UNIVERSITY OF TAMPERE

Development of CAbase and an Exon Analysis Pipeline for Visual assessment of Predicted Genes for the Carbonic Anhydrases

Master's Thesis

Lydia Isokangas
30 May, 2016

Acknowledgements

I am extremely thankful for my experiences in Finland and for the quiet joy of the Finnish people. It is truly incredible that the Finnish people and government are so willing to share so much with foreigners from all over the world!

Personally I would like to express my gratitude for my supervisor Martti Tolvanen who has taught me the skill of listening faster. Martti is the only person I know who can squeeze three facts in one sentence and still make sense. My thanks also extends to Seppo Parkilla for his interesting journal club meetings and infectious enthusiasm for the CAs, Catarina Ståhle-Nieminen for her constant and kind motivation for me via email and Matti Nykter who has incredibly fast email return times even when you are on the other side of the planet.

For all of my friends at university – you know who you are - thanks for all the good times and sharing of the pain!

And thanks to my children, Alex, Catherine and Andrew, who have understood that sometimes I have to study. Special thanks goes to my husband, Erik who has picked up everything that I have dropped with a smile and encouraged me to pursue just one more degree. And finally, thanks also to my mother for her unending enthusiasm for my studies and also to my mother-in-law for her loving support.

Master's Thesis

Place:	University of Tampere School of Medicine BioMediTech
Author:	Lydia Ruth Catherine Isokangas
Title:	Development of CAbase and an Exon Analysis Pipeline for Visual assessment of Predicted Genes for the Carbonic Anhydrases
Pages:	137
Supervisors:	Doctor Martti Tolvanen Professor Seppo Parkkila
Reviewers:	Professor Matti Nykter Doctor Martti Tolvanen
Date:	30 May, 2016

Abstract

Background and Aims

Humans are able to quickly recognize and evaluate visual patterns, thus this thesis aims to apply this feature to the analysis of aspects of the conservation of carbonic anhydrase proteins. This was facilitated through the creation of two pipelines:

- One to create a publically available specialized database to service the CA research world named CAbase, and,
- One to create a visual display of the aligned exons of the cDNA transcripts contained within CAbase with indicators to show the positions of start and stop codons along with the locations of the predicted signal and mitochondrial targeting peptides. This pipeline was named Exon_Analysis.

Carbonic anhydrases (CAs) are ubiquitous proteins that reversibly catalyse carbon dioxide into carbonic acid. Through the events of duplication, the CAs exist in at least 16 different isoforms and potentially up to 17 different isoforms.

Methods

The pipelines were created using freely available tools that included python, MySQL, various bioinformatic tools such as Clustal Omega, PRANK, BLAST and Pal2Nal.

The data for CAbase was extracted from Ensembl, NCBI, UniProt, RSCB PDB, UniGene and FlyBase. Additionally, calculated data from using SignalP and TargetP was also included. CAbase is hosted on the Amazon Web Server and can be accessed using any computer that has access to the Internet and has MySQL installed.

Exon_Analysis draws a scaled exon MSA schematic based on a PRANK MSA of the cDNA transcripts for a CA isoform. The exons and other indicators such as the start and stop codon, and the signal and target peptides are all drawn in different colours

and in their scaled locations. Thus it is possible to see the conserved nature of the exons within the coding regions and the aligned start and stop codons and the peptides for each CA isoform.

Results

CAbase is now publically available for anyone to use. However, it is still somewhat user unfriendly due to the requirement that user be familiar with SQL. CAbase facilitated the use of Exon_Analysis. This pipeline has enabled the quality control of the predicted genes one exon at a time. Evolutionary events such as the conservation of short exons within CAs VI, IX, XII, XIV, X and XI has been shown in the exon MSA schematics. Both tools could be adapted for use with other proteins.

Conclusion

CAbase is a specialized database for CA researchers that can continue to be adapted to their needs. It allows researchers to create specific queries without having to filter for unwanted proteins. Exon_Analysis creates an exon MSA schematic to visualise the relationship between the exons and other features of interest. This facilitates the quality control of predicted genes one exon at a time.

Abbreviations

BLAST	Basic Local Alignment Search Tool
CA	Carbonic Anhydrase
CARP	Carbonic Anhydrase Related Protein
CDS	Coding DNA sequence
cDNA	complementary DNA
DNA	Deoxyribonucleic Acid
EBI	European Bioinformatics Institute
EST	Expressed Sequence Tag
GPI	Glycophosphatidylinositol
HMM	Hidden Markov Model
HSP	High-scoring Segment Pair
MSA	Multiple Sequence Alignment
PDB	Protein Data Bank
PG	Proteoglygan (domain)
PRANK	Phylogeny-aware progressive alignment method
PTX	Pentraxin domain
RC	Reliability Class - given by TargetP
RNA	Ribonucleic Acid
SQL	Structured Query Language
TM	Transmembrane (domain)
TSL	Transcript Support Level

1 Introduction

The alpha carbonic anhydrases all catalyse the reversible reaction between carbon dioxide and carbonic acid. This reaction is essential to maintain the acid/base balance for all life forms. The CAs most commonly contain a zinc ion to facilitate this reaction, except for the CARPs which are highly conserved but lack the metal cofactor.

The alpha CAs exist in five subgroups; the cytoplasmic CAs (CAI, CAII, CAIII, CAVII and CAXIII), the membrane associated CAs (CAIV, CAIX, CAXII, CAXIV and CAXV) the CARPs (CAVIII, CAX and CAXI), the mitochondrial CAs (CAVA and CAVB) and the secreted CA, CAVI.

Within each of the groups certain features are conserved such as the numbers of exons in the coding regions, the existence of mitochondrial or signal peptides etc. A tool for visualising these features via the exon MSA schematic has been developed for this thesis.

2 Literature Review

2.1 Carbonic Anhydrases

In 1932 Brinkman et al. identified Carbonic anhydrase (CA) as a catalyst in solutions of haemoglobin in cow's blood and muscles. (Brinkman, Margaria, Meldrum, & Roughton, 1932) Since this early discovery knowledge of the carbonic anhydrases (CAs) has expanded greatly; we now know that CAs can be grouped into three major classes: the alpha, beta and gamma CAs. The alpha CAs are found in mammals, the beta CAs are found in plants and some prokaryotes and the gamma CAs are found in archaeabacteria. There are two other minor groups, the δ and ξ groups that have been found in diatoms. (Dutta & Goodsell, 2004) (Aggarwal, Boone, Kondeti, & McKenna, 2012) The existence of so many major and minor CA groups points to the possibility of independent evolution of the same function, i.e. convergent evolution. (Aggarwal, Boone, Kondeti, & McKenna, 2012)

All CAs are metalloenzymes that contain a zinc ion to rapidly and reversibly catalyze carbon dioxide into carbonic acid where the zinc ion binds the hydroxide ion (OH^-). (Brinkman, Margaria, Meldrum, & Roughton, 1932)

$$CO_2 + H_2O \leftrightharpoons HCO_3^- + H^+$$

CAs are found ubiquitously throughout life, and indeed in many forms throughout eukaryotic bodies. Within vertebrates, alpha CAs have evolved to over 10 different isoforms that are involved in different pathways or tissues. (Tolvanen M. , 2013) For example, CA1 is involved with blood pH regulation in mammals, thus ensuring that under the conditions of homeostasis the animal receives the signals to breathe and respiratory alkalosis does not occur. CA9 is found in the gastric mucosa and gall bladder in humans and is involved with maintaining the pH balance in those organs.

(Swenson, et al., 1991) (Morgan, Pastorekova, Stuart-Tilley, Alper, & Casey, 2007) The EBI Gene Expression Atlas (http://www.ebi.ac.uk/gxa/home) and the Kyoto Encyclopedia of Genes and Genomes: KEGG (http://www.genome.jp/kegg/kegg2.html) has a very thorough database of the expression and metabolic pathways of the carbonic anhydrases in many different life forms.

This thesis focuses on the alpha carbonic anhydrases. Within this group there are four subgroups: cytoplasmic, membrane, secreted, CARPs and mitochondrial CAs. The focus of the following sections will be on the human CA isoforms simply because there is more research on this area, except for CA XV which is not found in primates and CA XVII which is not found in mammals.

2.1.1 Cytoplasmic CAs

Cytoplasmic CAs function within the cell membrane. They lack a signal or mitochondrial targeting peptide. This group includes carbonic anhydrase I, II, III, VII and XIII. In humans CA I, CA II, CA III and CA XIII are all located on chromosome 8. Lowe found that *CA1* was created through a duplication event that then was duplicated again to produce *CA2* and *CA3*. *CA2* and *CA3* are transcribed in the opposite direction to CA1. (Lowe, Edwards, Edwards, & PH, 1991) The data extracted from Ensembl into CAbase show that the arrangement of *CA1*, *CA2* and *CA3* on chromosome 8 is a common theme amongst primates.

In a sequence study (Table 1) by Hassan et al. the similarity between the cytoplasmic CAs are high relative to the similarities between other isoforms such as the secreted CAs. This suggests that these CAs deviated relatively recently compared to the other CAs. (Hassan, Shajee, Waheed, Ahmad, & Sly, 2012) CA VII is not as closely evolutionary linked to CAs I to III and it is located on chromosome 16.

Table 1 : Sequence similarities (%) among human CAs calculated by using ClustalW2. The cytoplasmic CAs are coloured in blue, the mitochondrial CA in green and the membrane associated CAs in pink. (Hassan, Shajee, Waheed, Ahmad, & Sly, 2012)

CAs	CA I	CA II	CA III	CA IV	CA VA	CA VB	CA VI	CA VII	CA VIII	CA IX	CA X	CA XI	CA XII	CA XIII	CA XIV
CA I	100														
CA II	60	100													
CA III	53	58	100												
CA IV	30	33	31	100											
CA VA	47	50	45	23	100										
CA VB	46	52	43	23	58	100									
CA VI	31	33	32	26	27	24	100								
CA VII	50	56	49	31	48	49	34	100							
CA VIII	39	40	38	27	33	34	28	41	100						
CA IX	31	33	20	25	27	26	35	35	31	100					
CA X	26	28	27	23	24	24	19	27	27	20	100				
CA XI	25	30	28	22	24	26	18	26	28	21	50	100			
CA XII	35	34	31	26	27	25	32	37	28	36	21	19	100		
CA XIII	59	59	57	28	46	47	33	52	39	34	27	27	34	100	
CA XIV	34	35	34	25	29	25	32	35	27	37	19	23	40	37	100

2.1.2 Membrane CAs

There are five known membrane associated CAs (CA IV, CA IX, CA XII and CA XIV in humans with CA XV expressed in non-primates). Tolvanen et al. recently proposed another membrane CA, CA XVII that is not expressed in mammals. (Tolvanen, et al., 2012) CAs IX, XII and XIV all have between 289 to 410 amino acids which reside outside the cell, about 22 amino acids in the transmembrane region and about 24 to 26 amino acids that reside inside the cell. (Barker, 2013)

CA IV encodes a glycosylphosphatidyl-inositol anchor (GPI anchor) in the C-terminal of the protein that binds the CA to the cell membrane. Like most other CAs, the first exon encodes the signal peptide. Interestingly, CA4 is the principal sensor that allows mammals to taste sourness and CO_2 in carbonated beverages. (Chandrashekar, et al., 2009) (Hassan, Shajee, Waheed, Ahmad, & Sly, 2012)

CA IX is a tumour associated CA. Like CA IV, CA IX also encodes a signal peptide but then, unlike CA IV, a proteoglycan related region in the first exon follows it on chromosome 17. (Opavsky, et al., 1996) The proteoglycan region is followed by a highly conserved catalytic domain in exons 2 to 8 while exons 10 and 11 contain the sequence for the transmembrane region. CA9 most likely arose from exon shuffling events as indicated by the relationship between the predicted protein domains and the organization of exons. CA9 is thought to be one of the earliest CAs to belong to the animal alpha carbonic anhydrase group. (Opavsky, et al., 1996) (Bocchini & McKusick, 2014) Opavsky theorised that the proteoglycans in CAs might be involved in the modulating cell interactions while Supuran et al postulate that the proteoglycan might be involved with the processes of cell adhesion and differentiation. (Supuran, Scozzafava, & Conway, 2004) (Opavsky, et al., 1996)

CA XII, like CA IX, is a membrane CA that is found in tumours. It is found on chromosome 15, has a signal peptide in exon 1 and is encoded in 11 exons. (Supuran, Scozzafava, & Conway, 2004)

CA XIV has high sequence identity with the other extracellular CAs found in primates (CA XII, CA IX, CA VI and CA IV). It has a signal peptide in exon one, followed by the highly conserved active site and then the transmembrane site. Phylogenetic studies by Mori et al show that *CA14* is most closely related to *CA12*, and then followed by *CA9*, *CA6* and *CA4*. Mori et al. found that *CA14* is active on the plasma membrane whereas *CA9* is bound to either the plasma membrane or the nuclear membrane. (Mori, et al., 1999)

Mouse CA XV was found to be phylogenetically similar to mouse CA IV by Hilvo et.al. The phylogenetic studies also suggest that CA XV was the first alpha CA to be expressed in vertebrates, but not in primates. Like CA IV, CA XV has a GPI anchor, and performs similar functions to CA IV in the body. Hilvo et al. proposed that since CA IV has low activity in mice, CA XV is necessary to make up the shortfall in activity. Conversely, CA IV has high activity levels in primates that perhaps made CA XV redundant. (Hilvo, et al., 2005)

Wide ranging phylogenetic studies of GPI linked CAs conducted by Tolvanen et al. show that a duplication of an ancestor CA4 gave rise to CA15 in early vertebrates. Thereafter, another duplication event created CA17 sometime before the fish and tetrapods split. Mammals have lost CA17 but it is still expressed in fish and non-mammalian tetrapods. Like other membrane CAs, CA17 has a predicted signal peptide and a GPI anchor. It has not been included in this study as it has not yet been included in Ensembl as a gene.

2.1.3 Mitochondrial CAs

There are two mitochondrial carbonic anhydrases, *CA5A* and *CA5B*. Both contain the mitochondrial targeting peptide in the first exon. Hassan et al. postulate that the two isoforms were created by a duplication event. Phylogenetically the *CA5* and *CA7* proteins are the most closely related and deviated somewhere after the oldest CAs, *CA4* and *CA6* but before *CA1*, *CA2* and *CA3*. *CA5A* can be found on chromosome 16 in humans while *CA5B* is found on the X chromosome. (Hassan, Shajee, Waheed, Ahmad, & Sly, 2012)

2.1.4 Secreted CAs

The only CA that is secreted is *CA6*. It has a signal peptide in the first exon and no membrane anchoring. Along with *CA4*, *CA6* is one of the oldest mammalian CAs. In non-mammals *CA6* has a pentraxin domain encoded in the C-terminal of the protein. The function of this domain is unknown. (Patrikainen, 2012)

2.1.5 CARP

The three remaining CAs are the CA-related proteins VIII, X and XI. They do not express any catalytic activity because they lack one or more histidine residues that bind the zinc atom. As yet their functions remain unknown. (Aspatwar, 2014)

2.2 Tools

The analysis for this thesis was run on a MacBook Pro laptop using python 2.7 with both online and locally installed tools. The programs used would work equally well on a machine running any Linux based operating system.

2.2.1 Locating proteins in the cell using TargetP, SignalP and related tools

Olof Emanuelsson, Søren Brunak, Gunnar von Heijne & Henrik Nielsen

Emanuelsson et al developed two tools, SignalP and TargetP, to predict the existence and location of the cleavage sites of signal, mitochondrial and chloroplast targeting peptides. Knowledge of these peptides gives an indication of the sub cellular localisation of the protein and therefore points the way towards understanding the function of the protein. Both SignalP and TargetP are freely available on the Center for Biological Sequence Analysis' (CBS) website to use online or download (http://www.cbs.dtu.dk/services/TargetP/ and http://www.cbs.dtu.dk/services/SignalP-4.1/). As this thesis does not study plant proteins, no further mention will be made of the chloroplast targeting peptide.

Of the three N-terminal peptides the signal peptides are the best known and the most conserved with a typical length of 15 to 30 amino acids. Despite this there is no consensus sequence for the signal peptide. These peptides are characterised by

positively charged residues at the N-terminal, followed by a region with at least six hydrophobic residues and ending with polar uncharged residues at the -3 to -1 positions relative to the cleavage site.

SignalP outputs three values: the c-score, s-score and y-score. The s-score indicates where the signal peptide is and it is less than the values for the cleavage site (c and y scores). The c-score is the position of the raw cleavage site. The final cleavage site is found by using the average of the c-score and the steepest slope of the s-score and is contained in the y-score. These values are used in the exonAnalysis.py program and are graphed in the exon MSA schematic with arrows indicating in which exon the peptides are found.

The mitochondrial targeting peptide is not as well conserved as the signal peptide with lengths ranging from 6 to 85 residues. Most commonly the peptide is rich in Arg, Ser and Ala and with negatively charged residues rarely found. Very little sequence conservation has been found around the cleavage site; the most common motif has been an Arg in the -2 or -3 position. If the proteins are analysed post cleavage the evolutionary history of the mitochondria can be revealed. If SignalP is used on the proteins post cleavage, it predicts a signal peptide like signal reminiscent of the prokaryotic origins of the mitochondria.

The only value from TargetP that is graphed in the exon MSA schematic is TPlen - the predicted pre-sequence length. It is indicated with an orange arrow in the exon MSA schematic. When comparing the values of TPlen and y-score, they are typically within one or two amino acids of each other.

SignalP uses a combination of neural networks (NNs) and Hidden Markov Models (HMMs) to predict the existence and location of signal peptides. TargetP uses NNs to calculate the signal peptide score (SP score) and then a weight matrix to locate the cleavage site. With the greater conservation of the signal peptides and using default D-score values in the program, SignalP has a false positive rate of 1.2%. (Emanuelsson, Brunak, von Heijne, & Nielsen, 2007) TargetP achieves a false positive rate of 3% for a reliability class (RC) of 1 and a false positive rate of 47% when RC is 5. (Emanuelsson, Brunak, von Heijne, & Nielsen, 2007) The authors of these two programs explain that this discrepancy is due to the mitochondrial targeting peptides being less conserved than the signal peptides.

2.2.2 A model of evolution and structure for multiple sequence alignment

Ari Löytynoja and Nick Goldman

This paper examined a method for multiple sequence alignment that considered both the common biological functions and the evolutionary relationships between the sequences i.e. aligning both the homologous residues and the conserved structures thus avoiding some of the systematic errors of other alignment programs. Others had tried to combine the alignment of the structure of sequences and the homologous residues in the past, but this approach had been considered too computationally intensive or had not been able to be applied to enough sequences to be useful.

Löytynoja et al. used a two level HMM to implement a pairwise alignment algorithm that treated insertions and deletions as evolutionary distinct events.

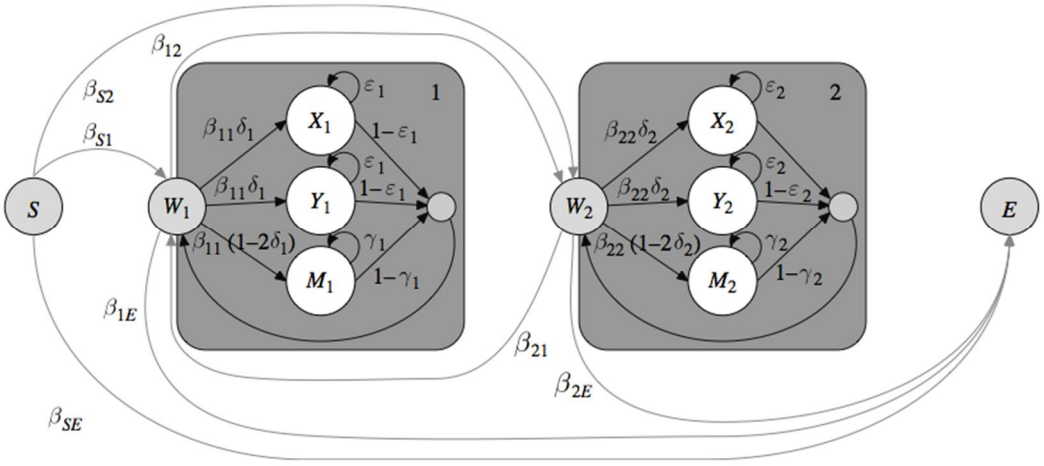

Figure 1 : The two level HMM with the light grey non-emitting Start and End states (S and E), and the two structure classes (grey boxes 1 and 2). The two structure classes are used to describe an evolutionary process. The grey and black arrows show the moves between and within the structure classes. The non-emitting, empty, grey circles are dummy states to clarify the moves between the emitting character states (X, Y and M) back to the non-emitting linker states W_j. (Löytynoja & Goldman, 2008)

Like all HMM's this model begins and ends with the start state S and the end state E, and has two or more middle states, which in this case are the structure states (grey boxes 1 and 2 in Figure 1). Within the structure state there are three emitting states:

- X_h that emits a character against a gap. The probability of staying in this state is ε_h, the probability of moving to the wait state W_h is $1-\varepsilon_h$, and the probability of entering this state from the wait state is $\beta_h\delta_h$.
- Y_h that emits a gap against a character. The probabilities of moving within the structure state are the same as that for the state X_h.
- M_h is the state that emits a character match. Here the probability of staying in the state is γ_h, the the probability of moving to the wait state W_h is $1-\gamma_h$, and the probability of entering this state from the wait state is $\beta_h(1-2\delta_h)$.

The three states within a "*structure class describes the character matching process within a given evolutionary process*" (Löytynoja & Goldman, 2008) (3914). The probabilities of movements between structure classes are predefined and fixed (β_{hn}). These structure class movements are used to distinguish between a slowly evolving site and a quickly evolving site; and also for detecting structures such as codons and exon boundaries. The phylogenetic awareness of this method is encoded in the value of δ_h which is defined by the deletion rate r_h and evolutionary time as the distance from the ancestral node to the two child node s of the sequences being aligned (known or calculated).

