--GCCT---GGCTTGCTGTCC 922
cow_dmwd        AGTTGCCTTTTCAAAGTGTCCAAA---TCATAAGTGATGGGTCCCCAGAGAG--CCCCCG 923
                **       *                  **  *  *    **     *    *   *

human_dmwd      AGG-CCC---AGG----CACCCCAGGCCGGTTGGAGGCCCCGACTTCC--ACTCTGG-AG 979
mouse_dmwd      AGG-CAGCTCCGCCCTCAACCCCTAAAATGTTTCTG-TCTCTAATCCTA-GCCCAGGCAG 979
cow_dmwd        AGGTCCCCTGAGT----GACCTGCTGCCAGCACAA-GCCTTGGCCTGCTCACTCGGCCAG 978
                *** *      *      ***        *        *            * * *  **

human_dmwd      AACC--GTCC-ACCCTGGAAAGAA 1000
mouse_dmwd      GAATGTGGCT-GCCCCGGC--CTG 1000
cow_dmwd        GGGG--GCCCTGGCATCCAGGGCC 1000
                      * *    *
```

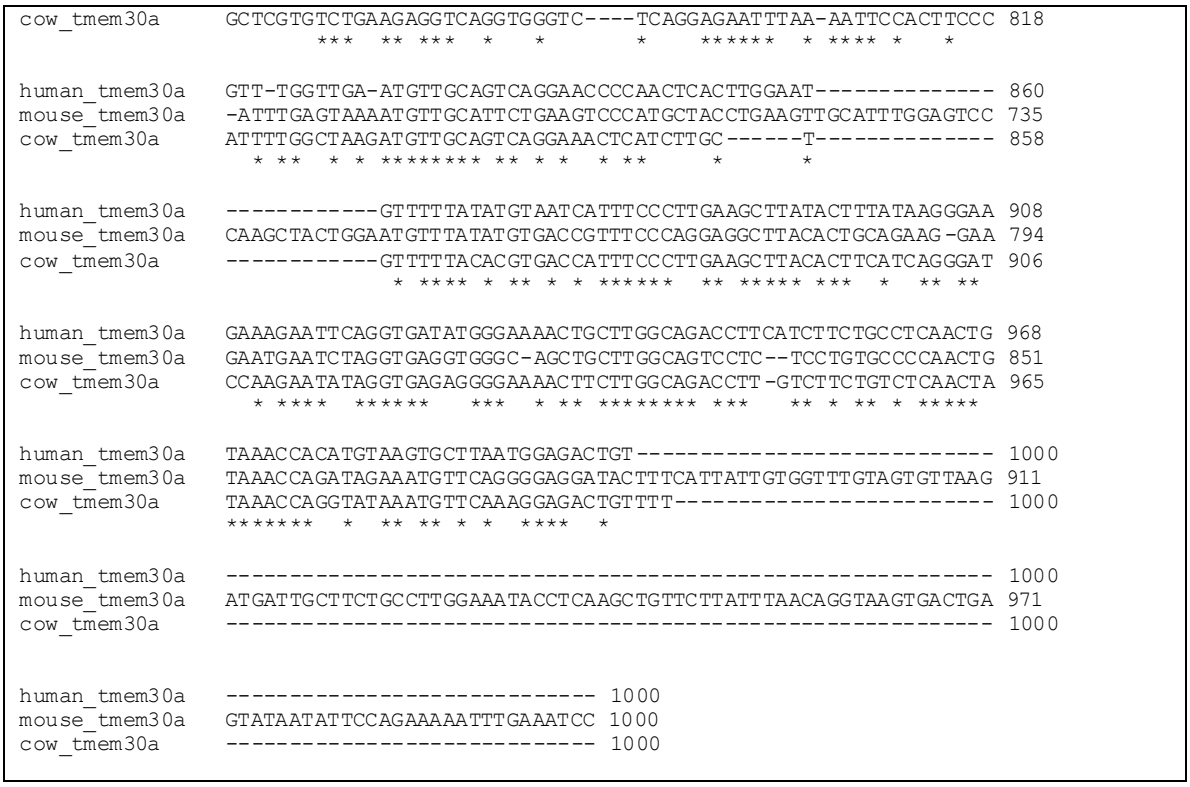
TMEM110:


human_tmem110 ---------- 0
mouse_tmem110 CGAGGAGGACCTCCGGAGACCTGTGAAAAAGAAGCACCGCTTCGGGCTGCCTGTATGACA 60
cow_tmem110 ---------T---------C---------------TC---T------------------ 5
platypus_tmem110 ---------C-ACA-----C---------------AC---C------------------ 8
human_tmem110 ---------- 0
mouse_tmem110 CATTCCCACGCTGCGGGTGACAGTCCTGGGGCCCAGCCCTGCAGCAACAGCGTCTCTGCC 120
cow_tmem110 ----GC-------------------C-----------C---------------CTCT--T 14
platypus_tmem110 ----TC-------------------C-----------CC-------------TCTCTCTT 21
human_tmem110 ------------------------------CCTCTTGCTGTCTCTGCCCGCTCTCGCCT 29
mouse_tmem110 CTCCTCTCTGCCCTCCTCTTCTACCTCTGCTCCTCCTGCTGTCTCTGCCCGCTCTCCACC 180
cow_tmem110 C---TACCT-----------------CCGCTCCTCCCGCGGCCTCGGCC----------- 43
platypus_tmem110 C---TCTCT-------------------------------------CCCATTCCTCCTTT 41
**
human_tmem110 GCCCCCAGACTACTGTGACTTAAAAAGAGGGAAGAGGAGCCAGCGCCCGAGGGGGCCAC- 88
mouse_tmem110 CGCCCCAGACGACTGTGACTTAAA-AGAGGGAAGAGGACCCA-TGCCCGAGGGGGCTGT- 237
cow_tmem110 GCTCCCA-GATACTGTGACCC-------GGGGAGGG-------GGGCCGCCGGCTCCACG 88
platypus_tmem110 TACCACCCTC-CCCACTTCCTG---A-TGGGAAGAGAACAGGTGATCTGAATGGGTCAC- 95
* * * * *** ** * * * *
human_tmem110 TGGCGGCTGGAGGTCCC-CATTCAGTTGCACTACAAA--CACTGACCAAATATGCAAGGA 145
mouse_tmem110 GGGCAGCTGGAG-TGCC-CACTCAGTTGCACTACAAA--CACTGACCAAATATGCAAGCG 293
cow_tmem110 TGGCAGCTGGAGGTCCC-CTCTCAGTTGCACTACAAA--CACTGACCAAATATGCAAGCG 145
platypus_tmem110 AGGCAGTTGAAGATTCTTAATTCAGTTGCACTACAAATATACTGACCAAATATGCAAGAG 155
*** * ** ** * * *** ********** ***********
human_tmem110 AGGAGCTGTGTTTGTTTGTTGTCGTCCCAGACAGTGTTGTGAGGGACCT-GAGGCCCTGC 204
mouse_tmem110 AAGAGCTTTGTTTGTTTGTTGTTGTCCCTTA-GGGGTTCTGTAAGACC------CCCTGT 346
cow_tmem110 AGGCGTTCTGTTTG--TGTTGTCGTCCCGGT-GGTGTTGCGAGGGACCCGGAAGCCCTGT 202
platypus_tmem110 AGGC-TTGTGTTTA--TCATCTTGTACGGAA-CTTCCGATGACGGTATT-GAGGACCAGG 210
* * * ***** * * * ** * * * * *
human_tmem110 CCCGTGTCCGACCACCGAGTGGCAAGGTGGA---AGGAAG-CACAGGCA-CACAGACCGT 259
mouse_tmem110 CC--TGTCTGACCA---GGTGGCATGGTGGA---AGGAAA-C---AGCA-CACAGACTGT 393
cow_tmem110 CCCGCGTCTTGCCGTGTGGCCTTGAGGCTG----AGGACT-TACCTGCA-CGTGGACCGT 256
platypus_tmem110 TTCGC-TGAGACGA----CTCACCT-GTGACTCGTTCACTCTGCAGGTGTTTTAGGCTTT 264
* * * * * * * * *
human_tmem110 GGGTGGGTCTCCTCACCGTGGCT---GTGGGCAGTGCGAACACATAACACCCTCG-GGCT 315
mouse_tmem110 TG-CGGGTCCTGGAGCTGAGGCT---GCCGGCCTTGTGGACAAGCGCTGTGCTTGAGGTT 449
cow_tmem110 GGGTGGAGCTTCTCGTCAGGGCTTCAGCGGGGAGTGTGAAGACGGAAAATGTTTGGGGGC 316
platypus_tmem110 GTGAGAAGCAGTCCTAT---------GTGGGGACCATCTGGAGGAAAAAAA-----AAAA 310
* * * * ** *
human_tmem110 AAAAGT-GACTC-GTTGAC-CAAGTTGGAACCGGAATGCTTTCTT-ACTCAAAATGGCTT 371
mouse_tmem110 AGCAGT-TGCTG-GTTGACCCAAGTTGGAACAGGAATGCTTTCTTAACTCAAAATGGCTT 507
cow_tmem110 T-GAGT-TACTC-ATTGAC-CAAGTTGGAGCTGTGACGCTTTCTT-ACTCGAAATGGCTT 371
platypus_tmem110 AAAAATCATCTGGACTGAC-CAAGATGGACCCAGTATGTTTTCTT-ACTCAAAGTGGCTT 368
* * ** **** **** **** * * * * ****** **** ** ******
human_tmem110 TTGTAACTATTGATTTCTGAAGCT----GGTTTTATGAGTTGTGACAGTG--TTACCAGG 425
mouse_tmem110 TTGTAACTCTTTATTTCTAAAGCC----AGTTTTATGAGCTGTGACAGTG--TTACTGTT 561
cow_tmem110 TTGTAACTATTGATTTCTATGGCT----GGTTTTATGAGCTGTCACAGTG--TTAACGTT 425
platypus_tmem110 TTGCAAC-ATTAATTCCTAGAGCGGGATAGTTTTATGGTTACAAGCTGTAACTGAATGTT 427
*** *** ** *** ** ** ******** * ** * *
human_tmem110 TTGGGGGTATGTGTTTATTTCCTACAAAGTACTTACGGGACTAATGGGCAAAACAGAGAT 485
mouse_tmem110 TTTTGAGTATGTGTTCATTTCCTACAAAGTATCTATGGGTCTAATGGGCAA-AC--AGAT 618
cow_tmem110 TTTTGAGTATGTGTTTATTTCCT-----------ATGGGACTAATGGGCAAAACATAGAT 474
platypus_tmem110 TTTAAAGTTTGTGTTTACTTCCTACAAGGTACCTATGGAACTAATAGGCAGAACAAAGAT 487
** ** ****** * ***** * ** ***** **** ** ****


human_tmem110 TTTTAAGTCTTCCGTATGCTGTTTGACTTTTATATTTTTAAGTTATATTTTCATACTATT 545
mouse_tmem110 TTTTAAGTCTTCCGTATACTGTTGGACTTTTATATTTTTAAGTTATATTTTCATACTATT 678
cow_tmem110 TTTTAAGTCTTCCGTATGCTGTT-GACTTTTATATGTTTAAGTTATATTTTCATACTATT 533
platypus_tmem110 TTTTAAGTCTTCTCGGTCCTGTTTGACTTTTATATTTTTAAGTTATATTTTCATACTGTT 547
human_tmem110 GTATTTAAAAACTCTTTTTA-G-------TCCC---CAAAGAAATGGGTTTATTTGCCTT 594
mouse_tmem110 GTATTTAAA-ACTCTTTTTA-A-------TCCC---CAAAGACATGGATTTGTTTGCCTT 726
cow_tmem110 GTATTTAAAAGCTCTTTTTT-A-------CCCC---CAAAGAAATGGGTTTGTTTGCCTT 582
platypus_tmem110 CTATTTAAGGACTTTGTTTTCCTCTCCTGCCTCCAGCAGAGAAAAAGATTCCCTAACCAT 607
human_tmem110 TCATGGGGTGTGGGCTGGCAGGAGG-----AAAAATCGGGAGTTTTTTA-TTTGGAATA- 647
mouse_tmem110 TCATGGGGAATGAGCTAGGGGGAGGTGGGTAAAAATGAGGAGTTTTGGAGGCAGGAATA- 785
cow_tmem110 TCATGGGACGTGAACTGGTGGGAGG----AAAAAGTTGAGAGTTTTTTTATTTGGAATA- 637
platypus_tmem110 TGTTTTGGATGGGGGGGTGGGGAGGGCACGGTAGGAGGAGAGATGTCACGGTCGGAATCG 667
human_tmem110 TTATTAGATGATGCCCTATGATAAGATGAGACAAATGATG--GGGAGGGAAGGAGGATGG 705
mouse_tmem110 TTGTTAAGTAACATCG-ATGGTGAGATAAACGGCCCTGGT--GGGAAGGTAGAGAGATAG 842
cow_tmem110 TCGTTAGGTAATCTGC---GATCCGATGCGGTAAATAATTCCAGTAGGAAGGGAGGTGGG 694
platypus_tmem110 GCGTCTTTTCATCTCC----------T-CGGTAAACAGCAAAGCCACAAAAGGAAG---- 712
human_tmem110 CCCTTCTCTAGAATCAGCAAACCCA----ATGGTTCTGTGAAG-GTCAGA-CCCAAGCTT 759
mouse_tmem110 CGCTTATCTGCGCTCA-CAGATCCA-----TA-----GTTAAG-GTCAGAACCTGCACAG 890
cow_tmem110 TCCTTATCTGAAATCAGCAAACCCAGTCAGTGGTTCTGCAGAG-GTCAGG-CCTGAGCAG 752
platypus_tmem110 CTTTCACCAGAAATTATGAAATAGGATTCCAT-----GTTAAGTGTGGGA-CACGAGGGG 766
human_tmem110 G-GAGCAGCTGGTCTGGATGCAGATCCAGGAGCTGCAGAGTGT--GGAACAGGAAAGGCT 816
mouse_tmem110 G-AGGCAGCTGC--------------------TCCAGTGTGG---GGACAGACAGAGCC 925
cow_tmem110 G-AGGCAGCTGGTCCAAATGTGGATCCAGG-GCCTGTGAGGGT--GAAACAGGAAGGGCT 808
platypus_tmem110 CTGGGCTGCATG--------CCCTACTCTGTGGCCCAGGGTCTTTGGAAAAACAGAAGGT 818
human_tmem110 CTGCCCAGGGCCCATGAGCTCTAGC------ATTCCTGCTGGCAGATTAGAGA----TCT 866
mouse_tmem110 CTGC-CAGGGCATACA--------------------AGCTGCCGTGTTGGAGG----CAA 960
cow_tmem110 CTGCCCAGAGCCCAGCAGCACTTGGCACT-TACTTCAGCTGGCAAGTTAGAG-----TCT 862
platypus_tmem110 GCTTCTGTAGATTTGAGCCTCATGC-ACTTTATCGTAAAAACCCAACTAGACTCCGGAGG 877
human_tmem110 TAGTTAAATTTGACCAAGAAAGGAGCTTAGCTAAGAGGTCTTTGTTTCCAGAGGACCCAA 926
mouse_tmem110 TGGTTAAACTTCACTC----AGGAGTT--------ACACGTTTGTGTCCAG--------- 999
cow_tmem110 TGGTGAAATGTGACTGAAAAGGGAGCA----TAAGAGGATATTGTGTCCAGAGAAATTGA 918
platypus_tmem110 TTGCAGGTTTTCAGCCCCTCAGGCGATCAGTC---ATGCAGTGACATTGACAGCACCTGC 934
human_tmem110 GACTACCAG---ACTCCTGT--GCAGTCTGCCTGTCTGCAGAGCCTCCAATCAC--GGTT 979
mouse_tmem110 ---------- 999
cow_tmem110 GACTCCAAGGCCACTCCTGTTTGTCCCCTGCATCTCTTCAGAACCTCCACATCTGAAGTT 978
platypus_tmem110 TGGATTCTAG--GCA-CTGG--GAATTCAACTTAC--AAAGAAG-------ACCG-ACTT 979
human_tmem110 TAAAGT--GGTCTTGGCCTCCCA 1000
mouse_tmem110 ----------A 1000
cow_tmem110 TAAAGCCCAGTCTTGGCCT-CCC 1000
platypus_tmem110 CCA--CCAAGAATTTGCAATCCT 1000

TMEM30A


human_tmem30a ---TCATGGATGGGAGGAAAAAATCCATTTTTGGGGATTGCTTACATCGCTGTTGGATCC 57
mouse_tmem30a ----------AC---------- 2
cow_tmem30a ATTTCATGGATGGGAGGAAAAAATCCATTTTTGGGGATTGCTTACATCACTATTGGATCC 60
human_tmem30a ATCTCCTTCCTTCTGGGAGTTGTACTGCTAGTAATTAATCATAAATATAGAAACAGTAGT 117
mouse_tmem30a ---------- 2
cow_tmem30a ATCTCCTTTCTTCTGGGAGTTGTACTGCTAGTAATTAATCATAAATATAGAAACAGTAGT 120
human_tmem30a AATACAGCTGACATTACCATTTAATTTTATATTATGAAAGCAAATCATCTGCATGTGCAT 177
mouse_tmem30a ----------ATCACCATTTAATTTTATATTCTGAAACCAAATCTACTGCATGTGCAT 50
cow_tmem30a AATACTGCTGACATTACCATTTAATTTTATATTCTGAAAACAAATTTACTGCATGTGCAT 180
human_tmem30a CAAGGCCAGTCCTATTCAACCTAGCTTTCGAATGCTGATA-TCTGGTTAGTATGTCATT- 235
mouse_tmem30a CAAGGCCAGTCCTGTTCAACCTAGCTTTTGAATGCTGATG-TCTGGTTAGTATGTCATT- 108
cow_tmem30a CAAGGCCAGTCCTATTCAACCTAGCTTTCAAATGCTGATGTTCTGGTTAGTATGTCATTT 240
human_tmem30a TTGAAGTTGGCACATAACTTTTCTAAAA---------AAAAGCAGTCTTTGTTGTTTGCT 286
mouse_tmem30a TTGAAGTTGGCACATAACTTTAAAAAAACAAAACAAACAAAACAGCCTTTGTTCTTTGCT 168
cow_tmem30a TTGAAGTTGGCACATAACATTTTTTAA----------AAAGCAGTCTTTGTCTTCTGCT 289
human_tmem30a TCTTCCCTACGGATGACTTCTAAAAATATATGACGGGTATAAAA-AAATTAGCTATATTG 345
mouse_tmem30a TCTTACATATGGATGACTTATGAAAATATATGATGGATATAA---AAATTAGCCATATTG 225
cow_tmem30a TCTTCCTTATGGATGACTTATGAAAATATATGACGGGTATAAAACAGATTAGCTATACTG 349
human_tmem30a ATCATATCAACACTGTAACTGCTGAAATGGCATTCTAATGTTTGCTTTTTATTCGGACAG 405
mouse_tmem30a ATTATATCAATATTGTAACTGCTAAAATGACATTCTAATGTCTGCTTTTTATTGGGACAG 285
cow_tmem30a ATCATATCAACACTGTAACTGTTGAAATGGCATTATGATGTTTCTTTTTTGTTGGGACAG 409
human_tmem30a GCCACATGATGCATAGAGCCTCTTTCATGTGACTTGTGTCTACTGCTTAAA-TCTTTATG 464