The results of using this method on genomic sequences show the alignment is able to align the protein coding exons and the insertions and deletions in the correct phase. When PRANK was compared against other alignment methods such as Clustalo and Clustalw, PRANK gave consistently more meaningful results without having to adjust the criteria for each individual protein or transcript. The other methods would align

some parts of the sequences well, but then introduce meaningless gaps to align fragments and thus mess up the phase of the sequences. This method was coded into a freely available program named PRANK, which can be downloaded from http://wasabiapp.org/software/prank/.

2.2.3 Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega

Fabian Sievers, Andreas Wilm, David Dineen, Toby J Gibson, Kevin Karplus, WeizhongLi, Rodrigo Lopez, Hamish McWilliam, Michael Remmert, Johannes Söding, Julie D Thompson, Desmond G Higgins

Clustal Omega is a protein-aligning tool that uses guide trees and the HHalign package to produce fast and accurate alignments. The authors of this package however admit that in tests that their model while being fast produces alignments that have comparable accuracy to other slower methods. While producing results for this thesis it was found to introduce errors when aligning all the known proteins for a CA. However when calculating small alignments, such as one translated protein against the full protein it was accurate and faster than PRANK. (Sievers, et al., 2011)

2.2.4 Basic local alignment search tool

Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J.

In this thesis the Basic Local Alignment Tool (BLAST) URL API, provided natively with Biopython 2.7, was used to search for potential replacement sequences from NCBI's Nucleotide and EST databases for sequences found to deviate from the exon structure of the majority of other CA sequences in the conserved exon schematic MSAs.

NCBI's BLAST uses heuristics to speed the search for the requested sequence(s). These heuristics can be understood as three steps:

1. After the BLAST algorithm reads the query, search parameters and the database, it breaks the query up into smaller 'words' made up of the areas of low complexity or repeats, which is used to find other areas of local similarity in the potential subject sequences.
2. If the algorithm finds two words on the same diagonal within a certain distance of each other, then the words are extended to test whether they are part of a high scoring alignment. The high scoring sequence pairs (HSPs) with scores above a threshold are used as seeds for gapped extensions that are again scored. If the extension alignment score drops below the threshold value, it is abandoned in favour of other sequences that have higher scores, thus reducing the search area.
3. Finally, the saved gapped extensions from the previous step are used as seeds to calculate the insertions and deletions and the E-value. If the E-value is low enough and the score high enough BLAST will return those sequences as results to the user. (Altschul, Gish, Miller, Myers, & Lipman, 1990)

2.2.5 Ensembl REST API: Ensembl Data for Any Language

Andrew Yates, Kathryn Beal, Stephen Keenan, William McLaren, Miguel Pignatelli, Graham R.S. Ritchie, Magali Ruffier, Kieron Taylor, Allessandro Vullo and Paul Flicek

The Ensembl Representational State Transfer (REST) API allows the use of Python to extract Ensembl data in the JSON and FASTA formats for easy interpretation and use. Ensembl data forms the core of the data for CAbase, thus making the Ensembl REST API a critical tool for this thesis. The REST API is free to access from http://rest.ensembl.org. (Yates, et al., 2015)

2.2.6 Python

Python Software Foundation

Apart from querying CAbase in SQL, the bulk of the coding for this thesis was done using Python 2.7 and the module Biopython. Python is freely available from python.org, which also publishes comprehensive and freely available online manuals. (Python Software Foundation, n.d.)

2.2.7 PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments

Mikita Suyama, David Torrents, Peer Bork

This tool was used to create a codon alignment of the protein and the cDNA sequence of a transcript. The codon alignment was used to calculate the corresponding locations of the signal or targeting peptide indicators in the DNA sequence so they could be drawn in the exon schematic MSA. Pal2Nal accepts both FASTA and CLUSTAL format sequences as inputs.

Pal2Nal works by reverse translating the protein sequence into regular expressions to deal with the redundancy of bases coding for the amino acids e.g. Arginine is encoded by CGU, CGC, CGA and CGG that becomes the regular expression (CG (U|C|A|G)). Deletions and insertions are represented by the number of nucleic acids at the site while frame shifts are represented using the same notation found in GeneWise.

The corresponding DNA sequence is searched for the matching sequence to the reverse translated regular expression sequence obtained from the protein sequence. (Suyama, Torrents, & Bork, 2006)

2.2.8 SQL, MySQL, Amazon Web Services

Structured Query Language (SQL) is a language developed for use with relational databases. SQL has been used to sort, manipulate and filter the data in the relational database, named CAbase, developed for this thesis to answer interesting questions such as whether all the signal peptide cleavage sites for the secreted CAs in exon 1 or 2. SQL is based on the mathematical theories of set theory and tuple relational calculus. (Wikipedia, 2015)

MySQL was used to create CAbase. MySQL is available as an Open Source product or it can be bought from Oracle for businesses. CAbase was uploaded to Amazon Web Services (AWS) on 18 October 2015, where it is hosted on an AWS MySQL server. Both Oracle and Amazon provide extensive documentation on their websites on how to use their products. (Oracle Corporation, 2015) (Amazon, 2015)

3 Methods

3.1 CAbase

CAbase is a specialized relational database that collates the carbonic anhydrase information from Ensembl, NCBI, UniProt, RSCB PDB, UniGene and FlyBase. CAbase also stores some calculated data such as locations of signal peptides, locations of indels for each exon and whether the protein is active in the cytosome, mitochondria or the membrane of the cell, courtesy of the programs SignalP (Petersen, Brunak, von Heijne, & Nielsen, 2011) and TargetP (Emanuelsson, Nielsen, Brunak, & von Heijne, 2000).

3.1.1 CAbase overall procedure

The flowchart of the process of populating CAbase (CAbaseGenerator.py) is shown in Figure 2. Updating the entire database requires at least 24hours because trawling several online databases for specific data needs a lot of wait times to avoid being blacklisted. To minimise the time needed, users can specify which part of CAbase to update using command line arguments as listed in Table 2. This feature is particularly useful when CAbaseGenerator.py crashes - there is user feedback indicating which part of the update process is complete.

Table 2 : CAbaseGenerator.py command line arguments.

Argument	Required	Meaning	Example
-e	Yes	Email address for NCBI	person@university.fi
-u	Yes	Database user name	CAbaseUser
-hst	Yes	Database host name	127.0.0.1
-db	Yes	Database name	CAbase
-genprot	No	Update genomic and protein data. If true all of the data stored for the selected CA(s) to be updated will be deleted and unrecoverable.	false
-ca	No	List the CAs to be updated. If nothing is listed all CAs will be updated.	CA1 CA2
-species	No	Restrict the update to only include the listed species. If nothing is listed all species' data is updated.	Homo sapiens
-id	No	Restrict the update of exons, transcripts and external references to an Ensembl genomic id.	ENSACAG00000002755
--update	No	Defines what type of data to update. The choices are: exon, transcript, extref (for external references), all or none. The default value is none.	none

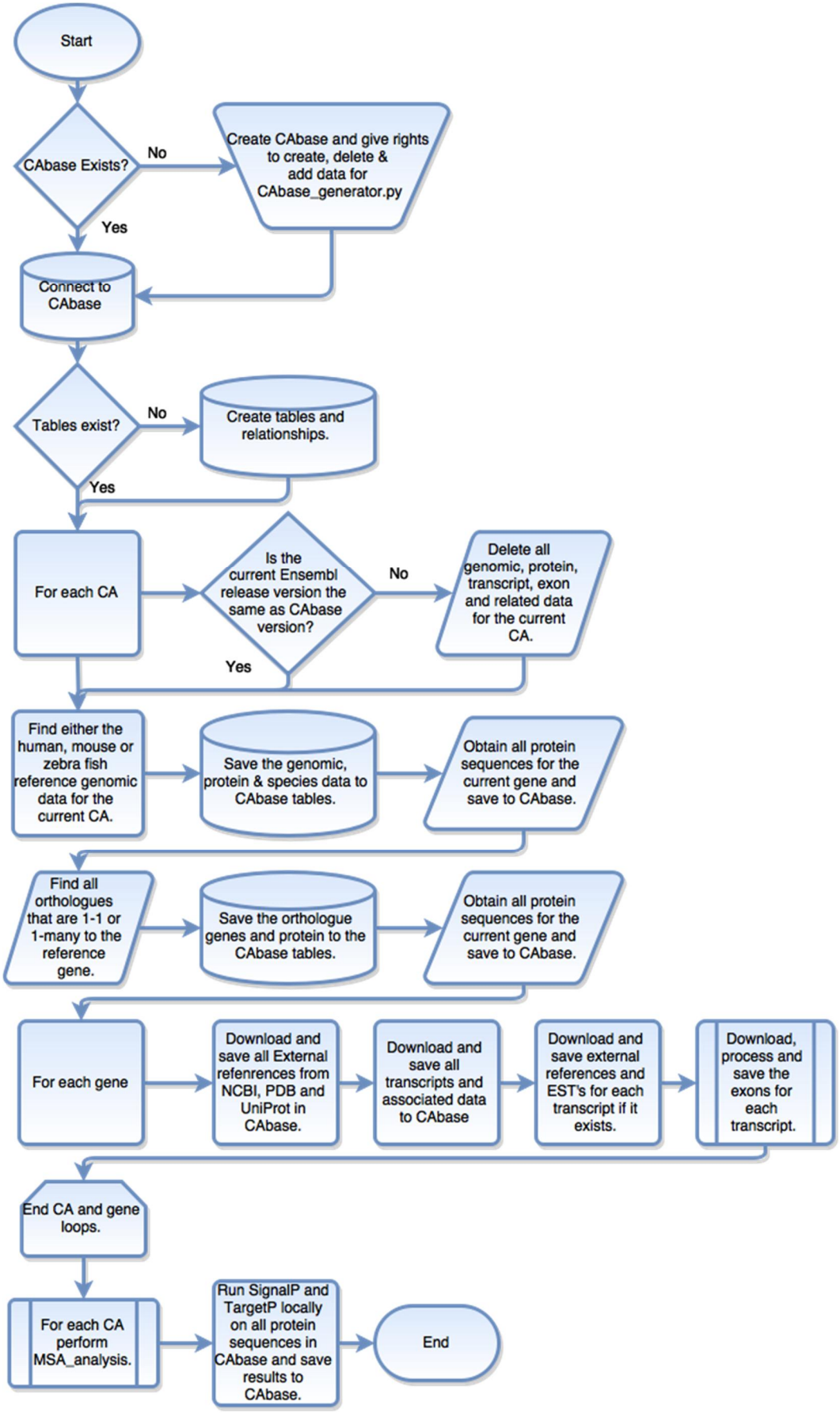

Figure 2 : Flow chart for CAbase_Generator.py for populating CAbase with data from Ensembl, UniProt, NCBI and PDB.

3.1.2 Protein Quality Check - MSA_Analysis

Additionally a modified quality check (MSA_Analysis) was calculated for each of the proteins and stored in the database (Barker, 2013). MSA_Analysis uses a reference protein sequence from either human, mouse or zebra fish to compare the other orthologues against. The check includes measures such as whether the sequence starts with M, how close the locations of the C and N terminals are to the locations in the aligned reference sequence, whether the sequence contains unknown bases (marked as 'X'), and whether the sequence contains insertions or deletions. This procedure is not 'smart' enough to deal with CAs such as CA6 where there is an extra pentraxin domain in non-mammalian species or CA9, which is shorter in non-mammalian species because they have no PG domain. The information generated purely indicates which protein sequences may be 'suspect' and is stored in the table EnsProteins. For ease of reading the protein MSA_Analysis procedure is shown in a flow chart format below in Figure 3.

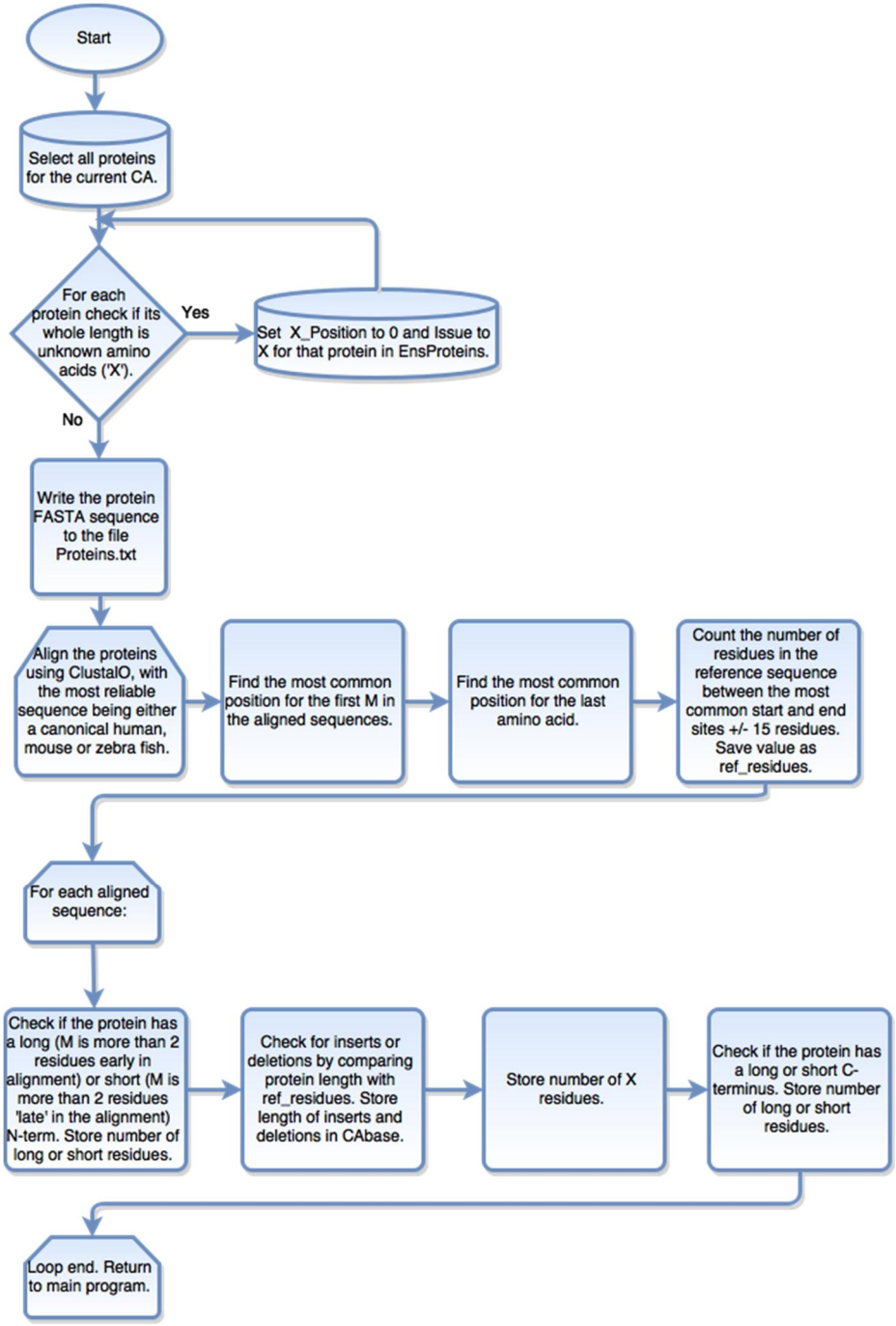

Figure 3 : MSA_analysis flow chart for assessing protein 'quality'.

3.1.3 Exon numbering

Ensembl provides an exon number with the translated coding exon sequences from their REST API for release 81. (Yates, et al., 2015), (Cunningham, et al., 2015). When downloading the un-translated exons for a transcript, Ensembl provides them in their transcribed order. With this information a procedure was scripted such that all the coding and 3' exons are numbered in their order of transcription from 1 to the last exon. The 5' non-coding exons are numbered from -1 to the first transcribed exon. The first coding exon is always numbered 1 even if it has a 5' UTR (as shown in Figure 4 where the purple bar is the 5'UTR of exon 1). The exon MSA schematic diagram provides a key for the colours of the exons at the bottom.

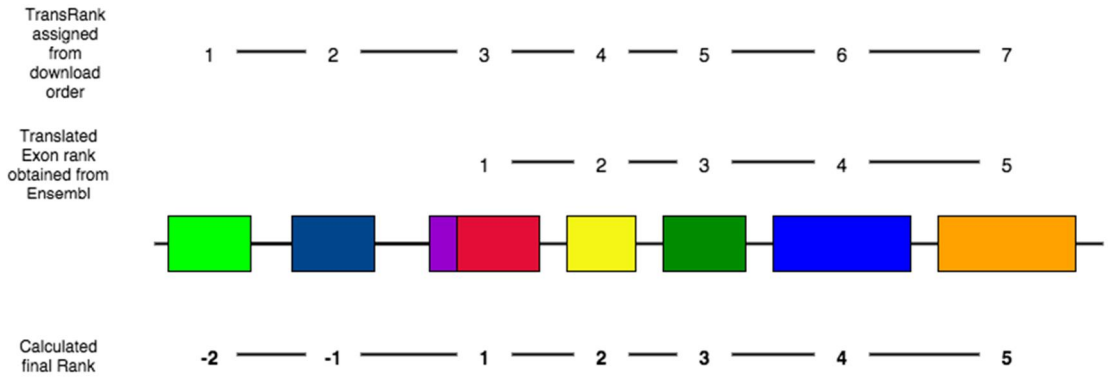

Figure 4 : Exon numbering schematic.

The flow chart for the exon numbering procedure is shown below in Figure 5.

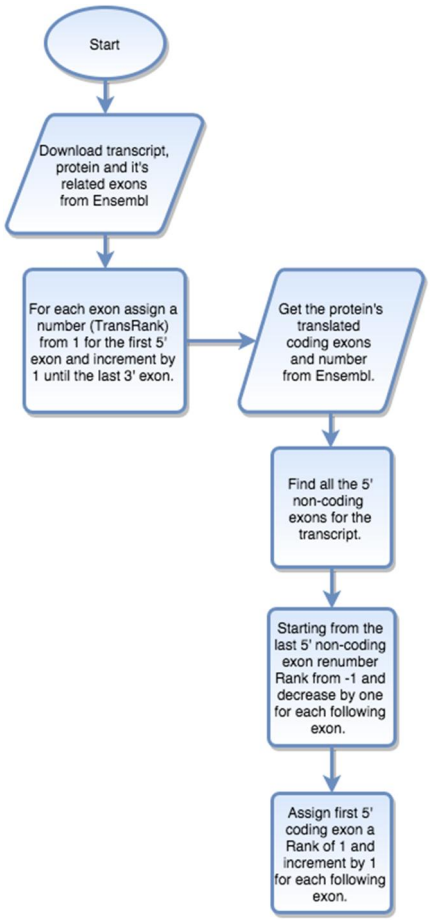

Figure 5 : Exon numbering flowchart

3.1.4 CAbase entity relationship diagram

The entity relationship diagram of CAbase is shown in Figure 6, below. This lists all the tables that make up CAbase and the relationships between the tables. The primary keys of each table are the column names shown in bold font.

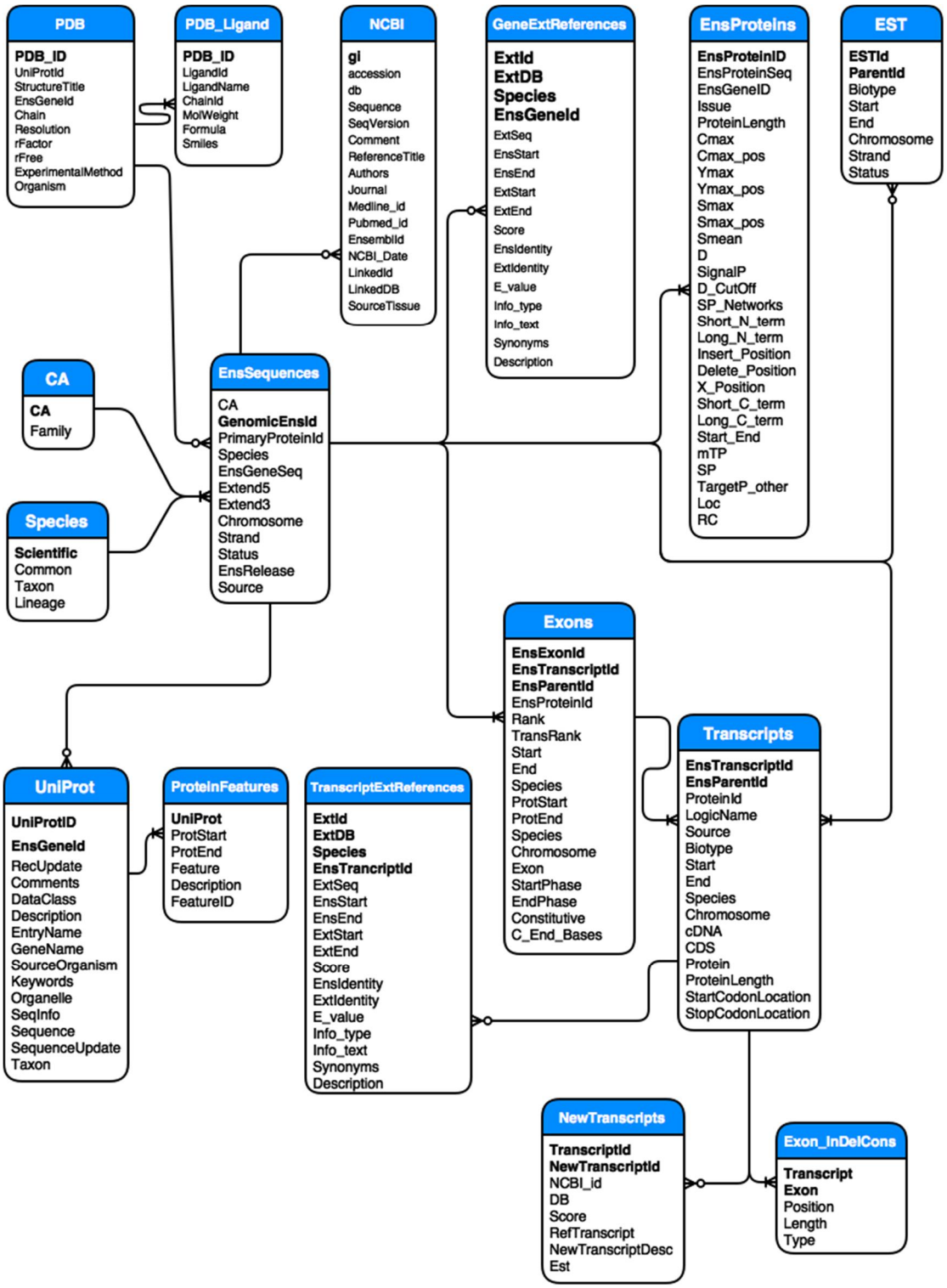

Figure 6 : CAbase entity relationship diagram

3.2 Exon Analysis Pipeline

A dedicated relational database, such as CAbase, organizes the data such that filtering and analyzing the data is much easier to implement using a structured query language (SQL). SQL is a powerful scripting language that allows the user to sort and manipulate data to show interesting relationships. The second part of this thesis examines an application (ExonAnalysis.py) of CAbase where the location of the various peptides (signal, target & mitochondrial) was found relative to the exon boundaries of each CA

isoform. In addition, the conservation of the exons of each transcript was tested in a pipeline with alternative EST or nucleotide sequences from NCBI proposed after BLASTing. (Altschul, Gish, Miller, Myers, & Lipman, 1990) To improve the visualization of the exon MSA schematic diagrams, the user could 'zoom' into interesting parts of the diagram using command line arguments. The overall flowchart for ExonAnalysis.py is shown in Figure 7 and the command line arguments are listed in Table 3.