mouse_tmem30a GCCATGTGATGCATAGAGCCTCTTTCCTATGAAATGCGTCTACTGCTTAAACTGTTGATG 345
cow_tmem30a GCTGTGTGATGCATAGAGCCTCTTTCATGTGAAATGCGTCTGCTGCTTATA-TGTTTATG 468
human_tmem30a CTGTGTTGATGATATTATATTGACATATGAAGCTGTATATGTGTATGTATTTTGTGGAGA 524
mouse_tmem30a CTGTGTTGATAA---CATATTGACA--TGATGCTGTATATGTGTGCCTACTGTGTGATGA 400
cow_tmem30a CTGTGTTGATGATAATATATTGACATATGATACTGTATATGTGTGCCTGTTGTGTGAAGA 528
human_tmem30a AAGGGATTACAAGATGTATGAGTATAATGACTTGCTAACCTTTCAGGATTCAGAGAAAGA 584
mouse_tmem30a AAGGGATTATGAGATGTATGAGTGTAATGACTTGCTAACCTTTGAAAATTTGGTTACAGT 460
cow_tmem30a ACGGGATTACAAGATGTATGAGTATAACGACTTGCTAACCTTTCAGGATCCAGAGTTACA 588
human_tmem30a TGAAGA--AAGACCATATCTAAATAATACACTTCATCATTTTCATGT-----GTATAAAT 637
mouse_tmem30a TCAGATAGAAGAAAGACTATAAATAAAACACTTCATCATTTTCATGTGTCGTGTGTAAAG 520
cow_tmem30a AAAAGATGAAGACCAAATCTAAATAAAACACTTCATCATTTTCATGTGTCGTGTGTAAAG 648
human_tmem30a GCTTAAAGTACCATCTTTGTTGAGGTGGTTCATGTATCCAGTTTATCCAGTACAGTTATT 697
mouse_tmem30a GCTTAAAGTCCC-TCCTTGTTGAGGTGGTTCATATGTTCAGTTGCTCTATTATGA----T 575
cow_tmem30a GCTTAAAGTACCATCTTTGTTGAGGTGATTCCTGTATTCAGCTTATCCAGTACAA----T 704
human_tmem30a T-GTCAAGCTTAGCTTTGATTTCAAAGGACACGCTTACCTTGTCTGGCATAAGAATTAAT 756
mouse_tmem30a TCTCCGATAATGACGTTGACTTCA-----CAC-TTTAGCTTGTACAACATAGAAATTAAT 629
cow_tmem30a T-GTCAAGCTTAGATTTGATTTCAAAGGACATGCTTAACCTGGCTAGCATAGGAATTAAT 763
human_tmem30a GCTCATGTCTGCAGTGGTTGGGTAGGTCCTGCTTAGGAGAATTAAAAAATTCCTCTTTCC 816
mouse_tmem30a A------TCTAAAGAGGTCAGTGGGTCT----TGGCTAGAATTTTA-AATTTCTTCTC-- 676


cow_tmem30a GCTCGTGTCTGAAGAGGTCAGGTGGGTC----TCAGGAGAATTTAA-AATTCCACTTCCC 818
human_tmem30a GTT-TGGTTGA-ATGTTGCAGTCAGGAACCCCAACTCACTTGGAAT---------- 860
mouse_tmem30a -ATTTGAGTAAAATGTTGCATTCTGAAGTCCCATGCTACCTGAAGTTGCATTTGGAGTCC 735
cow_tmem30a ATTTTGGCTAAGATGTTGCAGTCAGGAAACTCATCTTGC------T---------- 858
human_tmem30a ----------GTTTTTATATGTAATCATTTCCCTTGAAGCTTATACTTTATAAGGGAA 908
mouse_tmem30a CAAGCTACTGGAATGTTTATATGTGACCGTTTCCCAGGAGGCTTACACTGCAGAAG-GAA 794
cow_tmem30a ----------GTTTTTACACGTGACCATTTCCCTTGAAGCTTACACTTCATCAGGGAT 906
human_tmem30a GAAAGAATTCAGGTGATATGGGAAAACTGCTTGGCAGACCTTCATCTTCTGCCTCAACTG 968
mouse_tmem30a GAATGAATCTAGGTGAGGTGGGC-AGCTGCTTGGCAGTCCTC--TCCTGTGCCCCAACTG 851
cow_tmem30a CCAAGAATATAGGTGAGAGGGGAAAACTTCTTGGCAGACCTT-GTCTTCTGTCTCAACTA 965
human_tmem30a TAAACCACATGTAAGTGCTTAATGGAGACTGT---------- 1000
mouse_tmem30a TAAACCAGATAGAAATGTTCAGGGGAGGATACTTTCATTATTGTGGTTTGTAGTGTTAAG 911
cow_tmem30a TAAACCAGGTATAAATGTTCAAAGGAGACTGTTTT---------- 1000
human_tmem30a ---------- 1000
mouse_tmem30a ATGATTGCTTCTGCCTTGGAAATACCTCAAGCTGTTCTTATTTAACAGGTAAGTGACTGA 971
cow_tmem30a ---------- 1000
human_tmem30a ---------- 1000
mouse_tmem30a GTATAATATTCCAGAAAAATTTGAAATCC 1000
cow_tmem30a ---------- 1000

Appendix F. Inequality to check if lmers in the triplet share at least one common motif

Initial derivation was published in [Ho et al 2009]. Minimum numbers of identical positions between each *l*mer in the triplet and the common motif is (*l*-*d*), *l* = length of motif, *d* = maximum numbers of mutations.

Let's denote the number of P_i, P_{mn} and P_{nc} patterns by $|P_i,|$, $|P_{mn}|$ and $|P_{nc}|$ respectively.

For *l*mer1, the number of identical positions must satisfy this:

$$l\text{-}d \le |P_{i,}| + |P_{12}| + |P_{13}| + |P_{\text{nc assign to } l\text{mer1}}|$$

Similarly, for *l*mer2 and *l*mer3, it will be:

$$l\text{-}d \le |P_{i,}| + |P_{12}| + |P_{23}| + |P_{\text{nc assign to } l\text{mer2}}|$$

$$l\text{-}d \le |P_{i,}| + |P_{13}| + |P_{23}| + |P_{\text{nc assign to } l\text{mer3}}|$$

These three inequalities must hold simultaneously, so we summarize them together into one inequality:

$$3(l\text{-}d) \le 3|P_{i,}| + 2|P_{12}| + 2|P_{13}| + 2|P_{23}| + |P_{\text{nc assign to } l\text{mer1}}| + |P_{\text{nc assign to } l\text{mer2}}| + |P_{\text{nc assign to } l\text{mer3}}|$$

Since $|P_{mn}| = |P_{12}| + |P_{13}| + |P_{23}|$, and $|P_{nc}| = |P_{\text{nc assign to } l\text{mer1}}| + |P_{\text{nc assign to } l\text{mer2}}| + |P_{\text{nc assign to } l\text{mer3}}|$, we can simplify the above inequality through these two substitutions and divide both sides by 3. Hence, the precondition for a triplet to share at least one common motif is:

$$l - d \leq \left|P_{i}\right| + \left|P_{mn}\right| * 2/3 + \left|P_{nd}\right| * 1/3$$

Appendix G. 61 rules to discover neighboring motifs

These rules were initially published in [Ho et al 2009]. Note: Rule IDs are not in consecutive order. For the description of operations, refer to Table 4.1 in chapter 4.

Rule ID	Operation	Impact on Score Vector
1	Sac(P_{12})	[-1,-1,+1]
2	Compl(P_{12})	[-1,-1,0]
3	Sac_sac(P_{12}, P_{13})	[-2,0,0]
4	sac_compl(P_{12}, P_{13})	[-2,-1,0]
5	Sac_sac(P_{12}, P_{23})	[0,-2,0]
6	sac_compl(P_{12}, P_{23})	[-1,-2,0]
7	Sac_nc(P_{12}, (1,2))	[-2,0,1]
8	Sac_nc(P_{12}, (1,3))	[-2,-1,2]
9	Sac_nc(P_{12}, (1,0))	[-2,-1,1]
10	Sac_nc(P_{12}, (2,1))	[0,-2,1]
11	Sac_nc(P_{12}, (2,3))	[-1,-2,2]
12	Sac_nc(P_{12}, (2,0))	[-1,-2,1]
13	Sac_nc(P_{12}, (3,1))	[0,-1,0]
14	Sac_nc(P_{12}, (3,2))	[-1,0,0]
15	Sac_nc(P_{12}, (3,0))	[-1,-1,0]
81	Nc(1,0)	[-1,0,0]
84	Nc(1,2)	[-1,1,0]
85	Nc(1,3)	[-1,0,1]

24	Sac(P_{13})	[-1,1,-1]
25	Compl(P_{13})	[-1,0,-1]
27	sac_compl(P_{13}, P_{12})	[-2,0,-1]
28	Sac_sac(P_{13}, P_{23})	[0,0,-2]
29	sac_compl(P_{13}, P_{23})	[-1,0,-2]
30	Sac_nc(P_{13}, (1,2))	[-2,2,-1]
31	Sac_nc(P_{13}, (1,3))	[-2,1,0]
32	Sac_nc(P_{13}, (1,0))	[-2,1,-1]
33	Sac_nc(P_{13}, (2,1))	[0,0,-1]
34	Sac_nc(P_{13}, (2,3))	[-1,0,0]
35	Sac_nc(P_{13}, (2,0))	[-1,0,-1]
36	Sac_nc(P_{13}, (3,1))	[0,1,-2]
37	Sac_nc(P_{13}, (3,2))	[-1,2,-2]
38	Sac_nc(P_{13}, (3,0))	[-1,1,-2]
82	Nc(2,0)	[0,-1,0]
86	Nc(2,1)	[1,-1,0]
87	Nc(2,3)	[0,-1,1]
48	Sac(P_{23})	[1,-1,-1]
49	Compl(P_{23})	[0,-1,-1]
51	sac_compl(P_{23}, P_{12})	[0,-2,-1]
53	sac_compl(P_{23}, P_{13})	[0,-1,-2]
54	Sac_nc(P_{23}, (1,2))	[0,0,-1]
55	Sac_nc(P_{23}, (1,3))	[0,-1,0]
56	Sac_nc(P_{23}, (1,0))	[0,-1,-1]
57	Sac_nc(P_{23}, (2,1))	[2,-2,-1]

58	Sac_nc(P_{23}, (2,3))	[1,-2,0]
59	Sac_nc(P_{23}, (2,0))	[1,-2,-1]
60	Sac_nc(P_{23}, (3,1))	[2,-1,-2]
61	Sac_nc(P_{23}, (3,2)	[1,0,-2]
62	Sac_nc(P_{23}, (3,0))	[1,-1,-2]
83	Nc(3,0)	[0,0,-1]
88	Nc(3,1)	[1,0,-1]
89	Nc(3,2)	[0,1,-1]
71	Sac_sac(P_{12})	[-2,-2,0]
72	Sac_sac(P_{13})	[-2,0,-2]
73	Sac_sac(P_{23})	[0,-2,-2]
74	Sac_i_nc(P_i,(1,2))	[-2,0,-1]
75	Sac_i_nc(P_i,(1,3))	[-2,-1,0]
76	Sac_i_nc(P_i,(2,1))	[0,-2,-1]
77	Sac_i_nc(P_i,(2,3))	[-1,-2,0]
78	Sac_i_nc(P_i,(3,1))	[0,-1,-2]
79	Sac_i_nc(P_i,(3,2))	[-1,0,-2]
80	Sac_i(P_i)	[-1,-1,-1]

List of rules to test when the i-th lmer has excess score, each has 42 rules.

1st lmer	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,24,25,27,28,29,30,31,32, 33,34,35,36,37,38,71,72,73,74,75,76,77,78,79,80,81,84,85

2nd lmer	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,48,49,51,53,54,55,56,57, 58,59,60,61,62,28,71,72,73,74,75,76,77,78,79,80,82,86,87
3rd lmer	24,25,3,27,28,29,30,31,32,33,34,35,36,37,38,48,49,5,51,53, 54,55,56,57,58,59,60,61,62,71,72,73,74,75,76,77,78,79,80,83,88,89

Appendix H. Simulation data

This part was originally published in [Ho et al 2009]. We generated multiple sets of simulated sequences according to the <*l,d*> motif model formulated by Pevzner and Sze [Pevzer et al 2000]. Each dataset consists of 20 sequences, each 600 nucleotides long. All nucleotides occur equally likely. In each sequence, a single *l*-size *d*-mutant is planted at a random location. We have prepared datasets for a wide range of <*l,d*> motif models, i.e. <11,2>, <12,3>, <13,3>, <14,4>, <15,4>, <16,5>, <18,6>, <19,6>, <24,8>, <28,8>, <30,9>, <38,12>and <40,12>.

Appendix I. Run-time performance

This part was originally published in [Ho et al 2009]. We compared the performance of our method with three other methods with the same enumerative design philosophy, viz. MotifEumerator, RISOTTO and PMSprune. Source codes were downloaded from these sites, MotifEnumerator from http://faculty.cs.tamu.edu/shsze/motifenumerator/, RISOTTO from http://kdbio.inesc-id.pt/~asmc/pub/software/RISO/riso-me-src.zip, and PMSprune from http://www.engr.uconn.edu/~jid02003/Jaime/pmsprune.c. They were compiled in the Linux x86 platform according to the instructions documented in the respective websites.

Appendix J. Untranslated region sequence data

This part was originally published in [Ho et al 2009]. In addition to simulated data, we also prepared and tested several sets of real biological data that can be split into two groups, one 5' upstream of the start codon, i.e. 5' UTR and promoter; and the other from the 3' UTR. For the 5' UTR-promoter group, we chose four genes that are commonly tested in other motif finding algorithms [Blanchette et al 2002, Estkin et al 2002, Davila et al 2007], namely, preproinsulin, DHFR, metallothionine, and c-fos. Homologous regions from four species were included for analysis using the Homologene database (http://www.ncbi.nlm.nih.gov/sites/entrez?db=homologene) from NCBI. To obtain the upstream promoter region, BLAT [Kent 2002] was used to map the cDNA to the species genome provided by Genome browser [Kent et al 2002]. Based on the 5' starting point of the cDNA, we then extracted promoter sequence from the genome.

Another set of real biological sequences is taken from the 3' UTR where AU-rich elements (AREs), cytoplasmic polyadenylation elements (CPEs), and Pumillio binding elements (PBEs) were chosen. The AREs were derived from 30 experimentally validated human and mouse 3' UTRs [Chen et al 1995]. These genes were also confirmed by the ARE database ARED 2.0 (http://brp.kfshrc.edu.sa/ARED/) [Bakheet et al 2001]. Based on the accession numbers provided by ARED, we retrieved the cDNA sequences from NCBI's RefSeq database [Pruitt et al 2007]. The 5' end of the 3'UTR begins right after the stop codon, However, the 3' end of the 3' UTR is not obvious because we

have found that most of the cDNA sequences deposited in RefSeq database lacking a poly(A) tail. In order to accurately determine the 3' end of the 3’ UTR, we utilized expressed sequence tag (EST) data from the UCSC Genome Browser [Kent et al 2002]. We first mapped each cDNA to the genome using BLAT. The true end of the 3’ UTR should coincide with the endpoint of the EST. The more ESTs that end at the same spot as the cDNA, the higher confidence we have about the true end of the 3’ UTR. The set of sequences we obtained are variable in length ranging from 92 to 1608 bases bringing the total sequence space to 23,022 bases. For CPEs and PBEs, we have used the five cyclin genes, B1, B2, B3, B4, and B5 from *Xenopus laevis* [Pique et al 2008].