Table 3 : Command line arguments for ExonAnalysis.py

Argument	Required	Meaning	Example
-e	Yes	Email address for NCBI	person@university.fi
-u	Yes	Database user name	CAbaseUser
-hst	Yes	Database host name	127.0.0.1
-db	Yes	Database name	CAbase
-prot	No	The name of the protein family to analyse.	CA1
-prank	Yes	To choose whether to run PRANK or not. The default is yes.	no
-start	No	The position in the Clustal format MSA to start drawing the exon schematic MSA from. Default = 0.	1500
-end	No	The end position in the Clustal format MSA where the exon schematic MSA drawing would stop. Default = length of alignment.	7000
--file	No	Provide the file name of an alignment in Clustal format to use. It must have the extension ".txt"	Myalignment.txt
-rc	No	The reliability of the target peptides. Indicator arrows will not be drawn for peptides with RC greater than the given value. Default = 3.	4
--type	No	Draw the given type of exon schematic MSA from data in CAbase where full=draw all transcripts for the given protein family, conserved= draw all transcripts that are conserved in the protein family, signal=draw all transcripts with a signal or target peptide, new=draw conserved and transcripts found through BLAST, and, all= draw the full, signal, conserved and new type exon schematic MSA's for the given protein family. Default=conserved.	full

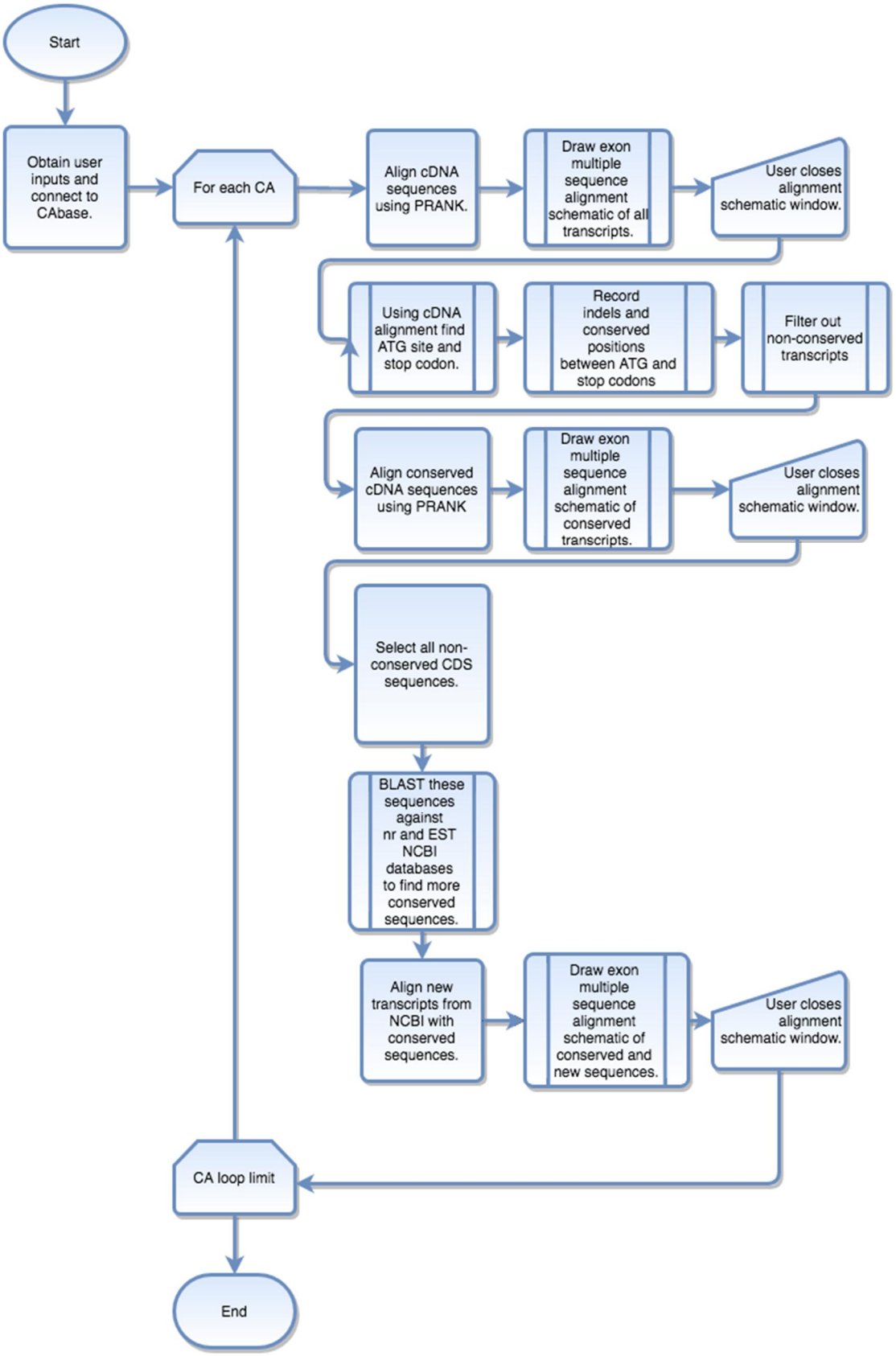

Figure 7 : ExonAnalysis.py pipeline flowchart.

3.2.1 Scaled Exon MSA schematic procedure

After connecting to CAbase all the cDNA sequences for each CA are aligned using Löytynoja's Phylogeny-aware multiple sequence alignment program PRANK.

(Löytynoja, Vilella, & Goldman, 2012) An exon MSA schematic is drawn of this alignment in a pop-up python tk-inter window that the user needs to close before the program progresses. This schematic shows the scaled length of each of the aligned exons and gaps in the alignment. This MSA of all the transcripts is used to assess the conservation of each transcript and to find the locations of the start and stop codons, and to record the locations and frequency of the indels in CAbase in the table Exon_InDelCons. The first exon schematic MSA of each CA typically shows an alignment with very many indels indicating that some of the cDNA sequences may be mispredicted.

The procedure to draw the exon schematic MSA can be used multiple times:

- For drawing the full MSA for all the transcripts of a CA,
- For drawing only the transcripts that have a signal or target peptide,
- For drawing only the transcripts that meet the conservation criteria of a CA, and,
- Finally, for drawing the conserved and proposed BLASTed EST and/or nucleotide replacement sequences for the transcripts that do not meet the conservation criteria.

Start and stop codons are indicated with arrows at their scaled locations in all of the exon MSA schematics, if known. The signal peptide and mitochondrial targeting peptides' locations found through TargetP and SignalP are also shown in their scaled locations with various arrows. (Petersen, Brunak, von Heijne, & Nielsen, 2011) (Emanuelsson, Nielsen, Brunak, & von Heijne, 2000) The legend of the arrow colours and exon colours are shown at the bottom of the exon schematic. The NCBI sequences found with BLAST are drawn in pink with no exon boundaries but with scaled gaps and rectangles for the aligned parts of the sequence. When the tk-inter windows are closed the schematics are saved as post-script files in the program's working directory (Table 4). Some of the exon MSA schematics are shorter than the number of transcripts for the isoform would suggest. This is due to some protein and nucleotide sequences not being translations of each other. Pal2Nal fails in these cases and ExonAnalysis.py stops attempting to draw these sequences in the MSA schematics.

Table 4 : The list of exon MSA schematic types generated by ExonAnalysis.py and the formats of the postscript filenames of each schematic.

Purpose	Filename form
Schematic of all transcripts in Ensembl for each CA isoform	Alignment_for_CAxAll transcript sequences available in Ensembl for CAx.ps
Schematic of all transcripts that are zoomed	Alignment_for_CAxAll transcript sequences available in Ensembl for CAx from yyy to zzzzz.ps
Schematic of all conserved transcripts for a CA isoform	Alignment_for_CAxAligned and Conserved CAx cDNA Sequences showing Start and Stop Codon Locations.ps
Schematic of all transcripts in the isoform with a signal peptide	Alignment_for_CAxSignalPeptide_cdna
Schematic with conserved and BLAST proposed sequences	Alignment_for_CAxWith_BLAST_Seqs_CAx

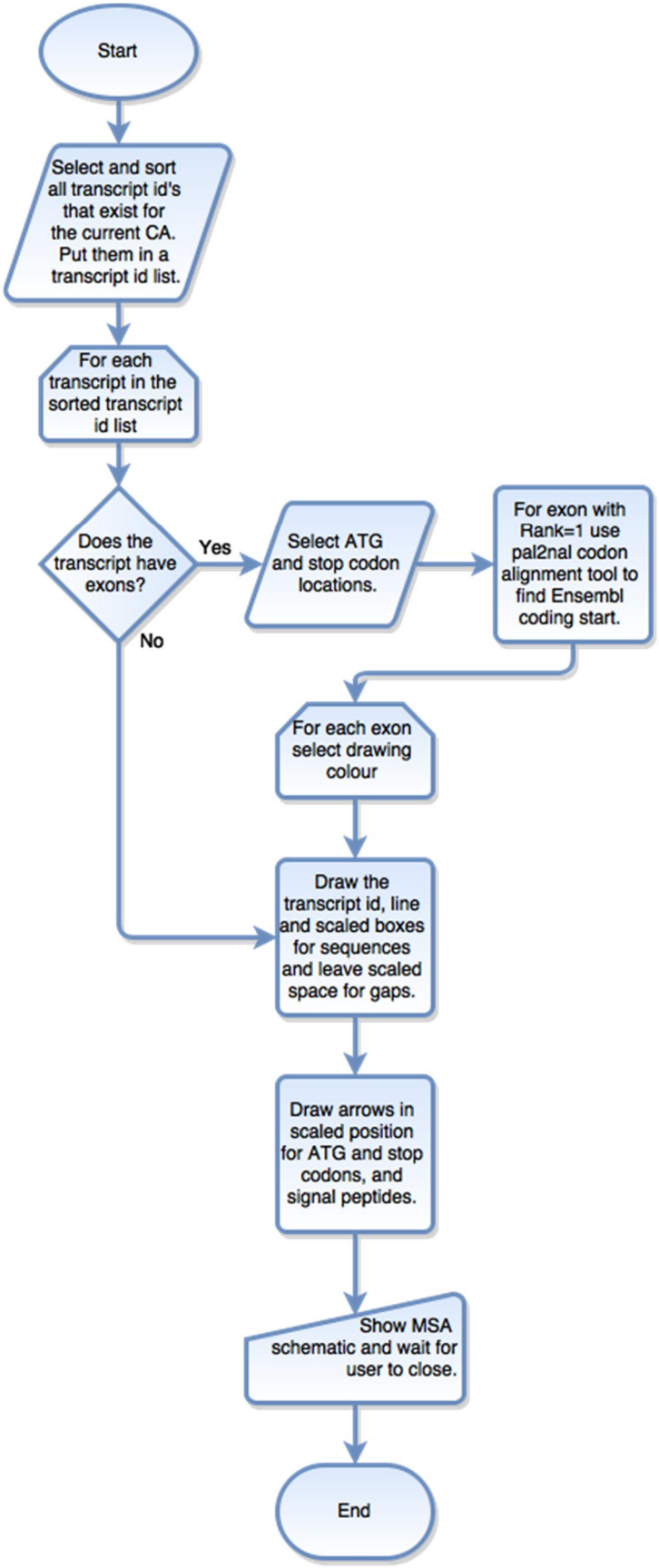

Figure 8 : Flowchart for drawing the exon MSA coloured schematic with arrows indicating the start and stop codons and the predicted signal peptide locations.

3.2.2 Conservation assessment of Transcripts

Transcripts are assessed as being conserved if they share some key properties with either the human or mouse CA reference sequences:

1. They have a start and a stop codon in similar locations in an MSA;
2. They have a similar coding length;
3. They don't have many/large indels in the coding part of the transcript; and,
4. They don't have unknown bases in the coding part.

All of these properties have been coded in ExonAnalysis.py.

If the CDS sequence does not start with 'ATG', the start codon is calculated from the PRANK aligned cDNA sequences. Finding the start and stop codon of each transcript depends on PRANK creating a codon and phylogenetically aware alignment. It is assumed that a start codon aligning with the reference sequence's start codon or the most popular start codon is in the correct frame.

3.2.2.1 Finding Start and Stop Codons

It is assumed that the start codon location in an alignment is conserved for all the transcripts in a CA isoform. Therefore the program starts by counting all the start codons aligning with the reference human or mouse sequence start codon. If there is no start codon aligning with this site the program finds the first ATG position in the alignment that occurs in at least 30% of the sequences. This location is used as a potential location of the start codon since the most popular start site can be skewed to an incorrect location in alignments where there are many sequences from one species (typically human).

The most common start codon is checked against the listed start codon in Ensembl. If the aligned start location is within +/- seven bases of the Ensembl nominated start site, it is saved in the Transcripts table in CAbase. A broad limit was chosen through experimentation to include as many conserved sequences as possible for the final analysis. If it is not the same and the Ensembl sequence has a TSL of 1 then the Ensembl location is considered to be more reliable. Otherwise the Ensembl start codon is indicated in the exon MSA schematic with the purple arrow while the most popular start codon location is indicated with the start of the red rectangle for exon 1. If no start codon can be found then the start codon location is NULL in CAbase. The stop codon is found by simply looking at each 3 bases after the start codon and checking if it is a stop codon. If it is found then it is saved in CAbase. A flow chart of this procedure is shown in Figure 9.

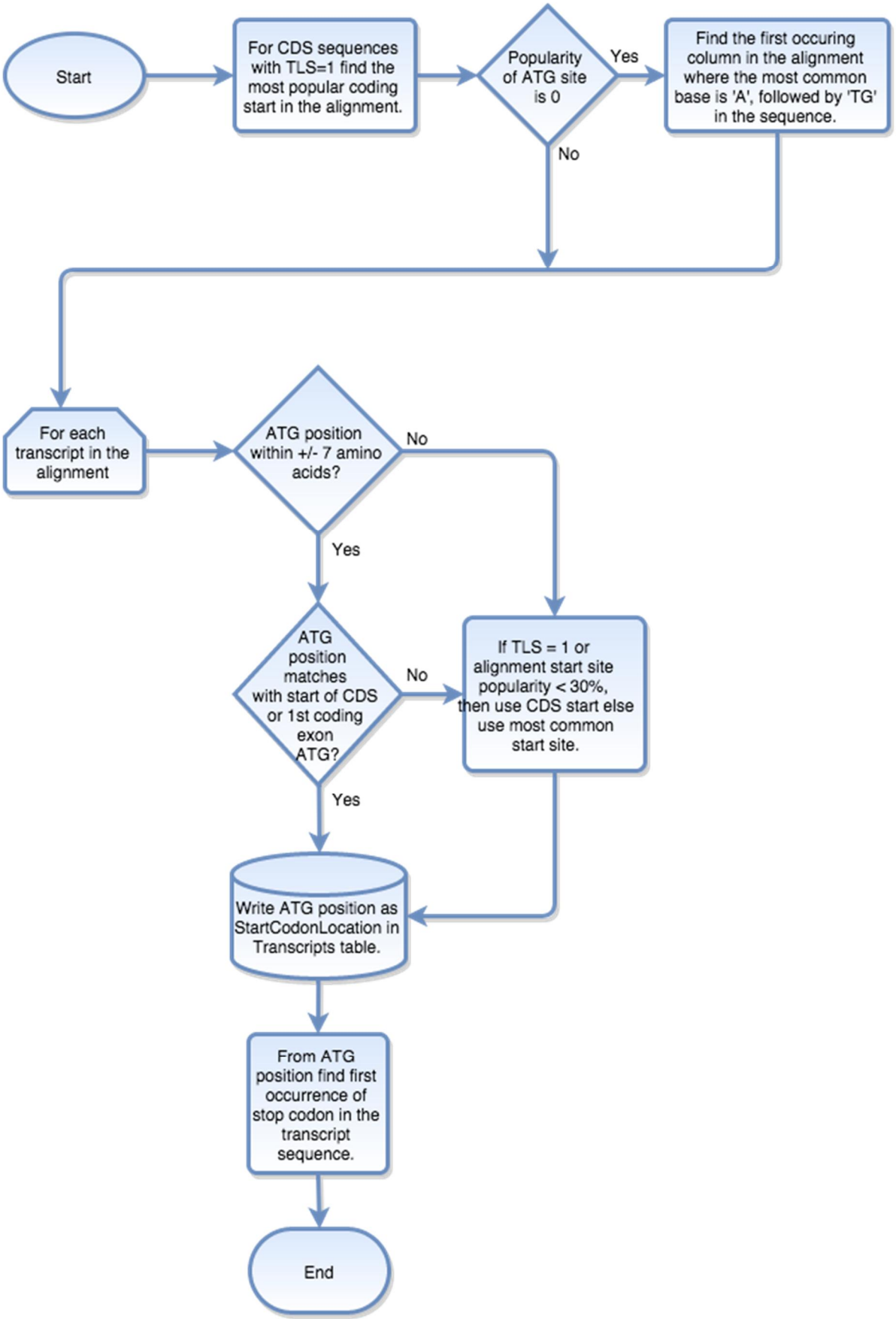

Figure 9 : Flow diagram for finding ATG and stop codons for each aligned cDNA sequence.

3.2.2.2 Is the Base Conserved, Inserted or Deleted?

For the next part of ExonAnalysis.py the conserved, inserted and deleted parts of each exon are found and recorded in the table Exon_InDelCons. Alignments of mispredicted transcripts typically display very large gaps or inserts and/or very many small indels within the coding part of the transcript. Sequences on either the 5' side of the start codon or the 3'side of the stop codon have a lot of variation as there is less selective pressure to maintain the sequence. With this in mind indels are only calculated between either the most common start site in the alignment or the transcript's start codon and the stop codon of the transcript.

Some CAs are less conserved than others and to allow for this situation the value of the 'Majority' changes. 'Majority' stores the number of sequences that align with the start codon of the reference human or mouse sequence. In aligned CAs that are more varied the value of 'Majority' is lower than in the more conserved CAs. Through experimentation it was found that a lower limit of 'Majority' should be greater than 25% of the length of the alignment.

The value of 'Majority' is compared against the frequency of the current base of the transcript being studied in the alignment. If the frequency of the current base is greater than the value of 'Majority' then it is likely that the base is conserved. The flow chart in Figure 10 shows the logic and procedure used to find and store the conserved, inserted and deletion locations within each exon in each transcript.

It has also been found that CAs with greater variability need to have stricter criteria when considering whether a base is inserted or deleted. For CAs that show less variation it has been found that if 90% of the other sequences do not have a gap or bases in the alignment at that location then the sequence has an indel at that location. In CAs with greater variation such as CA6 and CA9, an indel in the sequence under consideration is considered where 95% of the other sequences do not align with the sequence.