Appendix K. Sensitivity and specificity test

This part was originally published in [Ho et al 2009]. For the sensitivity and specificity test, we adopted the three-level (nucleotide, binding site, and motif) testing framework proposed by the Kihara group [Hu et al 2005]. Two sets of data, ECRDB70 and ECRDB62A, were downloaded from their website (http://dragon.bio.purdue.edu/pmotif). These data were originally derived from the RegulonDB database [Salgado et al 2004]. The ECRDB62A dataset comprises 713 intergenic sequences containing binding sites for 62 transcription factors in E. Coli K-12. We filtered out duplicated sequences, transformed reverse strands into forward direction, and dropped transcription factors with less than three binding sequences. The final reconstructed dataset contains 379 distinct sequences from 36 transcription factors. At the nucleotide and binding site level, four different assessments were performed. Sensitivity (*Sn*) is defined as $^{TP}/_{TP+FN}$, where *TP*, *FN* stands for true positive and false negative respectively. Specificity (*Sp*) is defined as $^{TP}/_{TP+FP}$, where *FP* is false positive. We followed two other assessments that were described in [Hu et al 2005] to combine *Sn* and *Sp*. Performance coefficient (*PC*) is defined as $^{TP}/_{TP+FP+FN}$, which was originally proposed in [Pevzner et al 2000,Tompa et al 2005]. The last assessment is called *F*-measure (*F*), which tends to penalize the imbalance of *Sn* and *Sp*. *F* is defined as $^{2*Sn*Sp}/_{Sn+Sp}$. Both *PC* and *F* fall into the range of [0,1], with value 1 indicating perfect prediction. In addition to the nucleotide and binding site levels, the Kihara group proposed two other

accuracy measurements viz. sequence accuracy (*sSr*) and motif accuracy (*mSr*). *sSr* is defined as Ns/N where *Ns* is the number of sequences having their motifs correctly predicted, and *N* is the total number of binding sequences of a transcription factor. The overall *sSr* is the average *sSr* of all transcription factors. *mSr* is defined as Np/M, where *Np* is the number of transcription factors with at least one correctly predicted binding site in the binding sequence set and *M* is the total number of transcription factors in the dataset. We compared our method with WEEDER [Pavesi et al 2004] and the top three best-performing methods previously evaluated in [Hu et al 2005]: MEME, BioProspector and MotifSampler. MEME, BioProspector, MotifSampler and WEEDER were download from http://meme.nbcr.net/downloads/, http://motif.stanford.edu/distributions/bioprospector/, http://homes.esat.kuleuven.be/~thijs/download/linux_3.2/MotifSampler and http://159.149.109.9/modtools/downloads/weeder1.3.1.tar.gz respectively. For WEEDER, we specified the organism to be E. Coli K12 “BEC” and the type of analysis “large”.

Appendix L. Transfection and Luciferase assays

This part was originally published in [Ho et al 2009]. Cell culture and transfections were done as previously described in [Goraczniak et al 2008]. For Luciferase assays, the cells were harvested after 24 hours and Luciferase measured using the Promega dual Luciferase kit (Promega, Madison, WI) measured on a Turner BioSystems Luminometer (Turner BioSystems, Sunnyvale, CA).

Appendix M. Alignment of mammalian PAS flanking regions

Due to data volume, unfiltered alignment data is accessible on the web only. To access, visit:

http://www.rci.rutgers.edu/~gundersn/conserved/human_mouse_cow_upstream_CFs.rpt.gz (5.9M)

http://www.rci.rutgers.edu/~gundersn/conserved/human_mouse_cow_platypus_upstream_CFs.rpt.gz (6.8M)

REFERENCES

Ahituv N, Zhu Y, Visel A, Holt A, Afzal V, Pennacchio LA, Rubin EM. 2007. Deletion of ultraconserved elements yields viable mice. PLoS Biol. 2007 Sep;5(9):e234.

Aravin AA, Lagos-Quintana M, Yalcin A, Zavolan M, Marks D, Snyder B, Gaasterland T, Meyer J, Tuschl T. 2003. The small RNA profile during Drosophila melanogaster development. Dev Cell. 2003 Aug;5(2):337-50.

Bailey TL, Elkan C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, pp. 28-36.

Bailey TL, Elkan C. 1995. Unsupervised Learning of Multiple Motifs in Biopolymers using EM. Machine Learning, 21(1-2):51-80.

Bakheet T, Williams BR, Khabar KS. 2006. ARED 3.0: the large and diverse AU-rich transcriptome. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D111-4.

Bartel DP, Chen CZ. 2004. Micromanagers of gene expression: the potentially widespread influence of metazoan microRNAs. Nat Rev Genet. 2004 May;5(5):396-400.

Beaudoing E, Freier S, Wyatt JR, Claverie JM, Gautheret D. 2000. Patterns of variant polyadenylation signal usage in human genes. Genome Res. 2000 Jul;10(7):1001-10.

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. 2004. Ultraconserved elements in the human genome. Science. 2004 May 28;304(5675):1321-5.

Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, Jaenisch R, Wagschal A, Feil R, Schreiber SL, Lander ES. 2006. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell. 2006 Apr 21;125(2):315-26.

Blanchette M, Tompa M. 2002. Discovery of regulatory elements by a computational method for phylogenetic footprinting. Genome Research, 5, 739-48.

Boelens WC, Jansen EJ, van Venrooij WJ, Stripecke R, Mattaj IW, Gunderson SI. 1993. The human U1 snRNP-specific U1A protein inhibits polyadenylation of its own pre-mRNA. Cell. 1993 Mar 26;72(6):881-92.

Brown JM, Bell TA 3rd, Alger HM, Sawyer JK, Smith TL, Kelley K, Shah R, Wilson MD, Davis MA, Lee RG, Graham MJ, Crooke RM, Rudel LL. 2008. Targeted depletion of hepatic ACAT2-driven cholesterol esterification reveals a non-biliary route for fecal neutral sterol loss. J Biol Chem. 2008 Apr 18;283(16):10522-34.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics. 2009 Dec 15;10:421.

Chen CY, Gherzi R, Ong SE, Chan EL, Raijmakers R, Pruijn GJ, Stoecklin G, Moroni C, Mann M, Karin M. 2001. AU binding proteins recruit the exosome to degrade ARE-containing mRNAs. Cell. 2001 Nov 16;107(4):451-64.

Chen CY, Shyu AB. 1995. AU-rich elements: characterization and importance in mRNA degradation. Trends Biochem Sci., 11, 465-70.

Chen CZ, Li L, Lodish HF, Bartel DP. 2004. MicroRNAs modulate hematopoietic lineage differentiation. Science. 2004 Jan 2;303(5654):83-6.

Chen F, Wilusz J. 1998. Auxiliary downstream elements are required for efficient polyadenylation of mammalian pre-mRNAs. Nucleic Acids Research. 1998 Jun 15;26(12):2891-8.

Cheng Y, Miura RM, Tian B. 2006. Prediction of mRNA polyadenylation sites by support vector machine. Bioinformatics. 2006 Jul 26.

Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature. 2005 Sep 1;437(7055):69-87.

Chuvpilo S, Zimmer M, Kerstan A, Glöckner J, Avots A, Escher C, Fischer C, Inashkina I, Jankevics E, Berberich-Siebelt F, Schmitt E, Serfling E. 1999. Alternative polyadenylation events contribute to the induction of NF-ATc in effector T cells. Immunity. 1999 Feb;10(2):261-9.

Cohen SN. 1995. Surprises at the 3' end of prokaryotic RNA. Cell. 1995 Mar 24;80(6):829-32.

Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: A sequence logo generator. Genome Research, 14, 1188-1190.

Czyzyk-Krzeska MF, Bendixen AC. 1999. Identification of the poly(C) binding protein in the complex associated with the 3' untranslated region of erythropoietin messenger RNA. Blood. 1999 Mar 15;93(6):2111-20.

Dalziel M, Nunes NM, Furger A. 2007. Two G-rich regulatory elements located adjacent to and 440 nucleotides downstream of the core poly(A) site of the

intronless melanocortin receptor 1 gene are critical for efficient 3' end processing. Mol Cell Biol. 2007 Mar;27(5):1568-80.