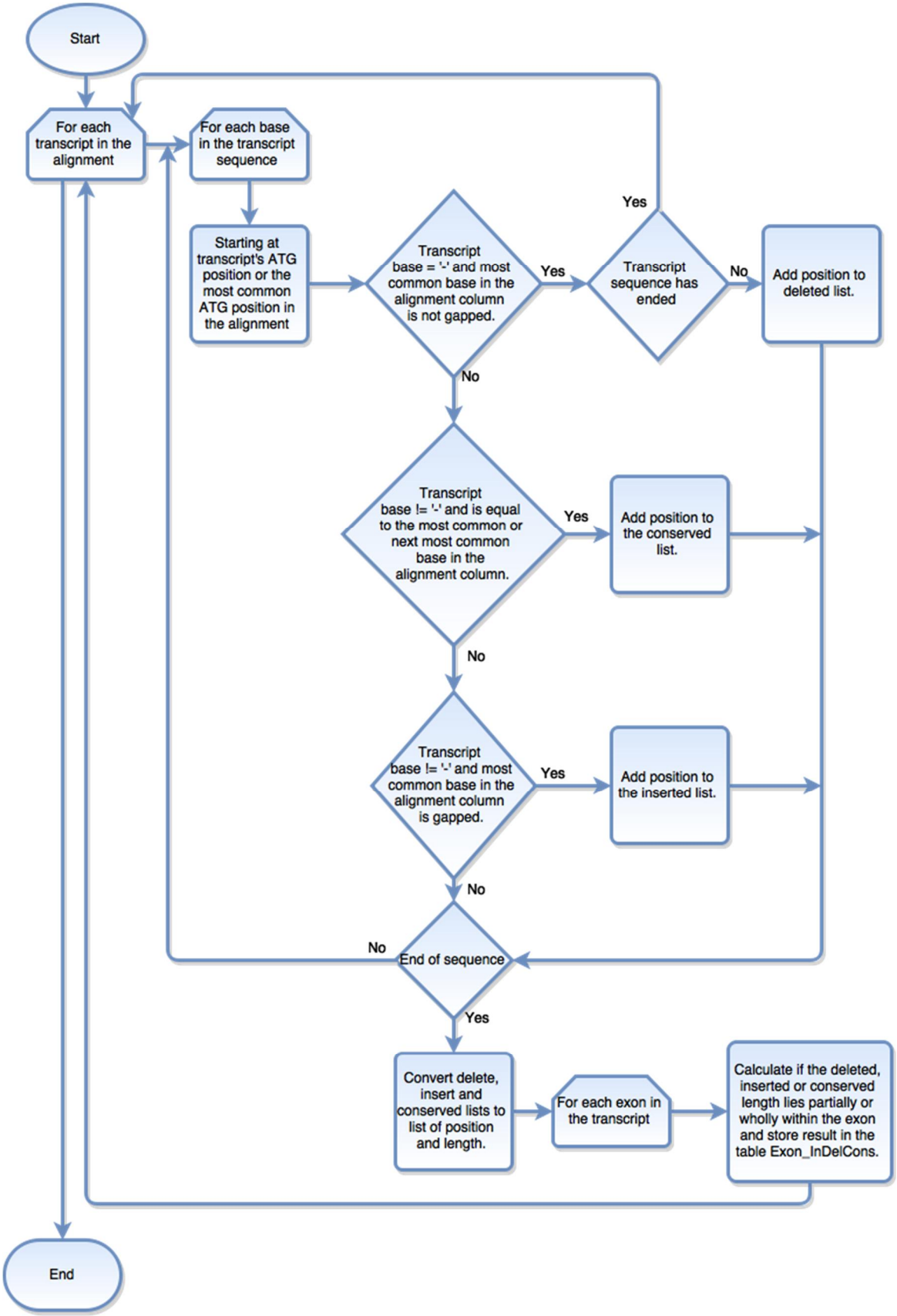

Figure 10 : Find and store indels and conserved lengths of sequence for each exon and record in CAbase.

3.2.2.3 Assessing coding Length

Conserved sequences display lengths of transcripts that fit within certain bounds. Through experimentation it has been found that the acceptable length of CDS sequences can be based on the mean length of the reference sequences with upper and lower length limits within 1.5 to 3 standard deviations (SD). In cases such as CA6 that have entire domains added in non-mammalian sequences or CA9 where entire domains are missing, it is important to have relaxed length criteria. The reference sequences are either human and/or mouse with a transcript support level (TSL) of 1.

Some CDS sequences from Ensembl are possibly mispredicted since they do not start with the expected ATG start codon and therefore include possible non-coding 5' and 3' bases. With this in mind the upper limits of the accepted conserved sequence lengths are more relaxed than the lower limits, which could indicate missing parts of the sequence, except in CA9 where it is known that non-mammalian sequences are missing entire proteoglycan domains.

In cases where the CAs have more conserved sequence lengths or many shorter sequences leading to smaller standard deviations of length (e.g. CA2, CA7, CA8, CA11, CA12, CA13) the standard deviation has been replaced by a fixed value of 100 and the minimum length was set at the (mean length - 50) while the maximum length was (mean length + 100). These bounds in sequence length were found through experimentation.

For most other CAs the minimum length was set at (mean length - SD) and the maximum length at (mean length + 2*SD), except for CA6 where non-mammalian sequences have the extra pentraxin domain. Here the maximum length is mean conserved length plus 3 standard deviations while the minimum length was the conserved mean minus one-quarter standard deviation.

Table 5 : CA filtering parameters based on CDS reference length.

CA	Mean CDS Ref Length	Standard Deviation	Minimum Length	Maximum Length
CA1	779	170	609	1119
CA2	783	100	733	883
CA3	783	131	652	1045
CA4	928	141	787	1210
CA5A	909	132	777	1173
CA5B	954	161	793	1276
CA6	871	326	545	1523
CA7	739	100	689	859
CA8	874	100	824	974
CA9	1119	250	869	1619
CA10	781	243	538	1267
CA11	987	115	872	1217
CA12	1049	100	999	1149
CA13	789	100	739	889
CA14	1014	133	881	1280
CA15	975	141	834	1257

The procedure for assessing whether a transcript was within the acceptable length bounds is shown in Figure 11.

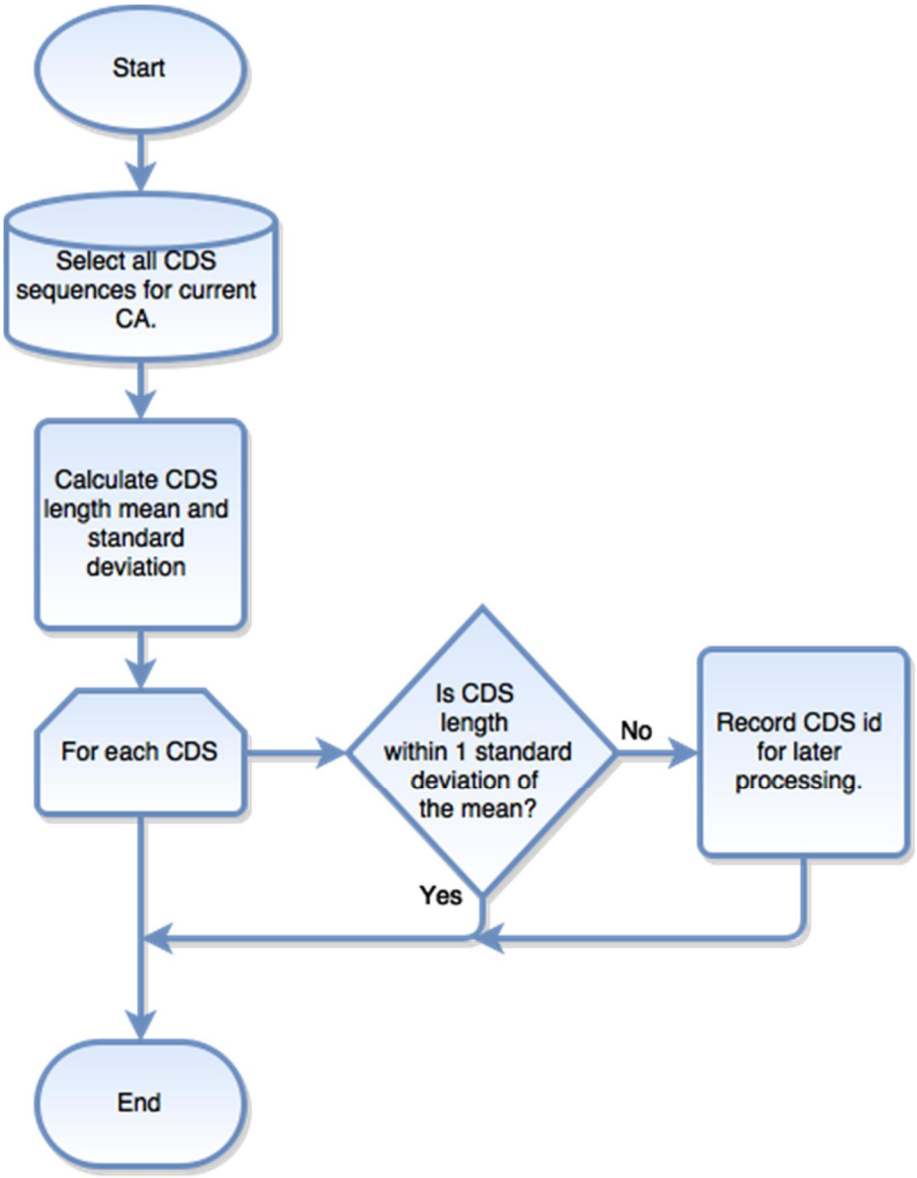

Figure 11 : Method for calculating conserved length of transcripts.

3.2.3 Filtering out non-conserved Transcripts

Using these calculated data points potentially mispredicted exons are filtered out so that only the conserved cDNA sequences are re-aligned using PRANK and the exon MSA schematic is redrawn. Filtering out the non-conserved sequences turned out to be quite challenging using SQL as many queries were composed of several sub-queries that created race conditions within Python thereby producing unstable results. It was necessary to store the results of sub-queries as lists and temporary tables within Python and then use simpler SQL queries based on those lists and temporary tables. This method is a lot slower but it produces consistent results. If the entire program could be written in SQL it would not be necessary to break up the complex queries, as they are stable within MySQL.

The filtering conditions are shown in the flow chart in Figure 12. Not all queries are applied to all CAs since some CAs are more variable than others. It was found that for some CAs even the conserved sequences have a lot of variation. In those cases only the coarse filters were applied to the sequences so that the normal variation could still be viewed in the schematic.

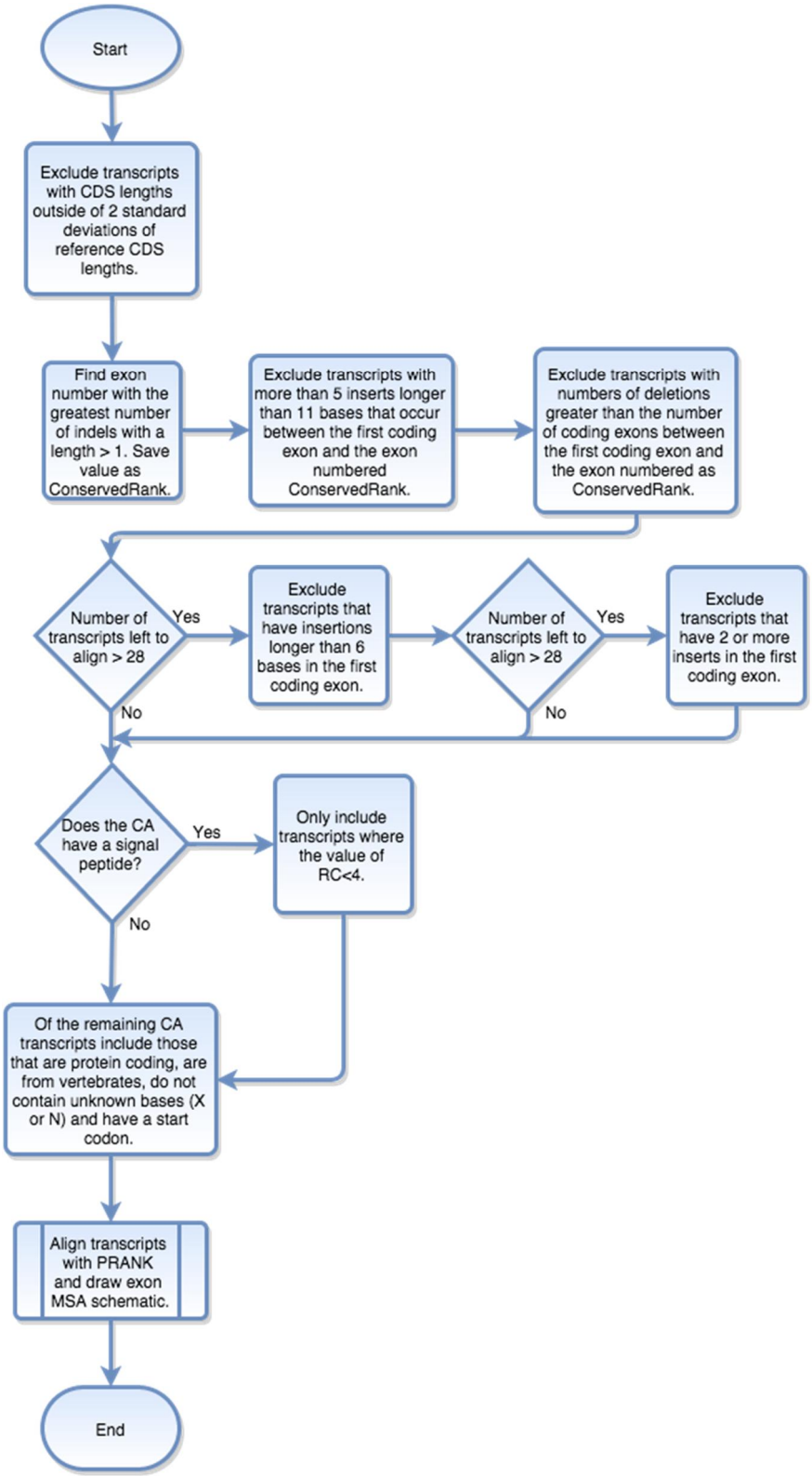

Figure 12 : Multi-step filter for conserved cDNA sequences.

3.2.4 Alternative Sequences from NCBI

The final step is to propose alternative sequences from NCBI's EST and nucleotide databases using BLAST before re-aligning the conserved sequences with the new proposed sequences in the exon MSA schematic. Only EST or nucleotide sequences that have a higher score than the sequence provided by Ensembl is shown in this exon MSA schematic. If there are no sequences that have a higher score than the original sequence then no exon MSA schematic window will be opened.

4 Results

The first result of this thesis is the creation and hosting of CAbase on the Amazon Web Server (Amazon, 2015). This relational database contains all vertebrate alpha CA data extracted from Ensembl, UniProt, NCBI, CBS and PDB. Calculated values, such as the location of the signal peptides are also stored for each CA protein.

The program Exon_Analysis.py visually displays the location of the signal peptide calculated indicators from SignalP and TargetP in the exon MSA schematic. This visualisation shows that for non-cytoplasmic CAs the signal peptide typically lies around the boundary of exons 1 and 2. Cytoplasmic CAs (CA1, CA2, CA3, CA7 and C13) do not have signal or target peptides. The python code that was written to create CAbase and Exon_Analysis is held in a public Dropbox folder found here: https://www.dropbox.com/l/sh/xYQqDdI4TGAi0TTfaJr9Hu .

The following pages contain the exon MSA schematics of the transcripts of each of the CA isoforms. Arrows indicate the locations of the signal and mitochondrial peptides calculated by SignalP and TargetP. The start and stop codons are also shown by arrows. At the bottom of each of the exon MSA schematics is a colour key for the exons and the arrows.

Some of the exon MSA schematics are hard to read because the CA isoform MSA is spread by PRANK thus generating an image with many vertical bars for each transcript. To minimise this issue somewhat, the exon MSA schematics are zoomed in to show the alignment between the start and stop codons in this thesis. For further detail Exon_Analysis.py can be run again to zoom into the part of the MSA that is of interest or the original images of the exon MSA schematics can be accessed in Dropbox here: https://www.dropbox.com/l/sh/xYQqDdI4TGAi0TTfaJr9Hu .

4.1 The cytoplasmic CAs - CA1, CA2, CA3, CA7 and CA13

The following exon MSA schematics (Figure 13 to Figure 26) contain the zoomed images of the PRANK alignments of all and the conserved transcripts between the start and stop codons for each of the CAs in this group. Cytoplasmic proteins do not have reliable predictions of signal or mitochondrial peptides; hence they are not shown within these schematics. If a transcript has been found by BLAST from the NCBI nucleotide or EST databases with a better conservation score than the original sequence from Ensembl, then that is shown in the exon MSA schematics Figure 15, Figure 18, Figure 23 and Figure 26. The basic statistics for each of the cytoplasmic CA

isoforms is in Table 6. Like the other CA groups, the cytoplasmic CAs also contain short exons that are listed in Table 7 and Table 8.

Table 6 : A list of the number of protein coding transcripts for each cytoplasmic CA isoform.

CA isoform	Number of protein coding transcripts
CA1	65
CA2	52
CA3	67
CA7	69
CA13	53

Table 7 : A summary of the CA cytoplasmic isoforms' short exons that are less than 33 residues long. The last three columns show the number of transcripts where the 2nd last exon is short, one or more of the last three exons are short and the number of transcripts where the short exon can be found between the 2nd and the 6th exons i.e. one or more of either the 3rd, 4th or 5th exons.

	Short Exon Transcripts	Protein Coding	% With Short Exon(s)	Last Exon Short	2nd last exon short	Short exon within last 3	Early short exon
Cytoplasmic Transcripts	54	306	18%	15	5	36	24
CA1	12	65	18%	7	2	14	3
CA2	10	52	19%	2	1	5	3
CA3	11	67	16%	1	1	6	7
CA7	12	69	17%	4	0	8	8
CA13	9	53	17%	1	1	3	3

Table 8 : A list of cytoplasmic transcripts with exons shorter than 33 bases (11 residues).

CA	EnsTranscriptId	Short Exon Number	Last Exon Number	Bases	Species	Start Codon Location
CA1	ENST00000517590	6	6	14	Homo sapiens	169
CA1	ENST00000519129	2	2	30	Homo sapiens	424
CA1	ENSMUST00000144327	4	4	24	Mus musculus	91
CA1	ENSAMET00000002770	6	8	9	Ailuropoda melanoleuca	0
CA1	ENSCJAT00000000674	10	10	4	Callithrix jacchus	77
CA1	ENSDNOT00000014904	3	13	25	Dasypus novemcinctus	6
CA1	ENSDNOT00000014904	6	13	21	Dasypus novemcinctus	6
CA1	ENSDNOT00000014904	9	13	24	Dasypus novemcinctus	6
CA1	ENSDNOT00000014904	13	13	19	Dasypus novemcinctus	6
CA1	ENSDNOT00000040923	5	11	21	Dasypus novemcinctus	0
CA1	ENSDNOT00000040923	8	11	24	Dasypus novemcinctus	0
CA1	ENSDORT00000008351	1	7	31	Dipodomys ordii	0
CA1	ENSPSIT00000018422	8	8	18	Pelodiscus sinensis	57
CA1	ENSSHAT00000011914	5	10	25	Sarcophilus harrisii	
CA1	ENSSHAT00000011914	9	10	27	Sarcophilus harrisii	
CA1	ENSTBET00000004637	6	14	6	Tupaia belangeri	0
CA1	ENSTBET00000004637	7	14	8	Tupaia belangeri	0
CA1	ENSTBET00000004637	8	14	22	Tupaia belangeri	0
CA1	ENSTBET00000004637	9	14	10	Tupaia belangeri	0
CA1	ENSTBET00000004637	10	14	20	Tupaia belangeri	0
CA1	ENSTBET00000004637	11	14	17	Tupaia belangeri	0
CA1	ENSTBET00000004637	12	14	17	Tupaia belangeri	0
CA1	ENSTBET00000004637	13	14	32	Tupaia belangeri	0

CA	EnsTranscriptId	Short Exon Number	Last Exon Number	Bases	Species	Start Codon Location
CA1	ENSTBET00000004637	14	14	24	Tupaia belangeri	0
CA1	ENSXETT00000030104	4	7	17	Xenopus tropicalis	
CA2	ENSACAT00000022410	1	8	27	Anolis carolinensis	
CA2	ENSCPOT00000008004	4	8	18	Cavia porcellus	0
CA2	ENSEEUT00000007234	3	12	4	Erinaceus europaeus	0
CA2	ENSEEUT00000007234	5	12	4	Erinaceus europaeus	0
CA2	ENSEEUT00000007234	6	12	14	Erinaceus europaeus	0
CA2	ENSEEUT00000007234	12	12	16	Erinaceus europaeus	0
CA2	ENSFALT00000005146	2	8	8	Ficedula albicollis	231
CA2	ENSMEUT00000016575	6	6	24	Macropus eugenii	0
CA2	ENSMLUT00000026737	6	8	22	Myotis lucifugus	
CA2	ENSMLUT00000026737	7	8	24	Myotis lucifugus	
CA2	ENSPVAT00000002689	5	7	15	Pteropus vampyrus	
CA2	ENSSART00000001834	2	6	29	Sorex araneus	
CA2	ENSTBET00000010545	3	7	11	Tupaia belangeri	0
CA2	ENSXETT00000030171	-1	6	30	Xenopus tropicalis	
CA3	ENSAMXT00000008428	5	17	4	Astyanax mexicanus	
CA3	ENSAMXT00000008428	6	17	10	Astyanax mexicanus	
CA3	ENSAMXT00000008428	7	17	23	Astyanax mexicanus	
CA3	ENSAMXT00000008428	8	17	10	Astyanax mexicanus	
CA3	ENSAMXT00000008428	9	17	16	Astyanax mexicanus	
CA3	ENSAMXT00000008428	10	17	25	Astyanax mexicanus	
CA3	ENSAMXT00000008428	12	17	16	Astyanax mexicanus	
CA3	ENSAMXT00000008428	13	17	17	Astyanax mexicanus	
CA3	ENSDORT00000001001	3	9	6	Dipodomys ordii	0
CA3	ENSDORT00000001001	4	9	4	Dipodomys ordii	0
CA3	ENSEEUT00000007123	1	7	28	Erinaceus europaeus	0
CA3	ENSEEUT00000007123	2	7	6	Erinaceus europaeus	0
CA3	ENSFCAT00000024616	4	10	13	Felis catus	0
CA3	ENSFCAT00000024616	5	10	25	Felis catus	0
CA3	ENSLACT00000002222	4	7	5	Latimeria chalumnae	
CA3	ENSLOCT00000015799	8	8	22	Lepisosteus oculatus	0
CA3	ENSMLUT00000028858	1	7	25	Myotis lucifugus	
CA3	ENSMLUT00000028858	4	7	6	Myotis lucifugus	
CA3	ENSOCUT00000033537	2	9	24	Oryctolagus cuniculus	
CA3	ENSPVAT00000007169	3	11	9	Pteropus vampyrus	0
CA3	ENSPVAT00000007169	4	11	31	Pteropus vampyrus	0
CA3	ENSPVAT00000007169	9	11	28	Pteropus vampyrus	0
CA3	ENSTGUT00000019143	1	7	28	Taeniopygia guttata	
CA3	ENSTSYT00000010492	7	9	6	Tarsius syrichta	0
CA3	ENSTSYT00000010492	8	9	9	Tarsius syrichta	0
CA7	ENSAMET00000018830	1	8	26	Ailuropoda melanoleuca	0
CA7	ENSAMET00000018830	2	8	14	Ailuropoda melanoleuca	0
CA7	ENSAMXT00000005405	5	9	13	Astyanax mexicanus	273
CA7	ENSAMXT00000005405	6	9	20	Astyanax mexicanus	273
CA7	ENSETET00000011872	4	12	8	Echinops telfairi	
CA7	ENSETET00000011872	6	12	23	Echinops telfairi	
CA7	ENSETET00000011872	7	12	2	Echinops telfairi	
CA7	ENSETET00000011872	10	12	6	Echinops telfairi	
CA7	ENSETET00000011872	12	12	3	Echinops telfairi	
CA7	ENSGACT00000020344	8	8	18	Gasterosteus aculeatus	55
CA7	ENSGACT00000020345	8	8	21	Gasterosteus aculeatus	59
CA7	ENSOANT00000019326	5	7	1	Ornithorhynchus anatinus	
CA7	ENSPVAT00000000384	5	9	3	Pteropus vampyrus	0
CA7	ENSPVAT00000000384	6	9	18	Pteropus vampyrus	0
CA7	ENSSHAT00000003720	1	7	17	Sarcophilus harrisii	
CA7	ENSSSCT00000029058	4	12	18	Sus scrofa	
CA7	ENSSSCT00000029058	7	12	3	Sus scrofa	
CA7	ENSTRUT00000044327	8	8	7	Takifugu rubripes	0

CA	EnsTranscriptId	Short Exon Number	Last Exon Number	Bases	Species	Start Codon Location
CA7	ENSTBET00000010632	1	12	18	Tupaia belangeri	
CA7	ENSTBET00000010632	2	12	3	Tupaia belangeri	
CA7	ENSTBET00000010632	3	12	6	Tupaia belangeri	
CA7	ENSTBET00000010632	4	12	15	Tupaia belangeri	
CA7	ENSTBET00000010632	5	12	3	Tupaia belangeri	
CA7	ENSTBET00000010632	6	12	30	Tupaia belangeri	
CA7	ENSTBET00000010632	7	12	6	Tupaia belangeri	
CA7	ENSTBET00000010632	8	12	15	Tupaia belangeri	
CA7	ENSXETT00000037904	2	9	23	Xenopus tropicalis	0
CA7	ENSXETT00000037904	3	9	30	Xenopus tropicalis	0
CA13	ENSCHOT00000002582	4	10	3	Choloepus hoffmanni	0
CA13	ENSCHOT00000002582	5	10	6	Choloepus hoffmanni	0
CA13	ENSCHOT00000002582	6	10	30	Choloepus hoffmanni	0
CA13	ENSFALT00000005135	1	7	31	Ficedula albicollis	0
CA13	ENSMICT00000003571	2	7	1	Microcebus murinus	
CA13	ENSOPRT00000014188	8	8	32	Ochotona princeps	0
CA13	ENSOANT00000003220	1	7	19	Ornithorhynchus anatinus	
CA13	ENSSHAT00000012595	1	7	7	Sarcophilus harrisii	0
CA13	ENSSART00000012981	7	8	4	Sorex araneus	0
CA13	ENSTBET00000004361	3	8	4	Tupaia belangeri	0
CA13	ENSVPAT00000003345	1	9	4	Vicugna pacos	
CA13	ENSVPAT00000003345	7	9	5	Vicugna pacos	

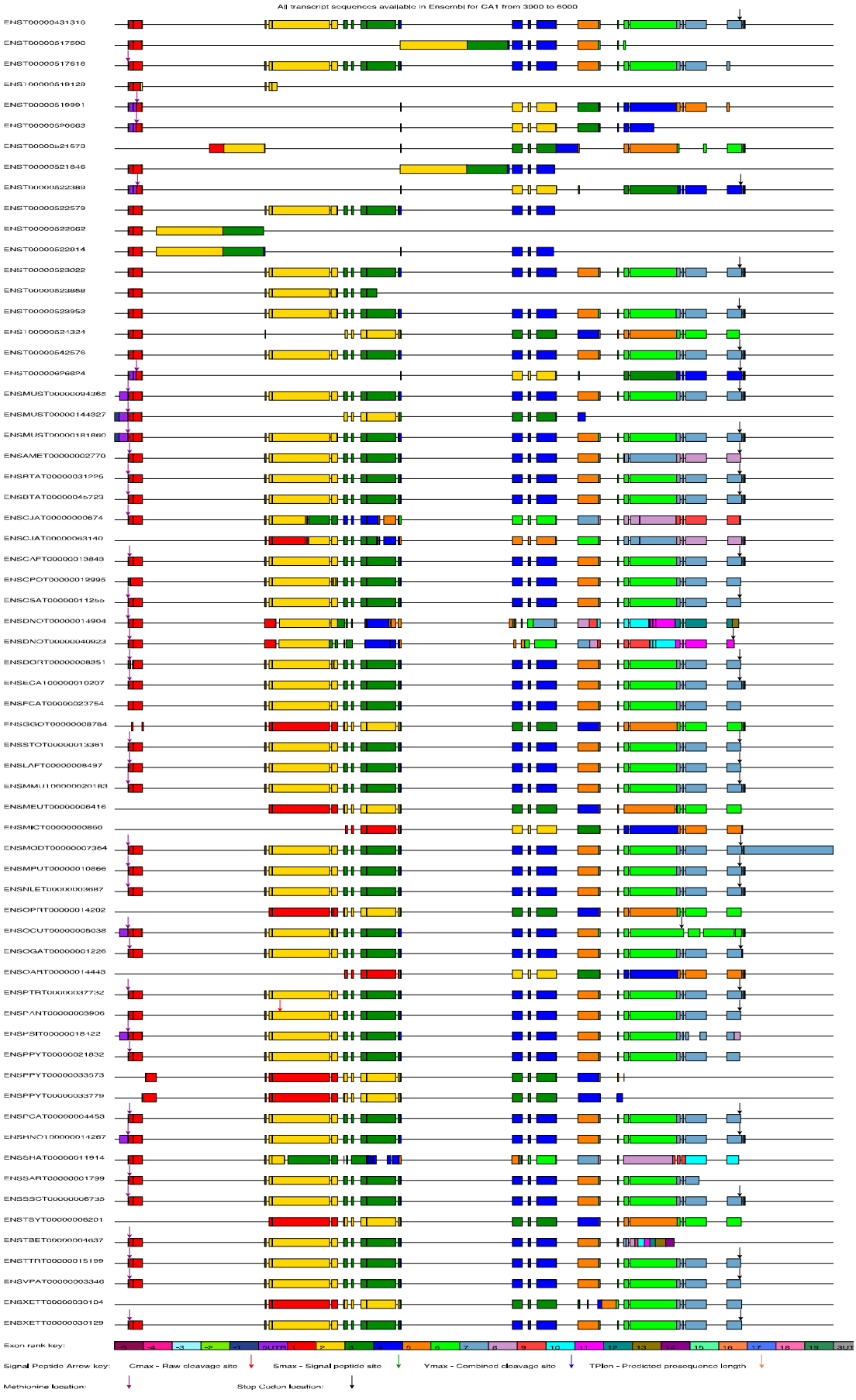

Figure 13 : Exon MSA schematic of all the Ensembl protein coding cDNA sequences of CA1. The zoom is from position 3900 to 6000 in the PRANK MSA.

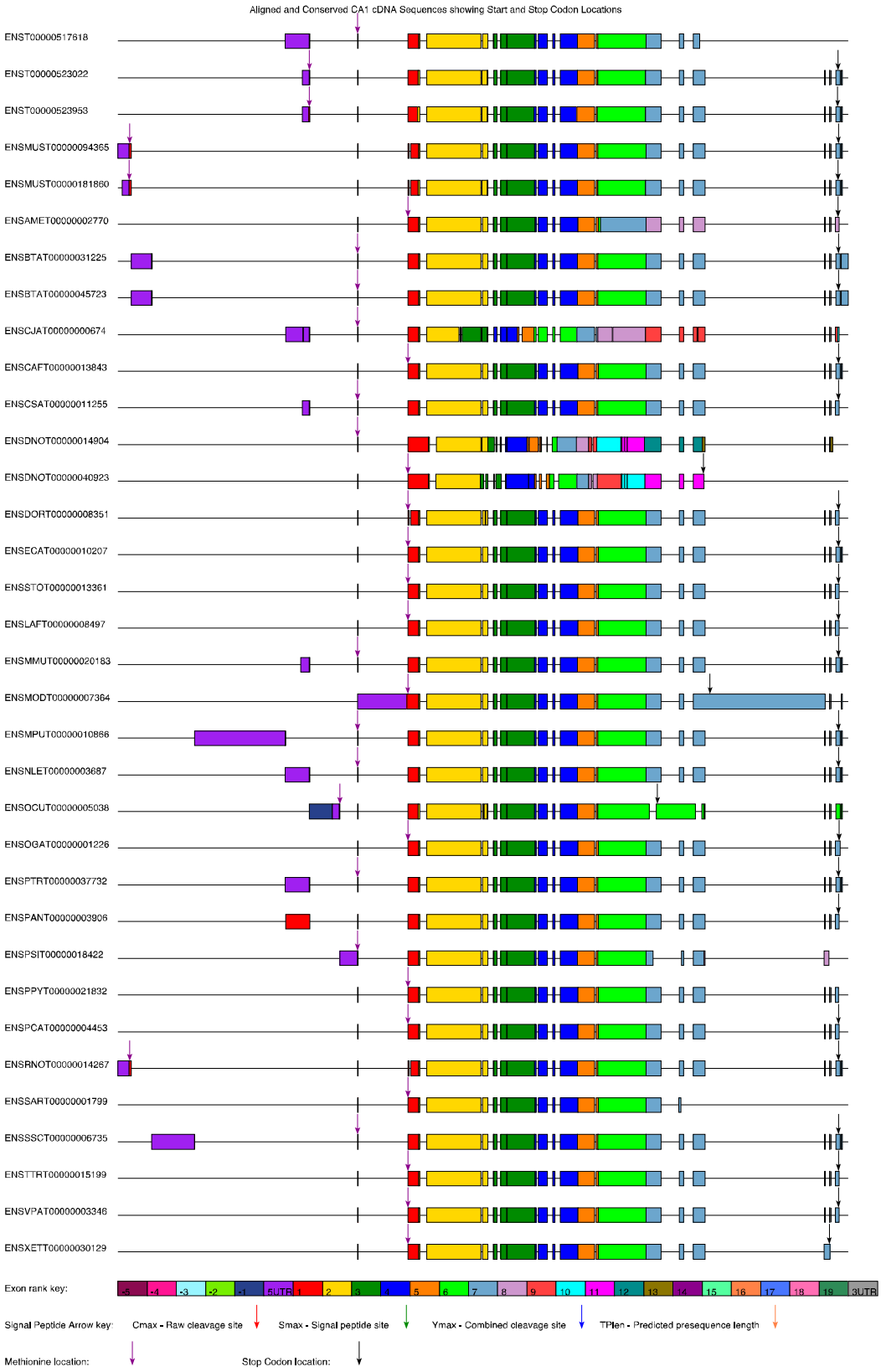

Figure 14 : Exon MSA schematic for conserved CA1 cDNA transcripts. Zoomed in from position 1150 to position 3520 in the MSA.

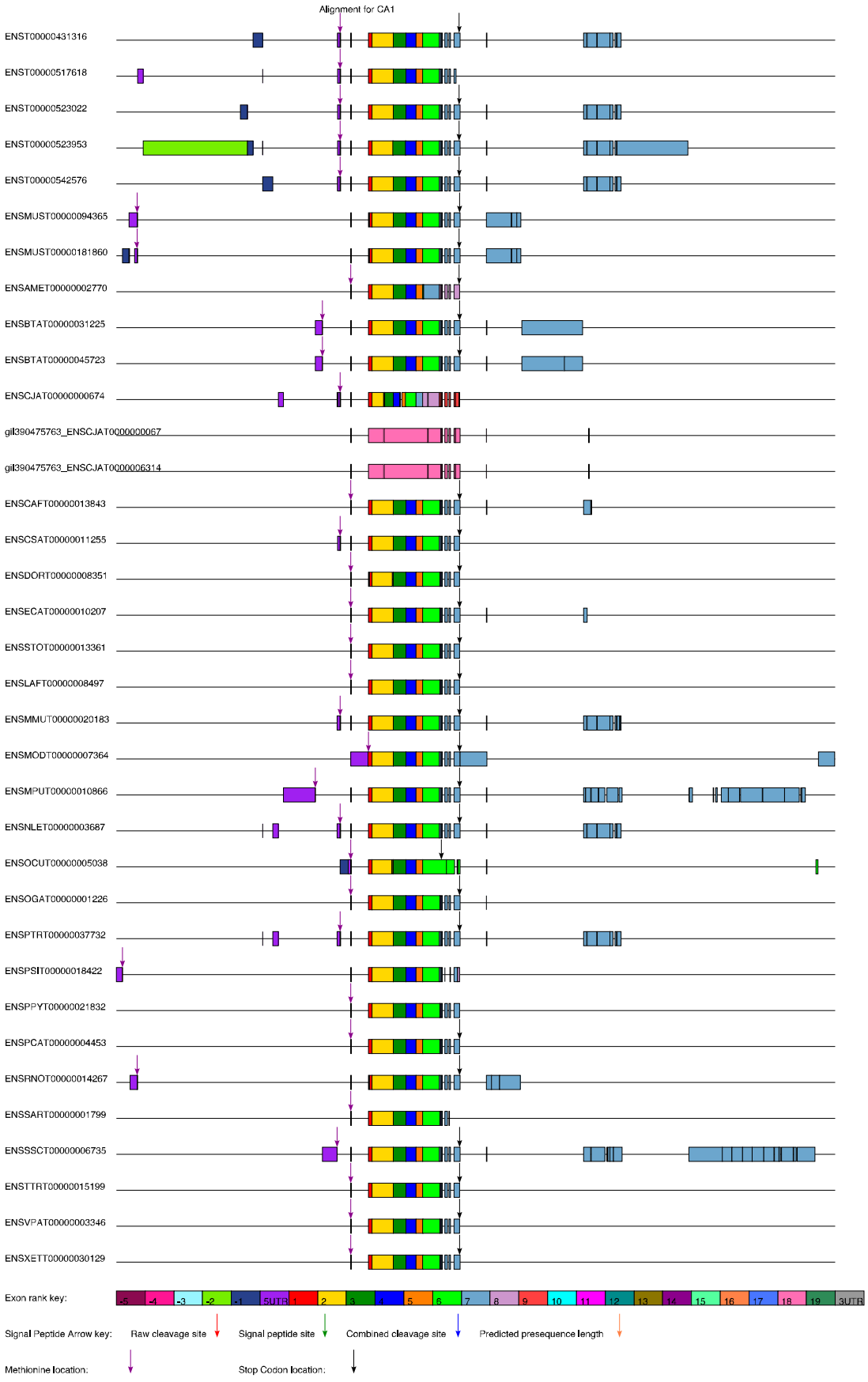

Figure 15 : Exon MSA schematic for entire CA1 conserved cDNA transcripts and suggested EST or nucleotide replacement sequences found through NCBI's BLAST program.

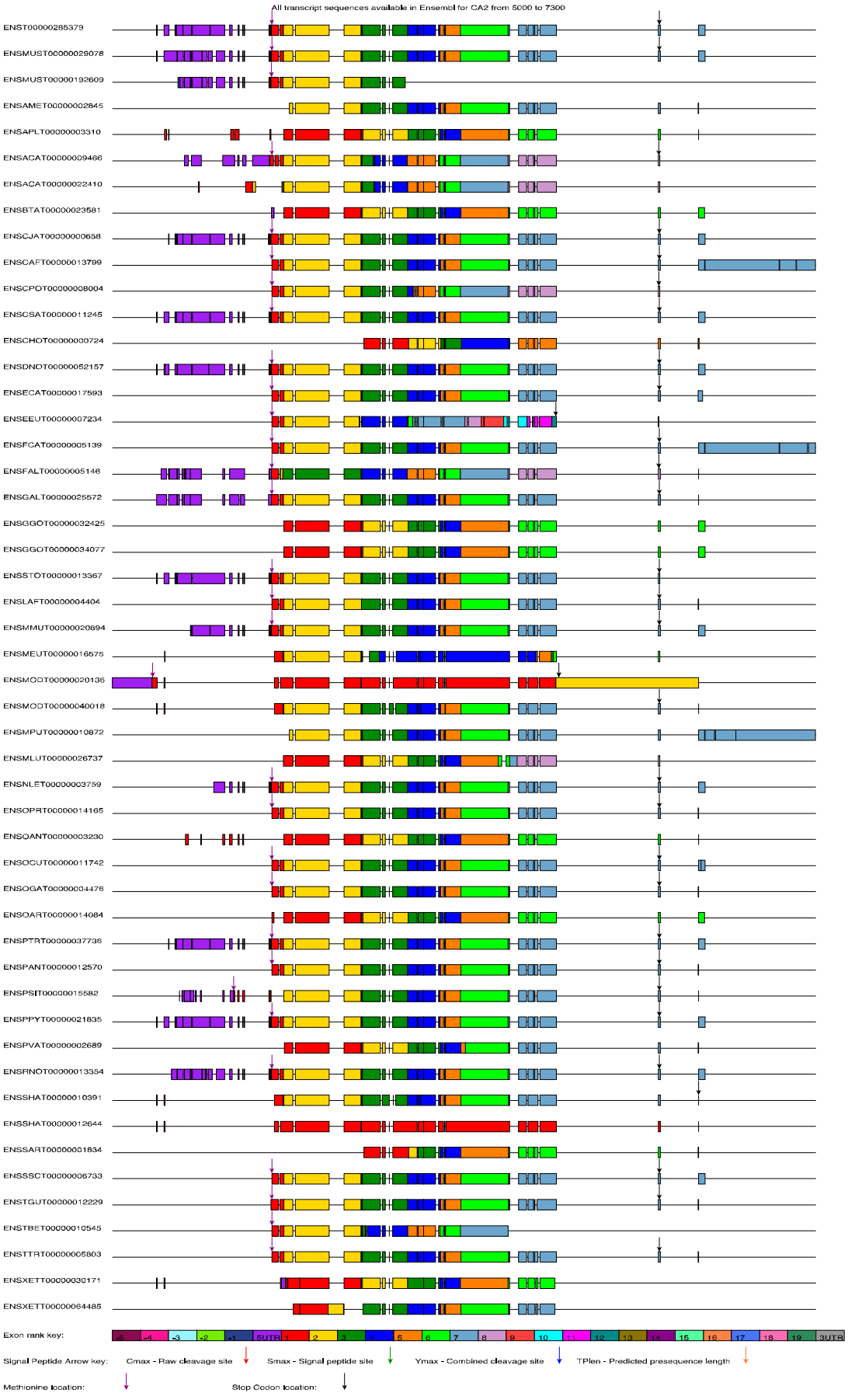

Figure 16 : Exon MSA schematic for all of the Ensembl protein coding cDNA sequences for CA2. The MSA is focused in from position 5000 to position 7300.

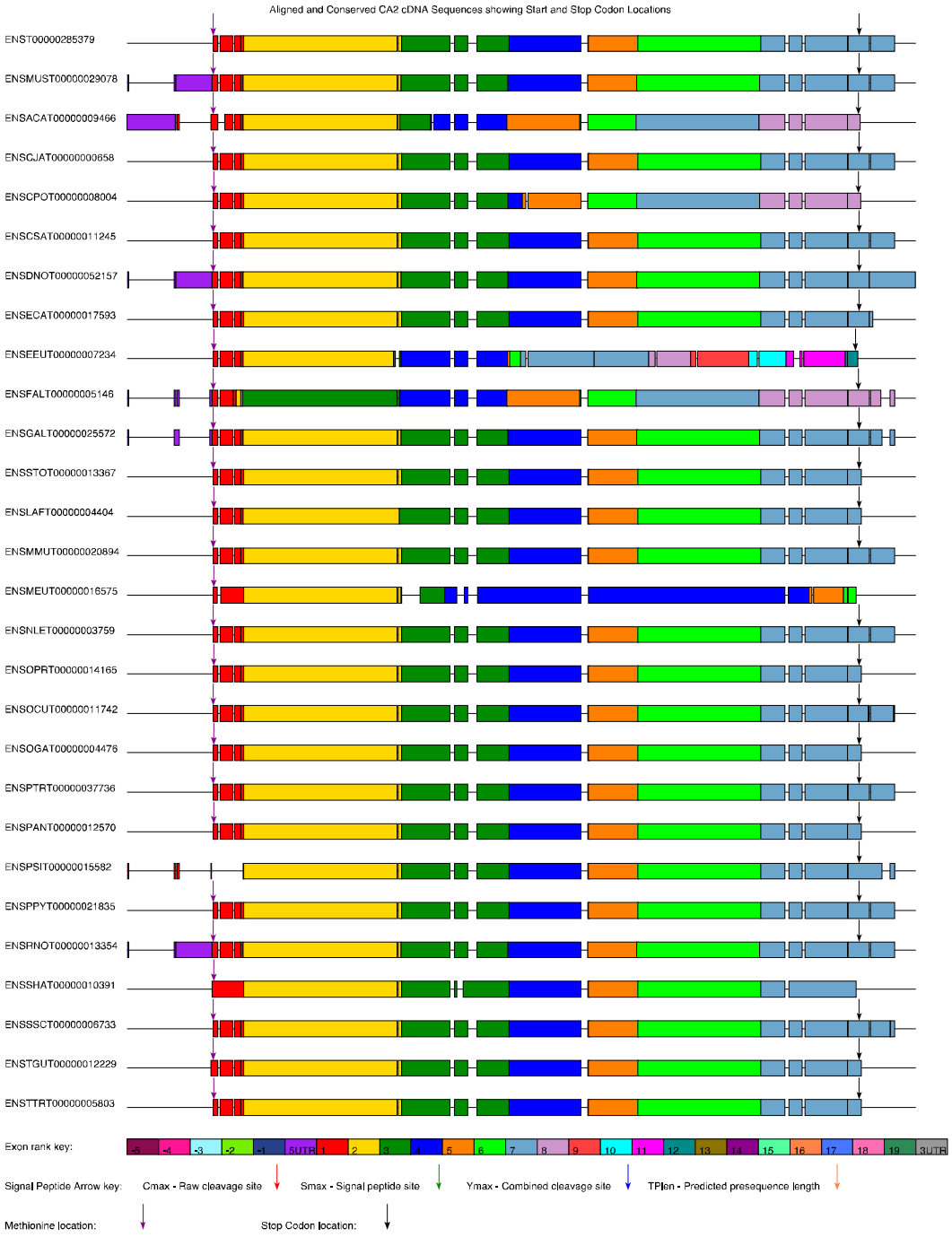

Figure 17 : Exon MSA schematic for conserved CA2 protein coding cDNA transcripts. Zoomed in from position 3100 to position 4100 in the MSA.