Danckwardt S, Kaufmann I, Gentzel M, Foerstner KU, Gantzert AS, Gehring NH, Neu-Yilik G, Bork P, Keller W, Wilm M, Hentze MW, Kulozik AE. 2007. Splicing factors stimulate polyadenylation via USEs at non-canonical 3' end formation signals. EMBO J. 2007 Jun 6;26(11):2658-69.

Danckwardt S, Gehring NH, Neu-Yilik G, Hundsdoerfer P, Pforsich M, Frede U, Hentze MW, Kulozik AE. 2004. The prothrombin 3'end formation signal reveals a unique architecture that is sensitive to thrombophilic gain-of-function mutations. Blood. 2004 Jul 15;104(2):428-35. Epub 2004 Apr 1.

Danckwardt S, Hartmann K, Katz B, Hentze MW, Levy Y, Eichele R, Deutsch V, Kulozik AE, Ben-Tal O. 2006. The prothrombin 20209 C-->T mutation in Jewish-Moroccan Caucasians: molecular analysis of gain-of-function of 3' end processing. J Thromb Haemost. 2006 May;4(5):1078-85.

Davila J, Balla S, Rajasekaran S. 2007. Fast and practical algorithms for planted (l, d) motif search. IEEE/ACM Trans Computational Biology & Bioinformatics, 4, 544-52.

De Val S, Chi NC, Meadows SM, Minovitsky S, Anderson JP, Harris lS, Ehlers ML, Agarwal P, Visel A, Xu SM, Pennacchio LA, Dubchak I, Krieg PA, Stainier DY, Black BL. 2008. Combinatorial regulation of endothelial gene expression by ets and forkhead transcription factors. Cell. 2008 Dec 12;135(6):1053-64.

Edmonds M. 2002. A history of poly A sequences: from formation to factors to function. Prog Nucleic Acid Res Mol Biol. 2002;71:285-389.

Edwards AO, Ritter R 3rd, Abel KJ, Manning A, Panhuysen C, Farrer LA. 2005. Complement factor H polymorphism and age-related macular degeneration. Science. 2005 Apr 15;308(5720):421-4. Epub 2005 Mar 10.

Eskin E, Pevzner PA. 2002. Finding composite regulatory patterns in DNA sequences. Bioinformatics, 18 Suppl 1, S354-63.

Fairbrother WG, Yeh RF, Sharp PA, Burge CB. 2002. Predictive identification of exonic splicing enhancers in human genes. Science. 2002 Aug;297(5583):1007-13.

Fritsche LG, Loenhardt T, Janssen A, Fisher SA, Rivera A, Keilhauer CN, Weber BH. 2008. Age-related macular degeneration is associated with an unstable ARMS2 (LOC387715) mRNA. Nat Genet. 2008 Jul;40(7):892-6.

Gehring NH, Frede U, Neu-Yilik G, Hundsdoerfer P, Vetter B, Hentze MW, Kulozik AE. 2001. Increased efficiency of mRNA 3' end formation: a new genetic

mechanism contributing to hereditary thrombophilia. Nat Genet. 2001 Aug;28(4):389-92.

van Gelder CW, Gunderson SI, Jansen EJ, Boelens WC, Polycarpou-Schwarz M, Mattaj IW, van Venrooij WJ. 1993. A complex secondary structure in U1A pre-mRNA that binds two molecules of U1A protein is required for regulation of polyadenylation. EMBO J. 1993 Dec 15;12(13):5191-200.

Goraczniak R, Gunderson SI. 2008. The regulatory element in the 3'-untranslated region of human papillomavirus 16 inhibits expression by binding CUG-binding protein 1. J. Biol Chem. 2008 Jan 25;283(4):2286-96. Epub 2007 Nov 27.

Goraczniak R, Behlke MA, Gunderson SI. 2009. Gene silencing by synthetic U1 adaptors. Nat Biotechnol. 2009 Mar;27(3):257-63.

Graber JH, Cantor CR, Mohr SC, Smith TF. 1999. In silico detection of control signals: mRNA 3'-end-processing sequences in diverse species. Proc Natl Acad Sci U S A. 1999 Nov 23;96(24):14055-60.

Graveley BR. 2000. orting out the complexity of SR protein functions. RNA. 2000 Sep;6(9):1197-211.

Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. 2008. miRBase: tools for microRNA genomics. NAR 2008 36(Database Issue):D154-D158.

Gromak N, West S, Proudfoot NJ. 2006. Pause sites promote transcriptional termination of mammalian RNA polymerase II. Mol Cell Biol. 2006 May;26(10):3986-96.

Guan F, Caratozzolo RM, Goraczniak R, Ho ES, Gunderson SI. 2007. A bipartite U1 site represses U1A expression by synergizing with PIE to inhibit nuclear polyadenylation. RNA. 2007 Dec;13(12):2129-40.

Gunderson SI, Beyer K, Martin G, Keller W, Boelens WC, Mattaj IW. 1994. The human U1A snRNP protein regulates polyadenylation via a direct interaction with poly(A) polymerase. Cell. 1994 Feb 11;76(3):531-41.

Gunderson SI, Vagner S, Polycarpou-Schwarz M, Mattaj IW. 1997. Involvement of the carboxyl terminus of vertebrate poly(A) polymerase in U1A autoregulation and in the coupling of splicing and polyadenylation. Genes Dev. 1997 Mar 15;11(6):761-73.

Hageman GS, Anderson DH, Johnson LV, Hancox LS, Taiber AJ, Hardisty LI, Hageman JL, Stockman HA, Borchardt JD, Gehrs KM, Smith RJ, Silvestri G, Russell SR, Klaver CC, Barbazetto I, Chang S, Yannuzzi LA, Barile GR, Merriam JC, Smith RT, Olsh AK, Bergeron J, Zernant J, Merriam JE, Gold B, Dean M, Allikmets R. 2005. A common haplotype in the complement regulatory gene

factor H (HF1/CFH) predisposes individuals to age-related macular degeneration. Proc Natl Acad Sci U S A. 2005 May 17;102(20):7227-32.

Haines JL, Hauser MA, Schmidt S, Scott WK, Olson LM, Gallins P, Spencer KL, Kwan SY, Noureddine M, Gilbert JR, Schnetz-Boutaud N, Agarwal A, Postel EA, Pericak-Vance MA. 2005. Complement factor H variant increases the risk of age-related macular degeneration. Science. 2005 Apr 15;308(5720):419-21. Epub 2005 Mar 10.

He L, Söderbom F, Wagner EG, Binnie U, Binns N, Masters M. 1993. PcnB is required for the rapid degradation of RNAI, the antisense RNA that controls the copy number of ColE1-related plasmids. Mol Microbiol. 1993 Sep;9(6):1131-42.

Hinman MN, Lou H. 2008. Diverse molecular functions of Hu proteins. Cell Mol Life Sci. 2008 Oct;65(20):3168-81.

Ho ES, Jakubowski CD, Gunderson SI. 2009. iTriplet, a rule-based nucleic acid sequence motif finder. Algorithms Mol Biol. 2009 Oct 29;4:14.

HomoloGene. 2009. NCBI HomoloGene database build 63. ftp://ftp.ncbi.nih.gov/pub/HomoloGene/build63/

Hu J, Lutz CS, Wilusz J, Tian B. 2005. Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation. RNA. 2005 Oct;11(10):1485-93.

Hu J, Li B, Kihara D. 2005. Limitations and potentials of current motif discovery algorithms. Nucleic Acids Research, 33(15), 4899-913.

Hutchins LN, Murphy SM, Singh P, Graber JH. 2008. Position-dependent motif characterization using non-negative matrix factorization. Bioinformatics. 2008 Dec 1;24(23):2684-90.

Jensen KL, Styczynski MP, Rigoutsos I, Stephanopoulos GN. 2006. A generic motif discovery algorithm for sequential data. Bioinformatics. 2006 Jan 1;22(1):21-8.

Ji Z, Lee JY, Pan Z, Jiang B, Tian B. 2009. Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. Proc Natl Acad Sci U S A. 2009 Apr 16.

Karnik P, Taljanidisz J, Sasvari-Szekely M, Sarkar N. 1987. 3'-terminal polyadenylate sequences of Escherichia coli tryptophan synthetase alpha-subunit messenger RNA. J Mol Biol. 1987 Jul 20;196(2):347-54.

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., Weber, R.J., Haussler,

D. and Kent, W.J. 2003. The UCSC Genome Browser Database. Nucleic Acids Research, 31(1), 51-54.

Kaufmann I, Martin G, Friedlein A, Langen H, Keller W. 2004. Human Fip1 is a subunit of CPSF that binds to U-rich RNA elements and stimulates poly(A) polymerase. EMBO J. 2004 Feb 11;23(3):616-26.

Kent WJ. 2002. BLAT - the BLAST-like alignment tool. Genome Res. 2002 Apr;12(4):656-64.

Kim HS, Kuwano Y, Zhan M, Pullmann R Jr, Mazan-Mamczarz K, Li H, Kedersha N, Anderson P, Wilce MC, Gorospe M, Wilce JA. 2007. Elucidation of a C-rich signature motif in target mRNAs of RNA-binding protein TIAR. Mol Cell Biol. 2007 Oct;27(19):6806-17.

Kuersten S, Goodwin EB. 2003. The power of the 3' UTR: translational control and development. Nat Rev Genet. 2003 Aug;4(8):626-37.

Labombarda F, González SL, Lima A, Roig P, Guennoun R, Schumacher M, de Nicola AF. 2009. Effects of progesterone on oligodendrocyte progenitors, oligodendrocyte transcription factors, and myelin proteins following spinal cord injury. Glia. 2009 Jun;57(8):884-97.

Lareau LF, Inada M, Green RE, Wengrod JC, Brenner SE. 2007. Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. Nature. 2007 Apr 19;446(7138):926-9.

Lee TI, Jenner RG, Boyer LA, Guenther MG, Levine SS, Kumar RM, Chevalier B, Johnstone SE, Cole MF, Isono K, Koseki H, Fuchikami T, Abe K, Murray HL, Zucker JP, Yuan B, Bell GW, Herbolsheimer E, Hannett NM, Sun K, Odom DT, Otte AP, Volkert TL, Bartel DP, Melton DA, Gifford DK, Jaenisch R, Young RA. 2006. Control of developmental regulators by Polycomb in human embryonic stem cells. Cell. 2006 Apr 21;125(2):301-13.

Legendre M, Gautheret D. 2003. Sequence determinants in human polyadenylation site selection. BMC Genomics. 2003 Feb 25;4(1):7.

Matlin AJ, Clark F, Smith CW. 2005. Understanding alternative splicing: towards a cellular code. Nat Rev Mol Cell Biol. 2005 May;6(5):386-98.

Meyers CD. 2001. Matrix Analysis and Applied Linear Algebra Book and Solutions Manual. SIAM: Society for Industrial and Applied Mathematics (February 15, 2001) ISBN 0898714540.

Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, Castle J, Bartel DP, Linsley PS, Johnson JM. 2005. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature. 2005 Feb 17;433(7027):769-73.

Liu D, Fritz DT, Rogers MB, Shatkin AJ. 2008. Species-specific cis-regulatory elements in the 3'-untranslated region direct alternative polyadenylation of bone morphogenetic protein 2 mRNA. J Biol Chem. 2008 Oct 17;283(42):28010-9.

Liu X, Brutlag DL, Liu JS. 2001. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Pac Symp Biocomput., 6:127-38.

Liu X, Brutlag DL, Liu JS. 2002. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. Nature Biotechnology, 20, 835-9.

Lu QR, Cai L, Rowitch D, Cepko CL, Stiles CD. 2001. Ectopic expression of Olig1 promotes oligodendrocyte formation and reduces neuronal survival in developing mouse cortex. Nat Neurosci. 2001 Oct;4(10):973-4.

Mandel CR, Kaneko S, Zhang H, Gebauer D, Vethantham V, Manley JL, Tong L.2006. Polyadenylation factor CPSF-73 is the pre-mRNA 3'-end-processing endonuclease. Nature. 2006 Dec 14;444(7121):953-6.

Mandel CR, Bai Y, Tong L. 2008. Protein factors in pre-mRNA 3'-end processing. Cell Mol Life Sci. 2008 Apr;65(7-8):1099-122.

Mayr C, Bartel DP. 2009. Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. Cell. 2009 Aug 21;138(4):673-84.

Ni JZ, Grate L, Donohue JP, Preston C, Nobida N, O'Brien G, Shiue L, Clark TA, Blume JE, Ares M Jr. 2007. Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. Genes Dev. 2007 Mar 15;21(6):708-18.

Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol. 2000 Sep 8;302(1):205-17.

Pavesi G, Mereghetti P, Mauri G, Pesole G. 2004. Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. Nucleic Acids Research. 2004 Jul 1;32(Web Server issue):W199-203.

Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, Plajzer-Frick I, Akiyama J, De Val S, Afzal V, Black BL, Couronne O, Eisen MB, Visel A, Rubin EM. 2006. In vivo enhancer analysis of human conserved non-coding sequences. Nature. 2006 Nov 23;444(7118):499-502.

Perez Canadillas JM, Varani G. 2003. Recognition of GU-rich polyadenylation regulatory elements by human CstF-64 protein. EMBO J. 2003 Jun 2;22(11):2821-30

Pevzner PA, Sze SH. 2000. Combinatorial approaches to finding subtle signals in DNA sequences. Proc Int Conf Intell Syst Mol Biol. 8, 269-78.

Phillips C, Pachikara N, Gunderson SI. 2004. U1A inhibits cleavage at the immunoglobulin M heavy-chain secretory poly(A) site by binding between the two downstream GU-rich regions. Mol Cell Biol. 2004 Jul;24(14):6162-71.

Piqué M, López JM, Foissac S, Guigó R, Méndez R. 2008. A combinatorial code for CPE-mediated translational control. Cell, 132(3), 434-48.

Pisanti N, Carvalho AM, Marsan L, Oliveira AL, Sagot MF. 2006. RISOTTO: Fast extraction of motifs with mismatches. Proceedings of the 7th Latin American Theoretical Informatics Symposium, 3887, 757-768.

Portnoy V, Schuster G. 2006. RNA polyadenylation and degradation in different Archaea; roles of the exosome and RNase R. Nucleic Acids Res. 2006;34(20):5923-31.

Pruitt KD, Tatusova, T, Maglott DR. 2007. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Research, 35 (Database issue):D61-5

Rat Genome Sequencing Project Consortium. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428, 493-521 (1 April 2004).

Rajasekaran S, Balla S, Huang CH. 2005. Exact algorithms for planted motif problems. Journal Computational Biology, 8, 1117-28.

Rajasekaran S, 2006. Algorithms for motif search. Handbook of Computational Biology edited by Srinivas Aluru, Chapman & Hall/CRC pp 37:1-21.

Rigoutsos I, Floratos A. 1998. Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. Bioinformatics. 1998;14(1):55-67.

Roth FP, Hughes JD, Estep PW, Church GM. (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nature Biotechnology, 10, 939-45.

Rott R, Zipor G, Portnoy V, Liveanu V, Schuster G. 2003. RNA polyadenylation and degradation in cyanobacteria are similar to the chloroplast but different from Escherichia coli. J Biol Chem. 2003 May 2;278(18):15771-7.

Sachchithananthan M, Stasinopoulos SJ, Wilusz J, Medcalf RL. 2005. The relationship between the prothrombin upstream sequence element and the G20210A polymorphism: the influence of a competitive environment for mRNA 3'-end formation. Nucleic Acids Res. 2005 Feb 17;33(3):1010-20.

Salgado, H., Gama-Castro, S., Martinez-Antonio, A., Diaz-Peredo, E., Sanchez-Solano, F., Peralta-Gil, M., Garcia-Alonso, D., Jimenez-Jacinto, V., Santos-Zavaleta, A., Bonavides-Martinez, C., Cllado-Vides, J. 2004. RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in Escherichia coli K-12 Nucleic Acids Research, 32, D303–D306

Salisbury J, Hutchison KW, Graber JH. 2006. A multispecies comparison of the metazoan 3'-processing downstream elements and the CstF-64 RNA recognition motif. BMC Genomics. 2006 Mar 16;7:55.

Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB. 2008. Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. Science. 2008 Jun 20;320(5883):1643-7.

Sarkar N. 1997. Polyadenylation of mRNA in prokaryotes. Annu Rev Biochem. 1997;66:173-97.

Segal E, Widom J. 2009. Poly(dA:dT) tracts: major determinants of nucleosome organization. Curr Opin Struct Biol. 2009 Feb;19(1):65-71.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001 Jan 1;29(1):308-11.

Siddharthan R, Siggia ED, van Nimwegen E. 2005. PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. PLoS Computational Biology, 7, e67.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005 Aug;15(8):1034-50.

Silver Key SC, Pagano JS. 1997. A noncanonical poly(A) signal, UAUAAA, and flanking elements in Epstein-Barr virus DNA polymerase mRNA function in cleavage and polyadenylation assays. Virology. 1997 Jul 21;234(1):147-59.

Slomovic S, Laufer D, Geiger D, Schuster G. 2005. Polyadenylation and degradation of human mitochondrial RNA: the prokaryotic past leaves its mark. Mol Cell Biol. 2005 Aug;25(15):6427-35.