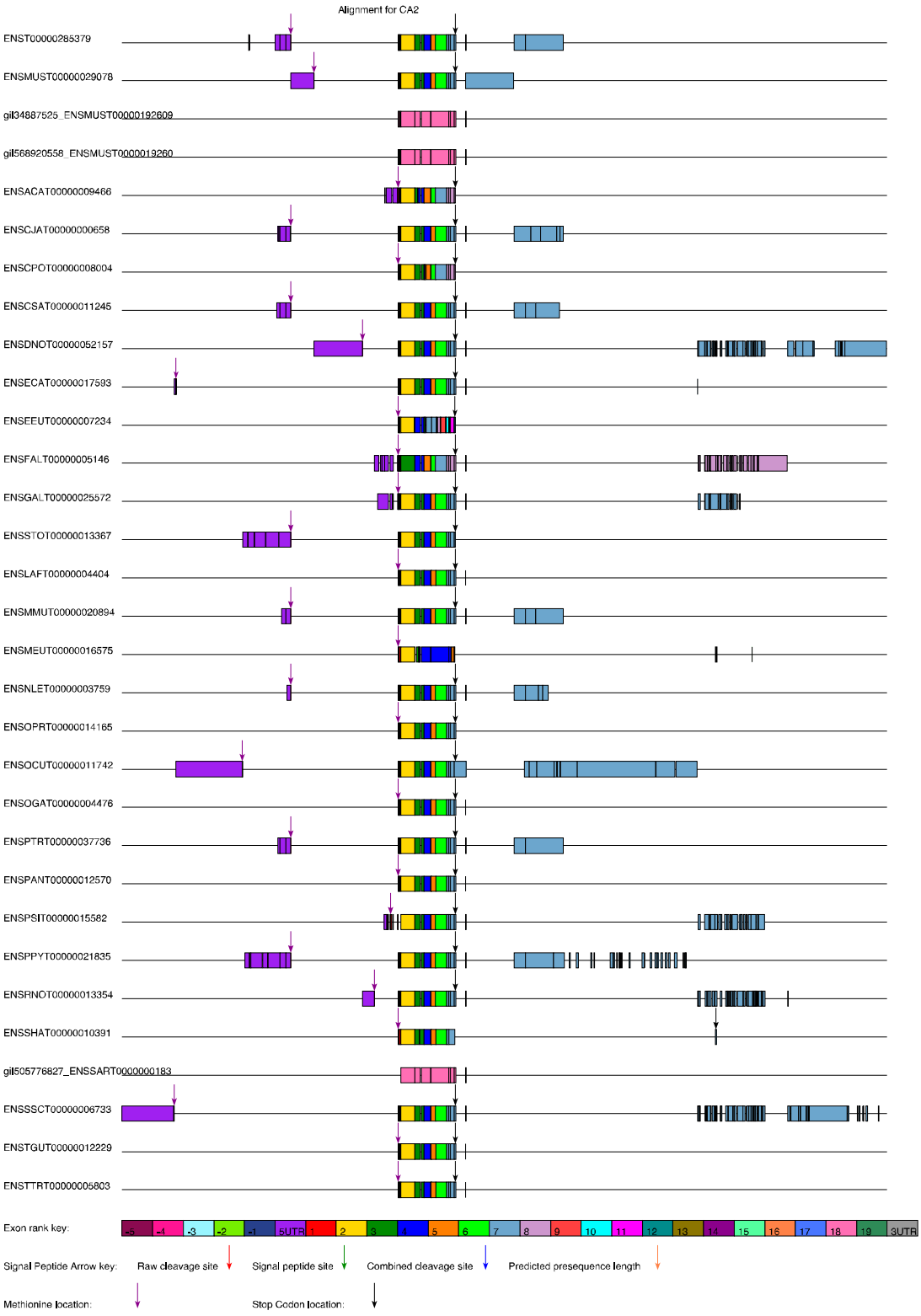

Figure 18 : Exon schematic for CA2 showing conserved CA sequences and the proposed NCBI EST and replacement nucleotide sequences found using BLAST for sequences not fitting the conserved criteria

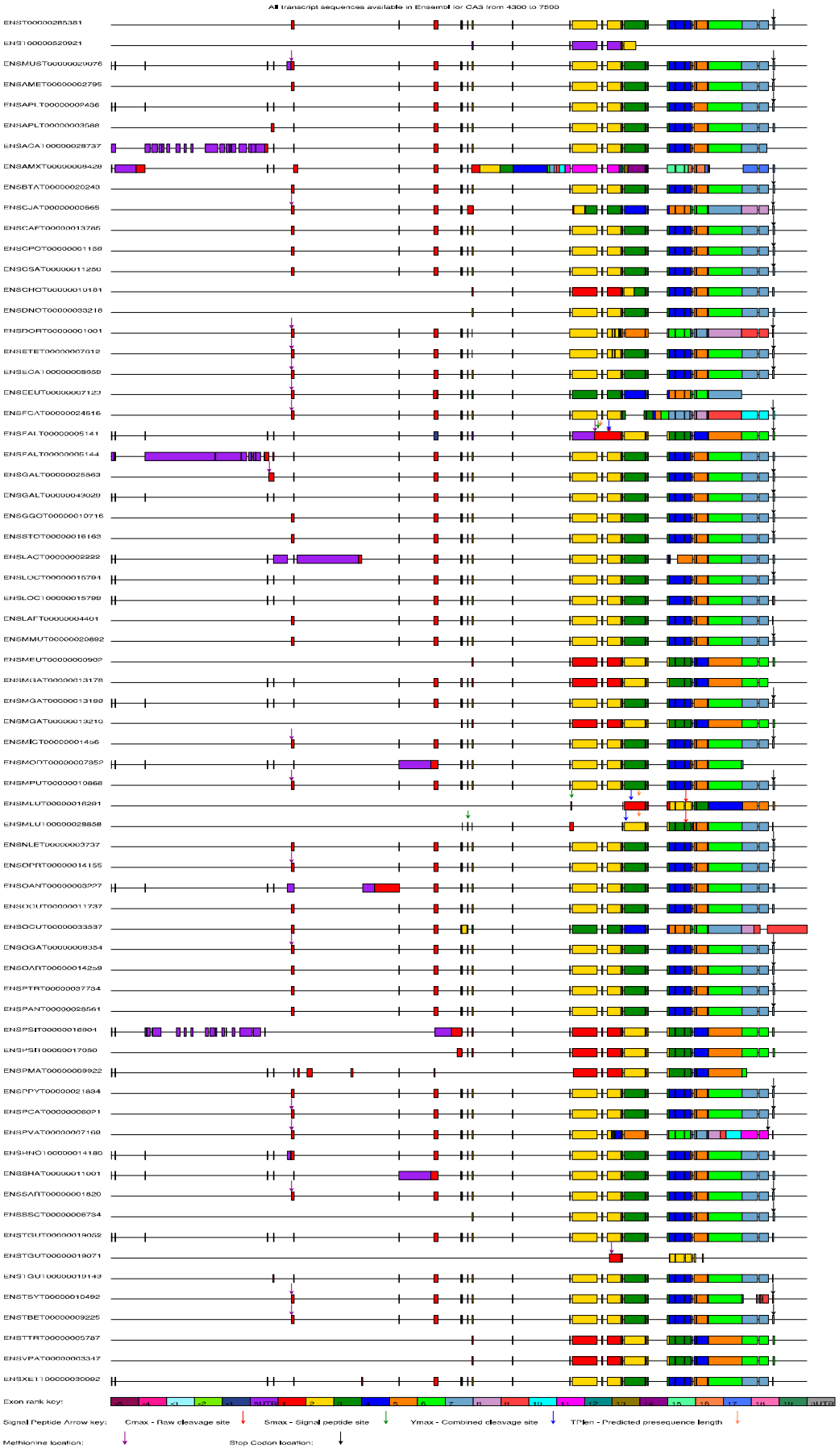

Figure 19 : All protein coding CA3 cDNA Ensembl transcripts in CAbase. The PRANK MSA is zoomed in from position 4300 to 7500 to capture the start and stop codons.

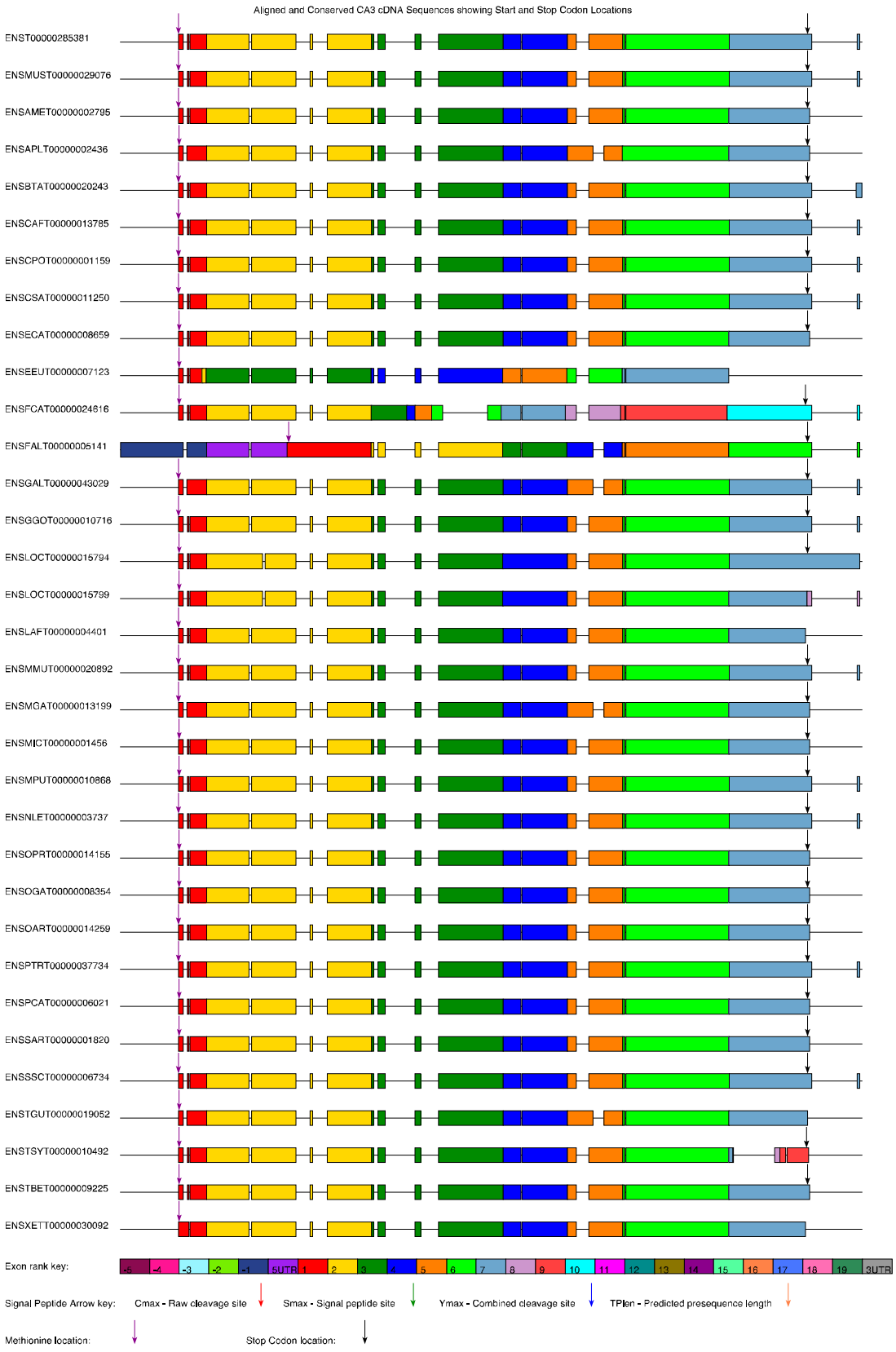

Figure 20 : Conserved exon MSA schematic for CA3 for protein coding cDNA transcripts. Zoomed in from position 2100 to position 3200 in the MSA.

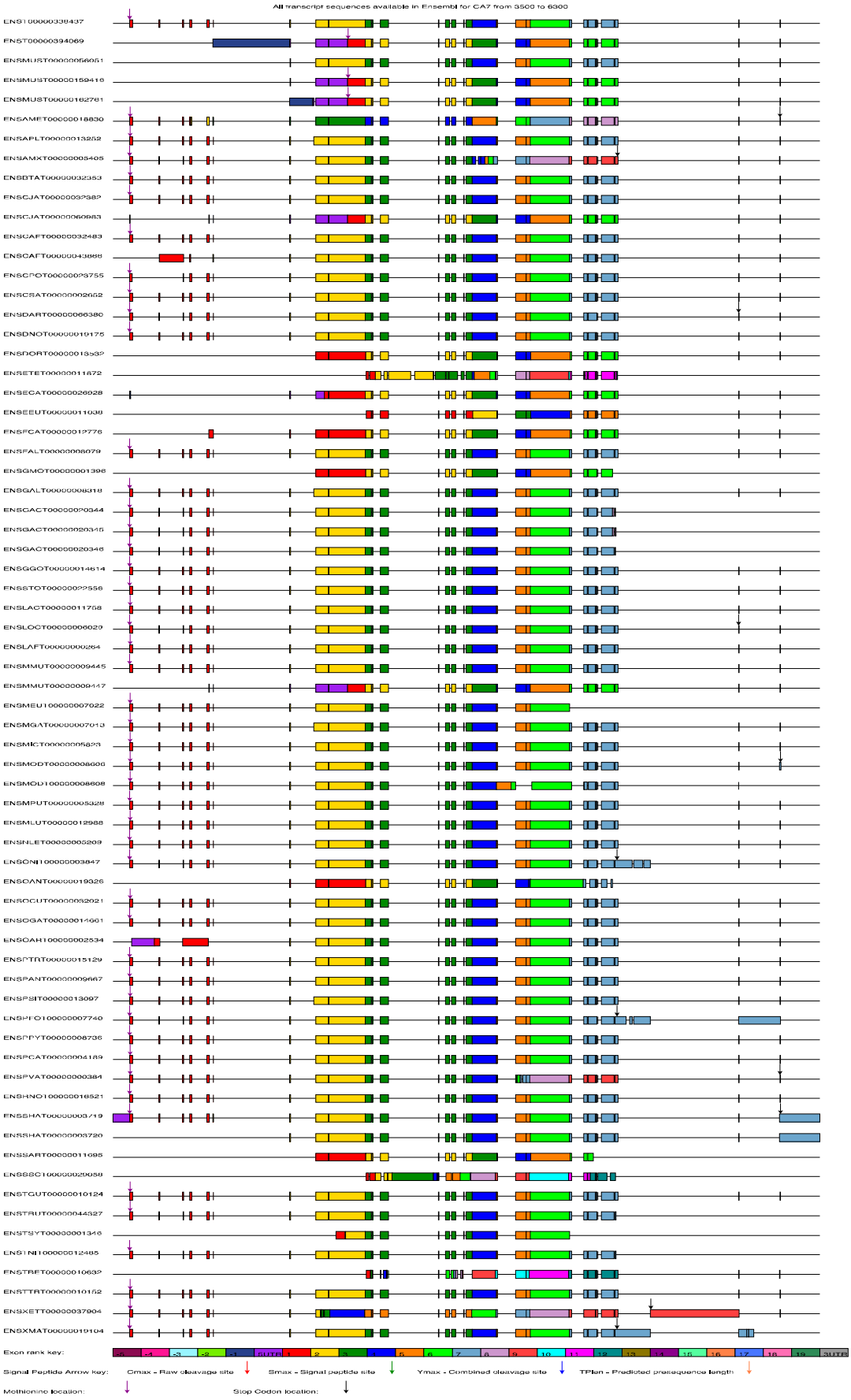

Figure 21 : All protein coding CA7 cDNA Ensembl transcripts in CAbase. The PRANK MSA is zoomed in from position 3500 to position 6300.

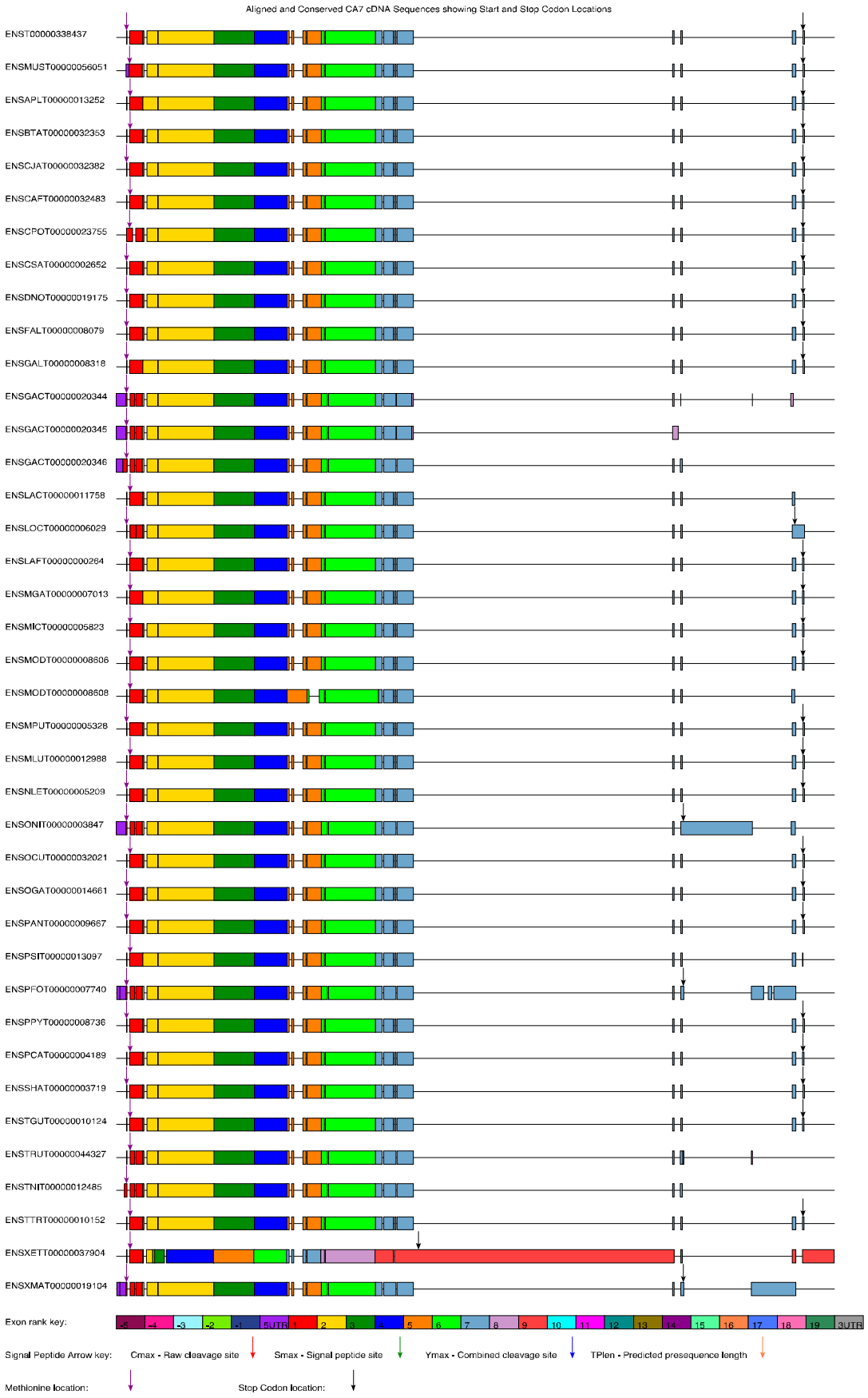

Figure 22 : Exon MSA schematic of conserved protein coding transcripts for CA7. The MSA is zoomed in from position 1900 to 4000.