Slomovic S, Portnoy V, Liveanu V, Schuster G. 2006. RNA Polyadenylation in prokaryotes and organelles; Different tails tell different tales. Critical reviews in plant sciences, 25:65-77, 2006.

Smit, AFA, Hubley, R, Green, P. 1996-2004. RepeatMasker Open-3.0. http://www.repeatmasker.org

Spies N, Nielsen CB, Padgett RA, Burge CB. 2009. Biased chromatin signatures around polyadenylation sites and exons. Mol Cell. 2009 Oct 23;36(2):245-54.

Sze SH, Zhao X. 2006. Improved pattern-driven algorithms for motif finding in DNA sequences. Proceedings of the 2005 Joint RECOMB Satellite Workshops on Systems Biology and Regulatory Genomics, Lecture Notes in Bioinformatics, 4023, 198-211.

Tabaska JE, Zhang MQ. 1999. Detection of polyadenylation signals in human DNA sequences. Gene. 1999 Apr 29;231(1-2):77-86.

Takagaki Y, Ryner LC, Manley JL. 1989. Four factors are required for 3'-end cleavage of pre-mRNAs. Genes Dev. 1989 Nov;3(11):1711-24.

Takagaki Y, Manley JL. 1997. RNA recognition by the human polyadenylation factor CstF. Mol Cell Biol. 1997 Jul;17(7):3907-14.

Takagaki Y, Manley JL. 1998. Levels of polyadenylation factor CstF-64 control IgM heavy chain mRNA accumulation and other events associated with B cell differentiation. Mol Cell 2(6):761–771.

Taljanidisz J, Karnik P, Sarkar N. 1987. Messenger ribonucleic acid for the lipoprotein of the Escherichia coli outer membrane is polyadenylated. J Mol Biol. 1987 Feb 5;193(3):507-15.

Thijs G, Marchal K, Lescot M, Rombauts S, De Moor B, Rouzé P, Moreau Y. 2002. A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. Journal of Computational Biology. 9, 447-64.

Tian B, Hu J, Zhang H, Lutz CS. 2005. A large-scale analysis of mRNA polyadenylation of human and mouse genes. Nucleic Acids Res. 2005 Jan 12;33(1):201-12.

Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, Makeev VJ, Mironov AA, Noble WS, Pavesi G, Pesole G, Régnier M, Simonis N, Sinha S, Thijs G, van Helden J, Vandenbogaert M, Weng Z, Workman C, Ye C, Zhu Z. 2005. Assessing computational tools for the discovery of transcription factor binding sites. Nature Biotechnology. 23(1):137-44.

Uitte de Willige S, de Visser MC, Houwing-Duistermaat JJ, Rosendaal FR, Vos HL, Bertina RM. 2005. Genetic variation in the fibrinogen gamma gene increases the risk for deep venous thrombosis by reducing plasma fibrinogen gamma' levels. Blood. 2005 Dec 15;106(13):4176-83.

Uitte de Willige S, Rietveld IM, De Visser MC, Vos HL, Bertina RM. 2007. Polymorphism 10034C>T is located in a region regulating polyadenylation of FGG transcripts and influences the fibrinogen gamma'/gammaA mRNA ratio. J Thromb Haemost. 2007 Jun;5(6):1243-9.

Varani L, Gunderson SI, Mattaj IW, Kay LE, Neuhaus D, Varani G. 2000. The NMR structure of the 38 kDa U1A protein - PIE RNA complex reveals the basis of cooperativity in regulation of polyadenylation by human U1A protein. Nat Struct Biol. 2000 Apr;7(4):329-35.

Venkataraman K, Brown KM, Gilmartin GM. 2005. Analysis of a noncanonical poly(A) site reveals a tripartite mechanism for vertebrate poly(A) site recognition. Genes Dev. 2005 Jun 1;19(11):1315-27.

de Vries H, Ruegsegger U, Hubner W, Friedlein A, Langen H, Keller W. 2000. Human pre-mRNA cleavage factor II(m) contains homologs of yeast proteins and bridges two other cleavage factors. EMBO J. 2000 Nov 1;19(21):5895-904.

Waggoner SA, Liebhaber SA. 2003. Regulation of alpha-globin mRNA stability. Exp Biol Med (Maywood). 2003 Apr;228(4):387-95.

Wang J, Hannenhalli S. 2005. Generalizations of Markov model to characterize biological sequences. BMC Bioinformatics. 2005 Sep 6;6:219.

Wang Z, Rolish ME, Yeo G, Tung V, Mawson M, Burge CB. 2004. Systematic identification and analysis of exonic splicing silencers. Cell. 2004 Dec 17;119(6):831-45.

Wickens M, Bernstein DS, Kimble J, Parker R. 2002. A PUF family portrait: 3'UTR regulation as a way of life. Trends Genet. 2002 Mar;18(3):150-7.

Wingender E, Dietze P, Karas H, Knüppel R. 1996. TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic Acids Research, 24 (1):238-41

Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K, Walter K, Abnizova I, Gilks W, Edwards YJ, Cooke JE, Elgar G. 2005. Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol. 2005 Jan;3(1):e7.

Wu C, Orozco C, Boyer J, Leglise M, Goodale J, Batalov S, Hodge CL, Haase J, Janes J, Huss JW 3rd, Su AI. 2009. BioGPS: an extensible and customizable

portal for querying and organizing gene annotation resources. Genome Biol. 2009;10(11):R130. Epub 2009 Nov 17.

Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M. 2005. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature. 2005 Mar 17;434 (7031):338-45.

Xu F, Cohen SN. 1995. RNA degradation in Escherichia coli regulated by 3' adenylation and 5' phosphorylation. Nature. 1995 Mar 9;374(6518):180-3.

Xu F, Lin-Chao S, Cohen SN. 1993. The Escherichia coli pcnB gene promotes adenylylation of antisense RNAI of ColE1-type plasmids in vivo and degradation of RNAI decay intermediates. Proc Natl Acad Sci U S A. 1993 Jul 15;90(14):6756-60.

Xu FF, Gaggero C, Cohen SN. 2002. Polyadenylation can regulate ColE1 type plasmid copy number independently of any effect on RNAI decay by decreasing the interaction of antisense RNAI with its RNAII target. Plasmid. 2002 Jul;48(1):49-58.

Yeap BB, Voon DC, Vivian JP, McCulloch RK, Thomson AM, Giles KM, Czyzyk-Krzeska MF, Furneaux H, Wilce MC, Wilce JA, Leedman PJ. 2002. Novel binding of HuR and poly(C)-binding protein to a conserved UC-rich motif within the 3'-untranslated region of the androgen receptor messenger RNA. J Biol Chem. 2002 Jul 26;277(30):27183-92.

Zhao J, Hyman L, Moore C. 1999. Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. Microbiol Mol Biol Rev. 1999 Jun;63(2):405-45.

Zhu H, Zhou HL, Hasman RA, Lou H. 2007. Hu proteins regulate polyadenylation by blocking sites containing U-rich sequences. J Biol Chem. 2007 Jan 6;282(4):2203-10.

CURRICULUM VITAE

Eric Sau-chum Ho

EDUCATION

1987	B.Sc.	National University of Singapore
1993	M.Sc.	The HK University of Science & Technology
2010	Ph.D.	Rutgers, The State University of New Jersey - New Brunswick

PUBLICATIONS

Peer Reviewed Journal Articles

Ho ES, Jakubowski CD, Gunderson SI. iTriplet, a rule-based nucleic acid sequence motif finder. *Algorithms Mol Biol.* 2009 Oct 29;4(1):14. (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2784457/)

Li Y, **Ho ES**, Gunderson SI, Kiledjian M. Mutational analysis of a Dcp2-binding element reveals general enhancement of decapping by 5'-end stem-loop structures. *Nucleic Acids Res.* 2009 Apr;37(7):2227-37.

Guan F, Caratozzolo RM, Goraczniak R, **Ho ES**, Gunderson SI. A bipartite U1 site represses U1A expression by synergizing with PIE to inhibit nuclear polyadenylation. *RNA.* 2007 Dec;13(12):2129-40.

Major Conference Presentations

Ho ES, Yang E, Gunderson SI, Androulakis I. Degenerative Sequence Motifs Identification AIChE 2006 Annual Meeting San Francisco USA (talk, http://aiche.confex.com/aiche/2006/preliminaryprogram/abstract_60932.htm)

Ho ES, Gunderson SI. Identification of Mammalian Polyadenylation Sites Using Logistic Regression ISMB ECCB 2009 Stockholm Sweden (peer reviewed poster, http://www.iscb.org/uploaded/css/43/11929.pdf)