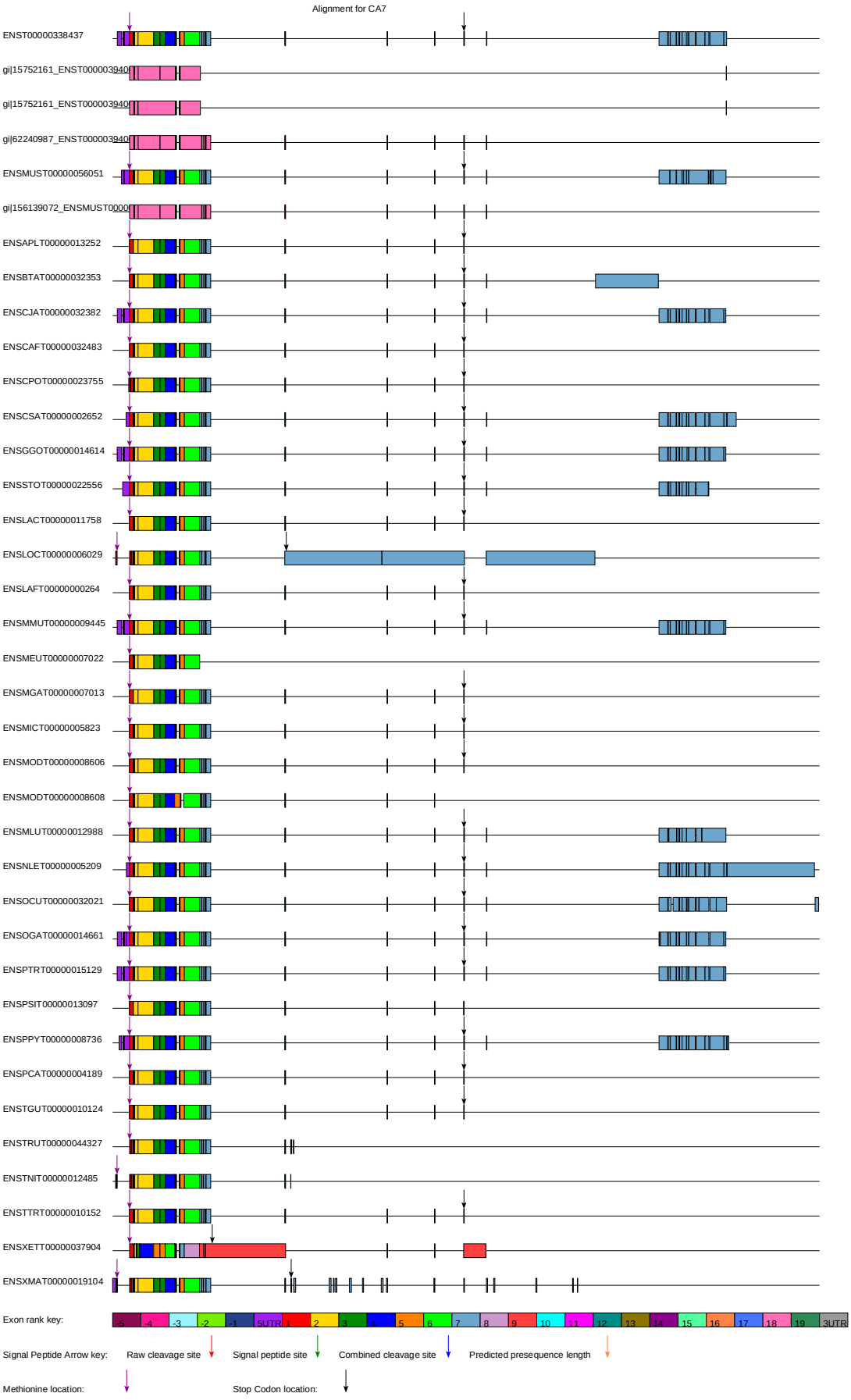

Figure 23 : Exon MSA schematic with conserved CA7 sequences and proposed NCBI nr and EST replacement transcripts for non-conserved Ensembl transcripts found by BLAST.

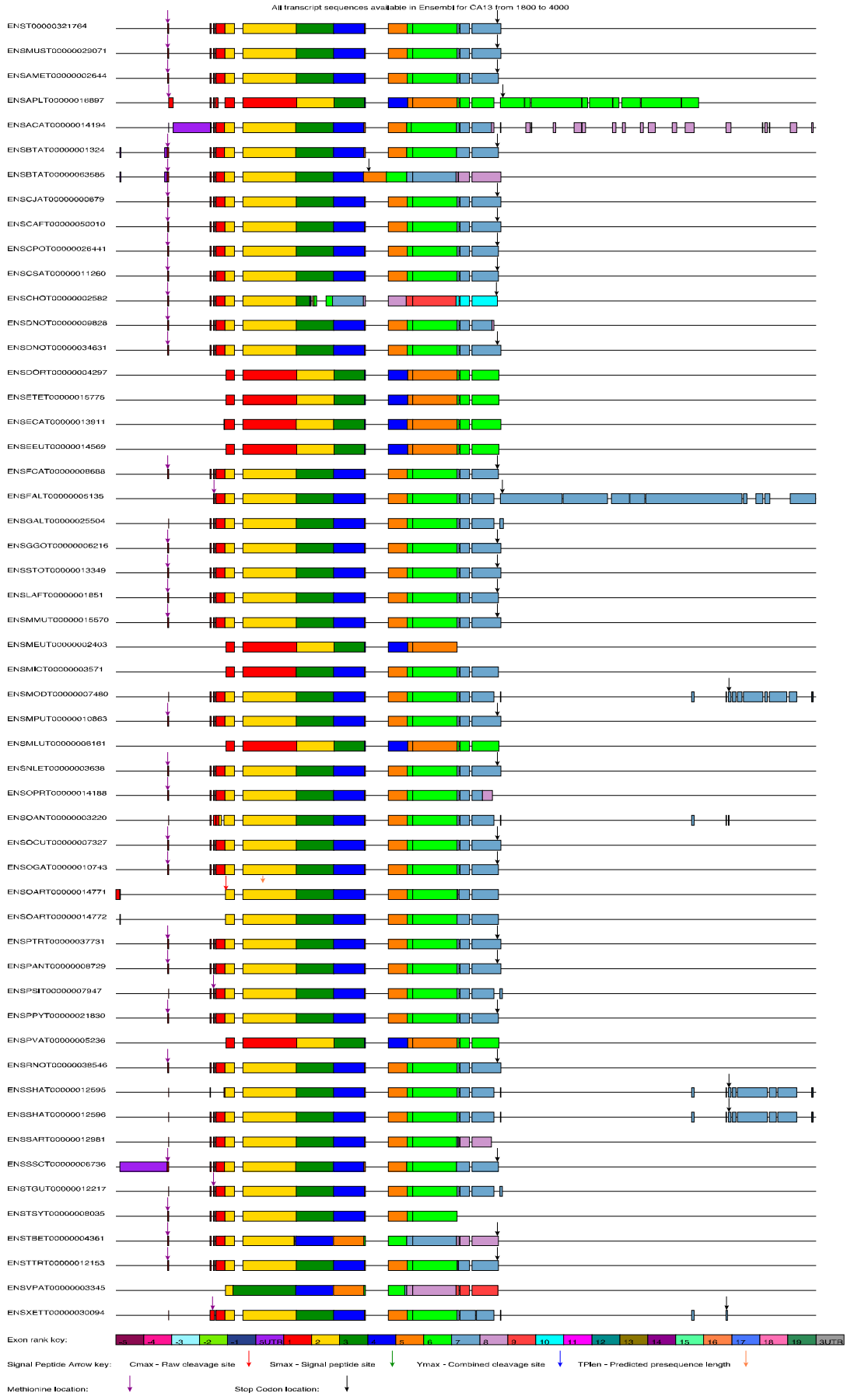

Figure 24 : Exon MSA schematic for all protein coding transcripts of CA13 in Ensembl. Zoomed in from position 1800 to 4000 in the PRANK MSA.

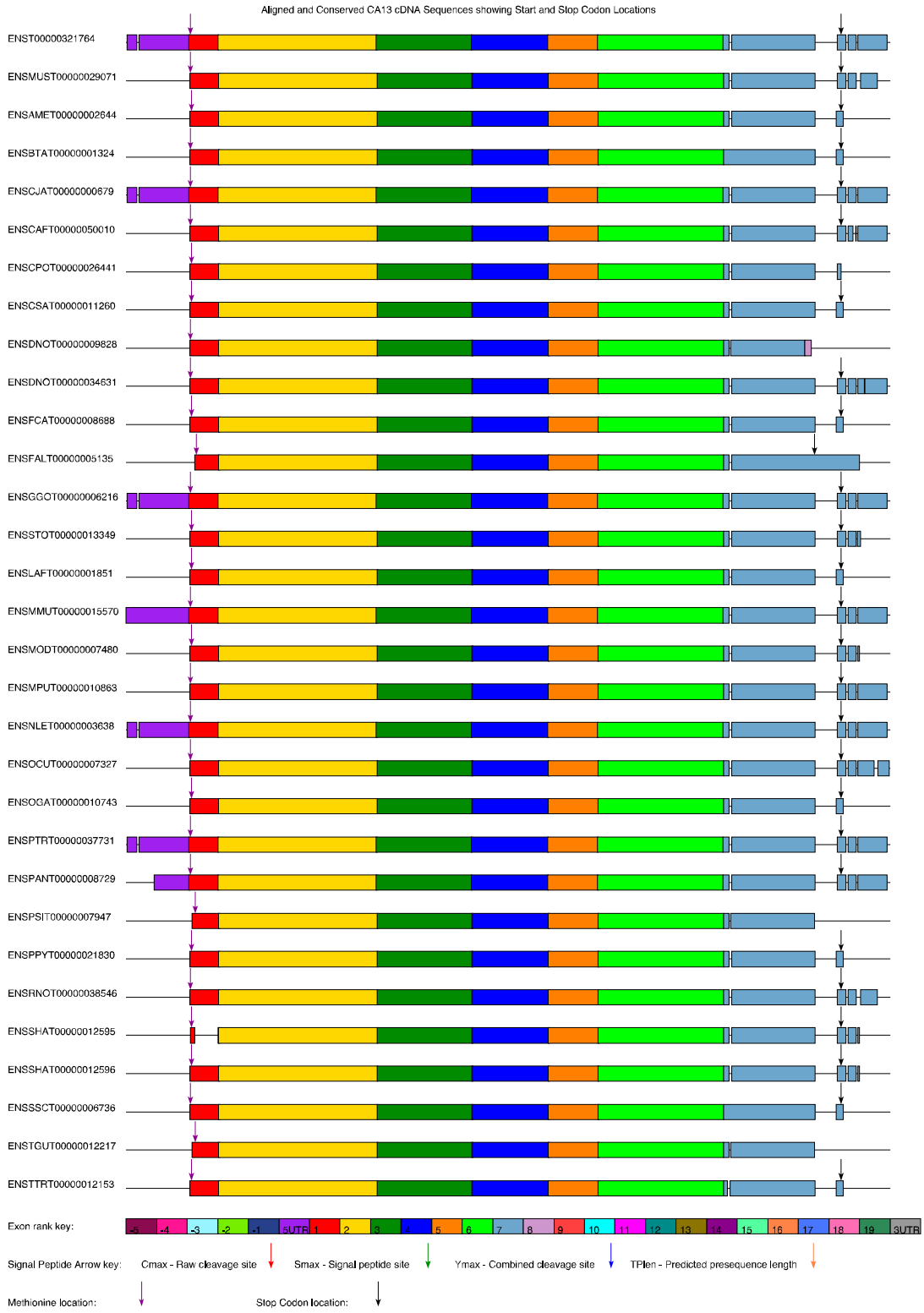

Figure 25 : Exon MSA schematic for CA13 showing conserved sequences. Zoomed in from position 1500 to 2450 in the PRANK MSA.

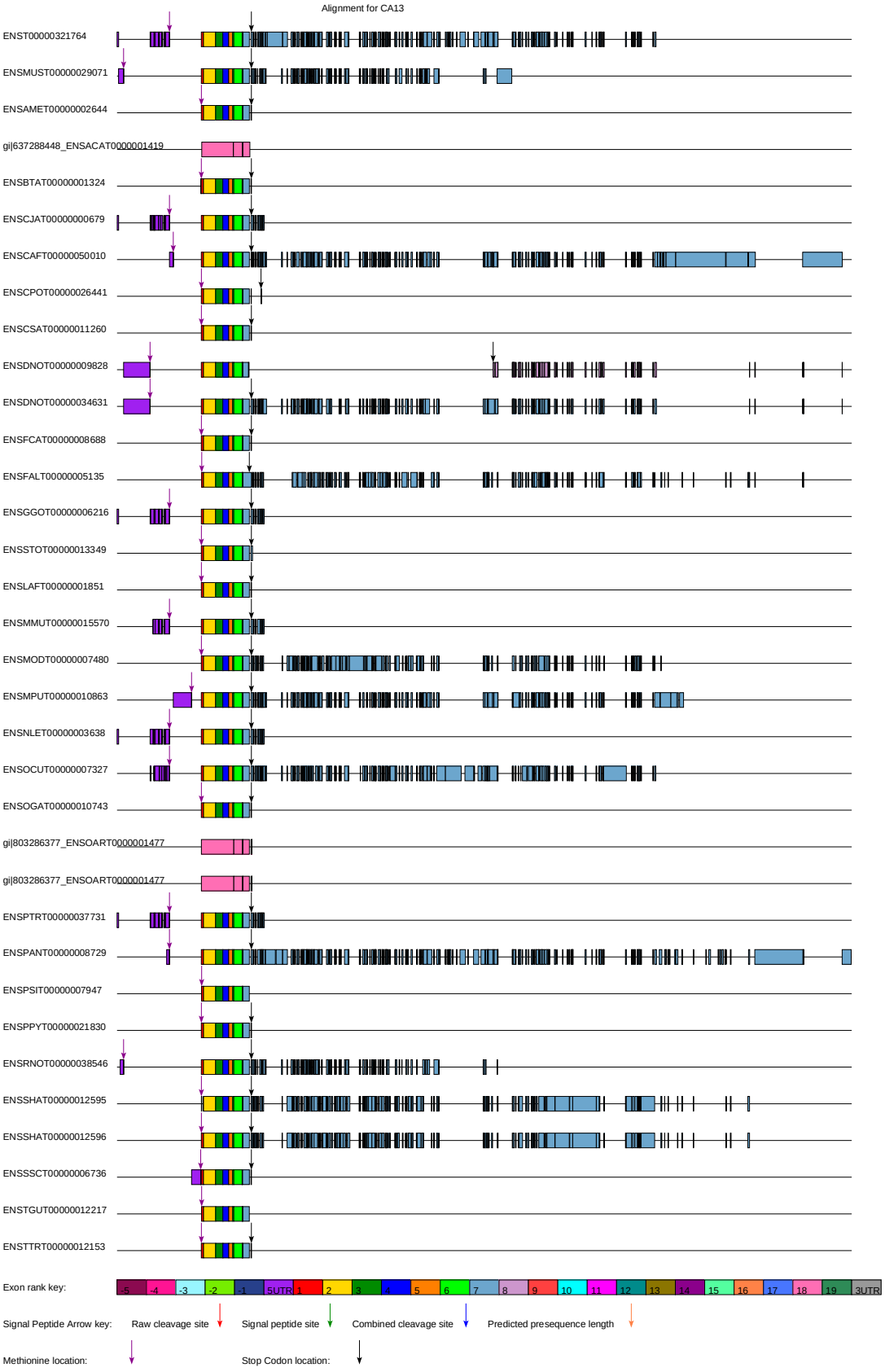

Figure 26 : Exon MSA schematic of conserved CA13 Ensembl sequences and proposed replacement sequences from the NCBI nr and EST databases found by BLAST.

4.2 The CA related proteins (CARPs) - CA8, CA10 and CA11

The following exon MSA schematics (Figure 27 to Figure 32) contain the zoomed images of various PRANK alignments of the CARPs. Not all CARPs have predicted signal peptides (Table 9) - only CA10 and CA11 have somewhat less than half of the transcripts with signal peptides that are detailed in Table 12 and Table 13. An exhaustive list of the short exons of 11 residues or less is contained in Table 11.

Table 9 : The signal peptide statistics for the CARPs.

CA	Number of transcripts in isoform	Exon containing cleavage site	Percentage
CA8	70		100%
CA10	50		56%
	37	1	42%
	2	2	2%
CA11	33		59%
	8	1	14%
	10	2	18%
	3	3	5%
	1	4	2%
	1	5	2%

Table 10 : A summary of the short exons that were 11 residues or less long found within the CARP transcripts. The last 3 columns (reading from left to right) list the number of transcripts where the 2nd last exon is short, at least one of the last three exons are short and the number of transcripts which have at least one of either the 3rd, 4th or 5th exons short (Early short exon).

	Short Exon Numbers	Protein Coding	% With Short Exon(s)	Last Exon Short	2nd last exon short	Short exon within last 3	Early short exon
All CARP Transcripts	76	199	38%	50	6	64	16
CA8	17	70	24%	0	2	5	6
CA10	41	89	46%	34	3	42	7
CA11	18	40	45%	16	1	17	3

Table 11 : An exhaustive list of CARP transcripts containing short exons that are less than 12 residues long.

CA	EnsTranscriptId	Short Exon Num	Last Exon Num	Bases	Species	Start Codon Location	SignalP	Loc	RC
CA8	ENSAMXT00000006325	1	10	17	Astyanax mexicanus	18	N	_	2
CA8	ENSAMXT00000006325	8	10	13	Astyanax mexicanus	18	N	_	2
CA8	ENSAMXT00000006325	9	10	20	Astyanax mexicanus	18	N	_	2
CA8	ENSCJAT00000017400	2	11	15	Callithrix jacchus	0	N	_	2
CA8	ENSCJAT00000017400	3	11	32	Callithrix jacchus	0	N	_	2
CA8	ENSETET00000002554	1	9	16	Echinops telfairi		N	_	1
CA8	ENSETET00000002554	2	9	14	Echinops telfairi		N	_	1
CA8	ENSEEUT00000001451	2	9	13	Erinaceus europaeus	0	N	_	1
CA8	ENSGMOT00000015239	1	11	22	Gadus morhua	21	N	_	1
CA8	ENSGMOT00000015239	4	11	10	Gadus morhua	21	N	_	1
CA8	ENSGGOT00000029327	2	10	24	Gorilla gorilla	0	N	_	2

CA	EnsTranscriptId	Short Exon Num	Last Exon Num	Bases	Species	Start Codon Location	SignalP	Loc	RC
CA8	ENSLOCT00000006192	2	10	28	Lepisosteus oculatus	579	N	_	5
CA8	ENSLAFT00000028157	2	8	24	Loxodonta africana	0	N	_	2
CA8	ENSMICT00000001196	1	9	30	Microcebus murinus		N	_	1
CA8	ENSOANT00000008004	1	9	18	Ornithorhynchus anatinus		N	S	5
CA8	ENSOCUT00000028785	2	13	24	Oryctolagus cuniculus	0	N	_	2
CA8	ENSOCUT00000028785	5	13	23	Oryctolagus cuniculus	0	N	_	2
CA8	ENSOCUT00000028785	6	13	19	Oryctolagus cuniculus	0	N	_	2
CA8	ENSOCUT00000028785	7	13	21	Oryctolagus cuniculus	0	N	_	2
CA8	ENSOCUT00000028785	8	13	15	Oryctolagus cuniculus	0	N	_	2
CA8	ENSOCUT00000028785	9	13	26	Oryctolagus cuniculus	0	N	_	2
CA8	ENSOART00000017424	1	8	22	Ovis aries		N	_	1
CA8	ENSSART00000007254	1	9	2	Sorex araneus		N	_	3
CA8	ENSSSCT00000006830	8	11	32	Sus scrofa	346	N	_	1
CA8	ENSSSCT00000006830	9	11	27	Sus scrofa	346	N	_	1
CA8	ENSSSCT00000006830	10	11	11	Sus scrofa	346	N	_	1
CA8	ENSTRUT00000009757	5	9	32	Takifugu rubripes	0	N	_	1
CA8	ENSTRUT00000009758	5	10	24	Takifugu rubripes	17	N	_	1
CA8	ENSTRUT00000009758	6	10	10	Takifugu rubripes	17	N	_	1
CA8	ENSTBET00000008598	2	9	11	Tupaia belangeri		N	_	1
CA8	ENSTBET00000008598	3	9	14	Tupaia belangeri		N	_	1
CA10	ENSAMET00000018635	8	8	20	Ailuropoda melanoleuca		N	_	2
CA10	ENSAMET00000018638	3	4	13	Ailuropoda melanoleuca		N	_	3
CA10	ENSAMXT00000005715	7	7	23	Astyanax mexicanus	0	N	_	4
CA10	ENSBTAT00000049111	9	9	23	Bos taurus	0	Y	S	1
CA10	ENSCPOT00000015643	9	9	20	Cavia porcellus	0	Y	S	1
CA10	ENSCHOT00000009034	1	15	3	Choloepus hoffmanni		N	_	5
CA10	ENSCHOT00000009034	3	15	15	Choloepus hoffmanni		N	_	5
CA10	ENSCHOT00000009034	4	15	3	Choloepus hoffmanni		N	_	5
CA10	ENSCHOT00000009034	5	15	13	Choloepus hoffmanni		N	_	5
CA10	ENSCHOT00000009034	6	15	6	Choloepus hoffmanni		N	_	5
CA10	ENSCHOT00000009034	7	15	26	Choloepus hoffmanni		N	_	5
CA10	ENSCHOT00000009034	8	15	7	Choloepus hoffmanni		N	_	5
CA10	ENSCHOT00000009034	13	15	24	Choloepus hoffmanni		N	_	5
CA10	ENSCHOT00000009034	14	15	27	Choloepus hoffmanni		N	_	5
CA10	ENSCINT00000021221	6	6	24	Ciona intestinalis		N	S	1
CA10	ENSDART00000074540	-1	9	32	Danio rerio	376	Y	S	1
CA10	ENSDORT00000011701	1	11	3	Dipodomys ordii		N	_	4
CA10	ENSDORT00000011701	2	11	18	Dipodomys ordii		N	_	4
CA10	ENSDORT00000011701	11	11	23	Dipodomys ordii		N	_	4
CA10	ENSETET00000015389	8	8	23	Echinops telfairi		N	_	2
CA10	ENSECAT00000013652	10	10	15	Equus caballus	512	Y	S	1
CA10	ENSEEUT00000001024	9	9	23	Erinaceus europaeus		N	_	2
CA10	ENSFCAT00000002278	9	9	20	Felis catus	0	Y	S	1
CA10	ENSFALT00000000851	1	8	30	Ficedula albicollis		N	_	2
CA10	ENSFALT00000000851	2	8	24	Ficedula albicollis		N	_	2
CA10	ENSGACT00000020141	7	7	23	Gasterosteus aculeatus	0	N	_	4
CA10	ENSGGOT00000015508	8	8	23	Gorilla gorilla	0	N	_	5
CA10	ENSMMUT00000019289	9	9	23	Macaca mulatta	634	Y	S	1
CA10	ENSMEUT00000012887	2	10	6	Macropus eugenii		N	_	2
CA10	ENSMEUT00000012887	9	10	21	Macropus eugenii		N	_	2
CA10	ENSMEUT00000012887	10	10	2	Macropus eugenii		N	_	2
CA10	ENSMGAT00000009462	5	5	23	Meleagris gallopavo		N	_	1
CA10	ENSMICT00000009761	1	10	24	Microcebus murinus		N	_	3
CA10	ENSMICT00000009761	2	10	6	Microcebus murinus		N	_	3
CA10	ENSMLUT00000016229	5	5	23	Myotis lucifugus		N	_	1
CA10	ENSOPRT00000012063	8	8	23	Ochotona princeps		N	_	2
CA10	ENSOGAT00000004993	9	9	23	Otolemur garnettii		N	_	5
CA10	ENSOART00000004411	5	5	23	Ovis aries		N	_	2

CA	EnsTranscriptId	Short Exon Num	Last Exon Num	Bases	Species	Start Codon Location	SignalP	Loc	RC
CA10	ENSPANT00000000051	6	6	23	Papio anubis	0	Y	M	5
CA10	ENSPMAT00000003458	3	3	23	Petromyzon marinus		N	_	2
CA10	ENSPMAT00000009940	4	4	20	Petromyzon marinus		N	_	3
CA10	ENSPCAT00000002023	1	11	27	Procavia capensis		N	_	2
CA10	ENSPCAT00000002023	3	11	1	Procavia capensis		N	_	2
CA10	ENSPCAT00000002023	8	11	5	Procavia capensis		N	_	2
CA10	ENSPCAT00000002023	9	11	15	Procavia capensis		N	_	2
CA10	ENSPVAT00000013905	1	11	18	Pteropus vampyrus		N	_	4
CA10	ENSPVAT00000013905	11	11	23	Pteropus vampyrus		N	_	4
CA10	ENSRNOT00000073315	4	4	30	Rattus norvegicus	0	Y	S	1
CA10	ENSRNOT00000075163	9	9	20	Rattus norvegicus	0	Y	S	1
CA10	ENSSART00000001115	8	8	23	Sorex araneus		N	_	2
CA10	ENSSSCT00000019155	2	13	20	Sus scrofa	12	N	_	5
CA10	ENSSSCT00000019155	3	13	9	Sus scrofa	12	N	_	5
CA10	ENSSSCT00000019155	4	13	18	Sus scrofa	12	N	_	5
CA10	ENSSSCT00000019155	6	13	11	Sus scrofa	12	N	_	5
CA10	ENSSSCT00000019155	7	13	26	Sus scrofa	12	N	_	5
CA10	ENSSSCT00000019155	8	13	13	Sus scrofa	12	N	_	5
CA10	ENSSSCT00000019155	9	13	16	Sus scrofa	12	N	_	5
CA10	ENSSSCT00000019155	10	13	18	Sus scrofa	12	N	_	5
CA10	ENSTRUT00000007886	6	6	23	Takifugu rubripes		N	_	4
CA10	ENSTSYT00000003011	7	11	26	Tarsius syrichta	0	N	_	4
CA10	ENSTSYT00000003011	8	11	12	Tarsius syrichta	0	N	_	4
CA10	ENSTSYT00000003011	11	11	23	Tarsius syrichta	0	N	_	4
CA10	ENSTNIT00000008996	6	6	23	Tetraodon nigroviridis		N	_	3
CA10	ENSTBET00000014768	8	8	23	Tupaia belangeri		N	_	2
CA10	ENSTTRT00000001682	9	9	23	Tursiops truncatus	0	N	_	3
CA10	ENSVPAT00000008163	4	12	3	Vicugna pacos	0	N	S	4
CA10	ENSVPAT00000008163	12	12	23	Vicugna pacos	0	N	S	4
CA10	ENSXETT00000002775	1	9	21	Xenopus tropicalis		Y	S	1
CA10	ENSXETT00000002775	9	9	26	Xenopus tropicalis		Y	S	1
CA10	ENSXMAT00000016670	7	7	23	Xiphophorus maculatus		N	_	3
CA11	ENSMUST00000126106	5	5	19	Mus musculus				
CA11	ENSMUST00000137294	4	4	19	Mus musculus				
CA11	ENSAMET00000016276	9	9	26	Ailuropoda melanoleuca	0	Y	S	3
CA11	ENSBTAT00000008890	9	9	26	Bos taurus	331	Y	S	3
CA11	ENSCAFT00000006320	9	9	26	Canis familiaris	0	Y	S	3
CA11	ENSCPOT00000012860	9	9	26	Cavia porcellus	27	Y	S	5
CA11	ENSDNOT00000007145	9	9	26	Dasypus novemcinctus	0	Y	S	3
CA11	ENSDORT00000010600	9	9	26	Dipodomys ordii	0	Y	S	2
CA11	ENSECAT00000019400	7	8	16	Equus caballus	195	Y	S	1
CA11	ENSEEUT00000009287	1	16	16	Erinaceus europaeus		Y	S	5
CA11	ENSEEUT00000009287	2	16	9	Erinaceus europaeus		Y	S	5
CA11	ENSEEUT00000009287	3	16	17	Erinaceus europaeus		Y	S	5
CA11	ENSEEUT00000009287	4	16	14	Erinaceus europaeus		Y	S	5
CA11	ENSEEUT00000009287	5	16	8	Erinaceus europaeus		Y	S	5
CA11	ENSEEUT00000009287	6	16	3	Erinaceus europaeus		Y	S	5
CA11	ENSEEUT00000009287	8	16	30	Erinaceus europaeus		Y	S	5
CA11	ENSEEUT00000009287	16	16	26	Erinaceus europaeus		Y	S	5
CA11	ENSSTOT00000021195	9	9	26	Ictidomys tridecemlineatus	42	Y	S	4
CA11	ENSLACT00000005065	8	8	23	Latimeria chalumnae	96	N	S	5
CA11	ENSLAFT00000035430	9	9	26	Loxodonta africana	0	Y	S	2
CA11	ENSMEUT00000000667	2	8	19	Macropus eugenii	0	Y	S	2
CA11	ENSMICT00000016328	9	9	26	Microcebus murinus	0	Y	S	3
CA11	ENSMPUT00000003886	9	9	26	Mustela putorius furo	556	Y	S	2
CA11	ENSPANT00000006422	9	9	26	Papio anubis	791	Y	S	3
CA11	ENSPVAT00000009529	3	7	22	Pteropus vampyrus		N	M	1
CA11	ENSPVAT00000009529	7	7	26	Pteropus vampyrus		N	M	1

Table 12 : SignalP and TargetP results for CA10

TranscriptId	Cmax_Pos	Ymax_Pos	Smax_Pos	TPlen	Loc	RC
ENST00000285273	2	2	1	2	S	1
ENST00000442502	2	2	1	2	S	1
ENST00000451037	2	2	1	2	S	1
ENST00000575181	2	2	1	2	S	1
ENSMUST00000042943	2	2	1	2	S	1
ENSMUST00000092780	2	2	1	2	S	1
ENSMUST00000107858	2	2	1	2	S	1
ENSMUST00000107859	2	2	1	2	S	1
ENSMUST00000107861	2	2	1	2	S	1
ENSMUST00000107863	2	2	1	2	S	1
ENSACAT00000017448	2	2	1	2	S	1
ENSBTAT00000049111	2	2	1	2	S	1
ENSCJAT00000022193	2	2	1	2	S	1
ENSCJAT00000022198	2	2	1	2	S	1
ENSCAFT00000044667	2	2	1	2	S	1
ENSCPOT00000015643	2	2	1	2	S	1
ENSCSAT00000003558	1	1	1	1	M	5
ENSDART00000074540	2	2	1	2	S	1
ENSDART00000133188	2	2	1	2	S	2
ENSDNOT00000034568	2	2	1	2	S	1
ENSECAT00000013652	2	2	1	2	S	1
ENSFCAT00000002278	2	2	1	2	S	1
ENSGALT00000004742	2	2	1	2	S	1
ENSSTOT00000016380	2	2	1	2	S	1
ENSLACT00000010843	2	2	1	2	S	1
ENSLOCT00000013627	2	2	1	2	S	1
ENSMMUT00000019289	2	2	1	2	S	1
ENSMPUT00000015741	2	2	1	2	S	1
ENSNLET00000010168	2	2	1	2	S	1
ENSONIT00000024573	2	2	1	2	S	1
ENSOCUT00000002346	2	2	1	2	S	1
ENSPTRT00000017287	2	2	1	2	S	1
ENSPANT00000000051	1	1	1	1	M	5
ENSPSIT00000016603	2	2	1	2	S	1
ENSPFOT00000006415	2	2	1	2	S	1
ENSPPYT00000009660	2	2	1	2	S	1
ENSRNOT00000073315	2	2	1	2	S	1
ENSRNOT00000075163	2	2	1	2	S	1
ENSXETT00000002775	2	2	2	2	S	1

Table 13 : SignalP and TargetP results for CA11

TranscriptId	Cmax_Pos	Ymax_Pos	Smax_Pos	TPlen	Loc	RC
ENSMUST00000032129	3	3	2	2	S	1
ENSAPLT00000016929	2	2	2	2	S	2
ENSBTAT00000017203	2	2	2	2	S	1
ENSCPOT00000014323	2	2	2	2	S	1
ENSDORT00000013219	1	1	1	1	S	3
ENSDORT00000014384	1	1	1	1	S	2
ENSDORT00000015978	5	5	4	4	S	2
ENSECAT00000024766	2	2	2	2	S	1
ENSFALT00000000232	1	1	1	1	S	1
ENSMGAT00000000662	1	1	1	1	S	1
ENSMODT00000014980	1	1	1	1	S	1
ENSMLUT00000003668	2	2	2	2	S	5
ENSOCUT00000010226	3	3	2	2	S	1
ENSOGAT00000012180	1	1	1	1	S	2
ENSOART00000021610	2	2	2	2	S	1
ENSOART00000021611	2	2	2	2	S	1
ENSPANT00000006292	2	2	2	2	S	1
ENSPSIT00000014466	2	2	2	2	S	1
ENSPCAT00000013787	1	1	1	1	S	2
ENSPVAT00000015874	4	4	3	3	S	4
ENSRNOT00000012218	3	3	2	2	S	1
ENSSART00000007643	1	1	1	1	S	2
ENSXETT00000017970	2	2	3	2	S	1

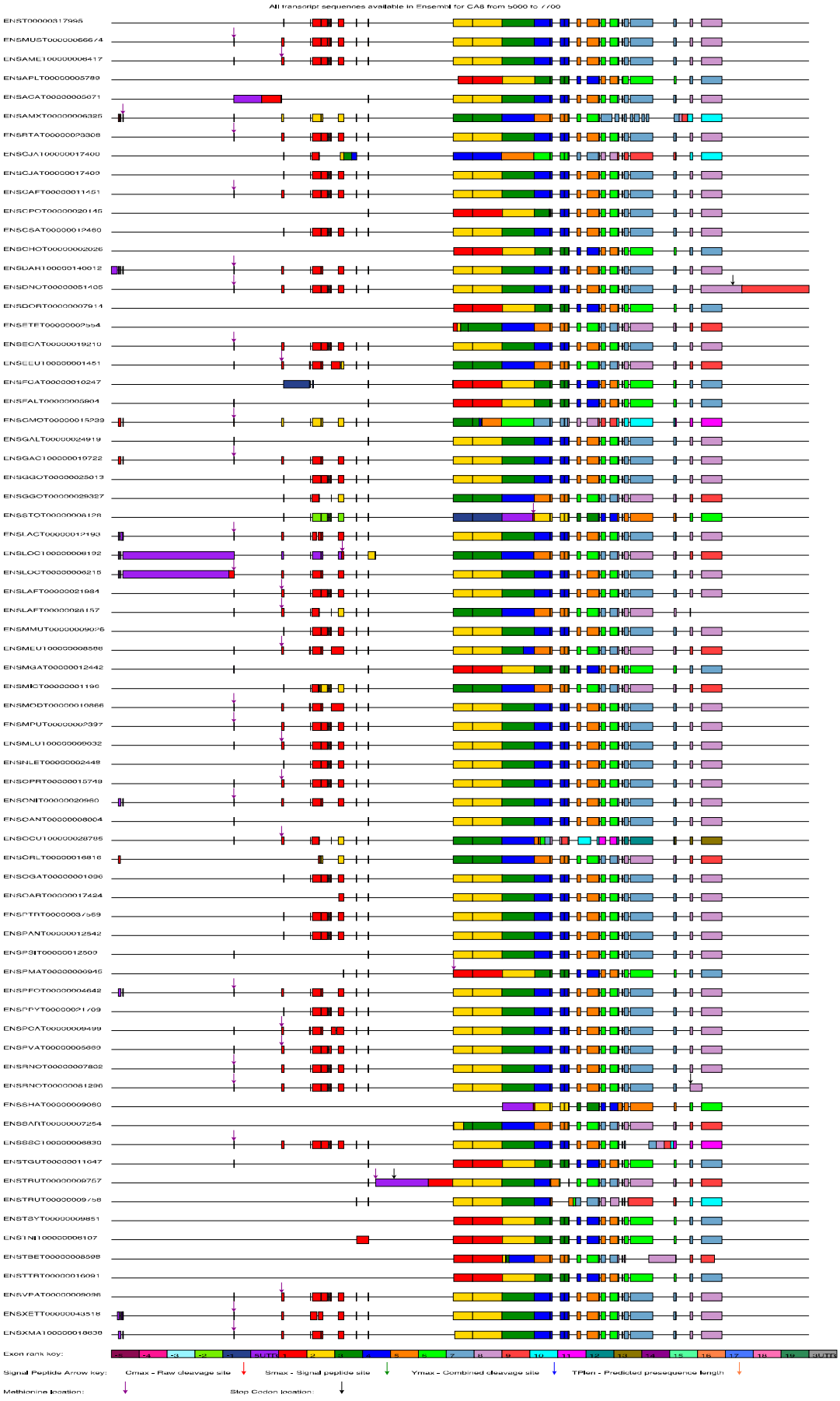

Figure 27 : Exon MSA schematic of all Ensembl cDNA sequences for CA8. The PRANK MSA is zoomed in from position 5000 to 7700.

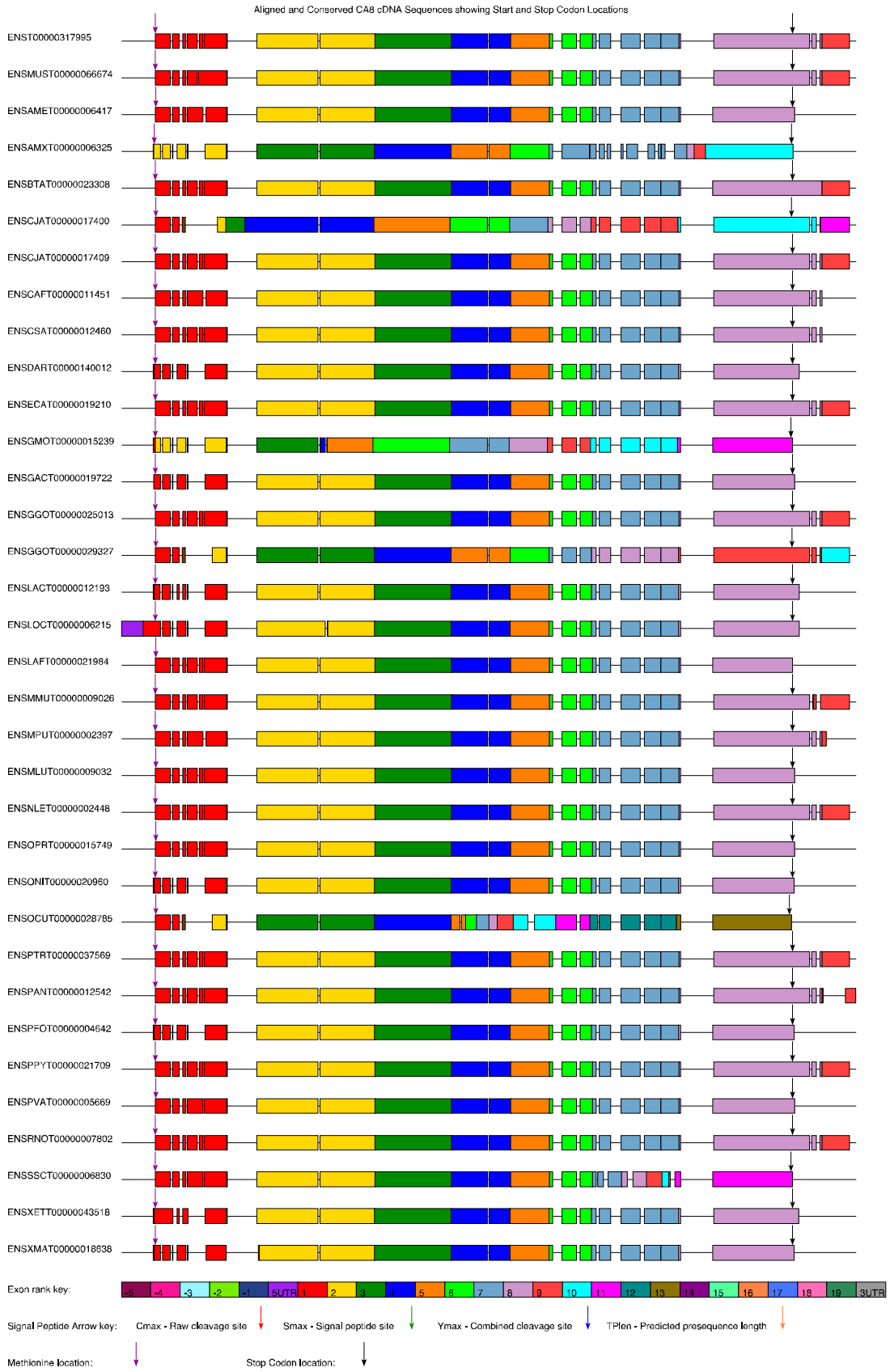

Figure 28 : Exon MSA schematic of conserved protein coding transcripts of CA8. The PRANK MSA is zoomed in from position 2900 to 4100.

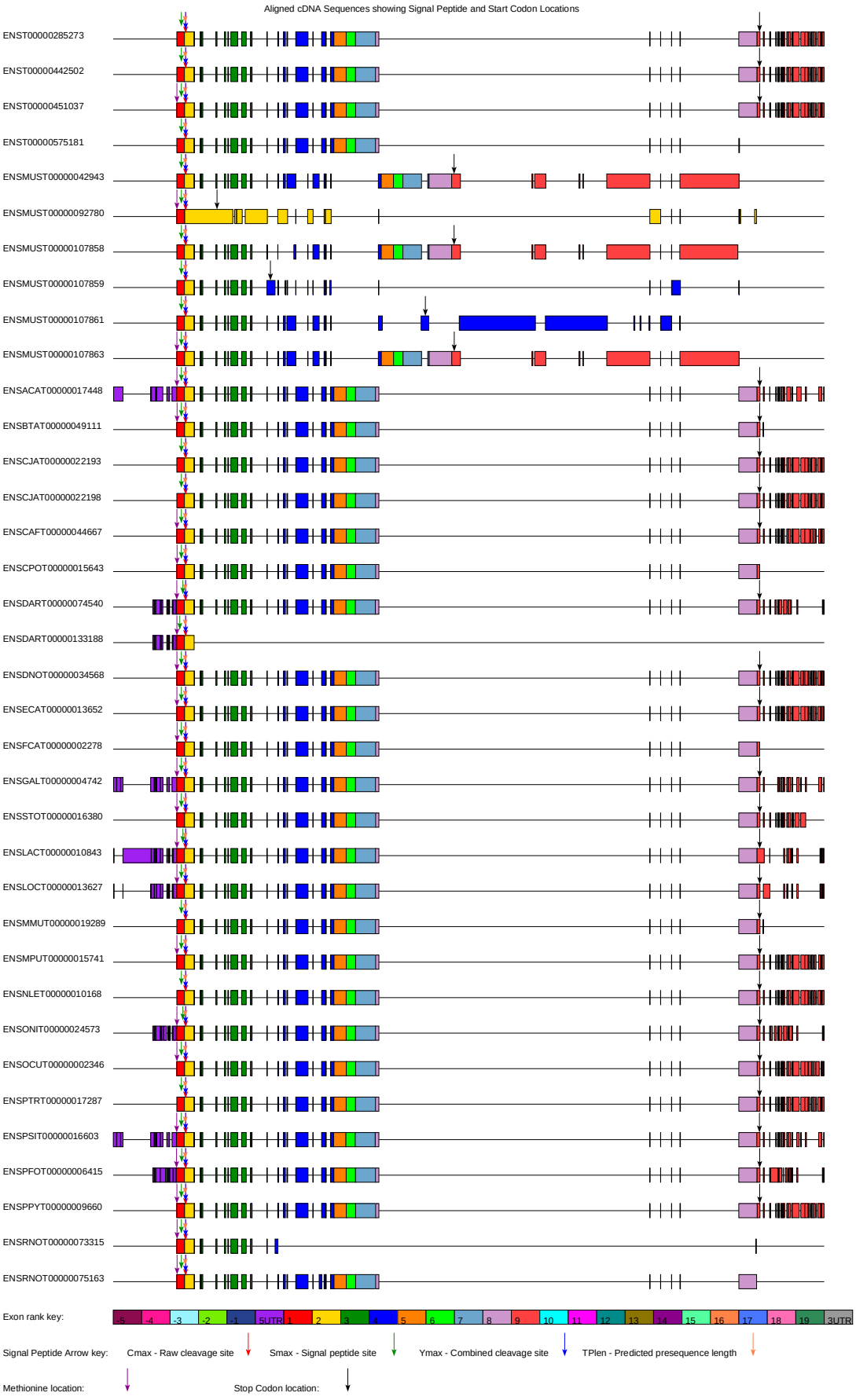

Figure 29 : Exon MSA schematic for CA10 transcripts that have signal peptides and start and stop codons. The PRANK MSA is zoomed in from position 5000 to 10500.

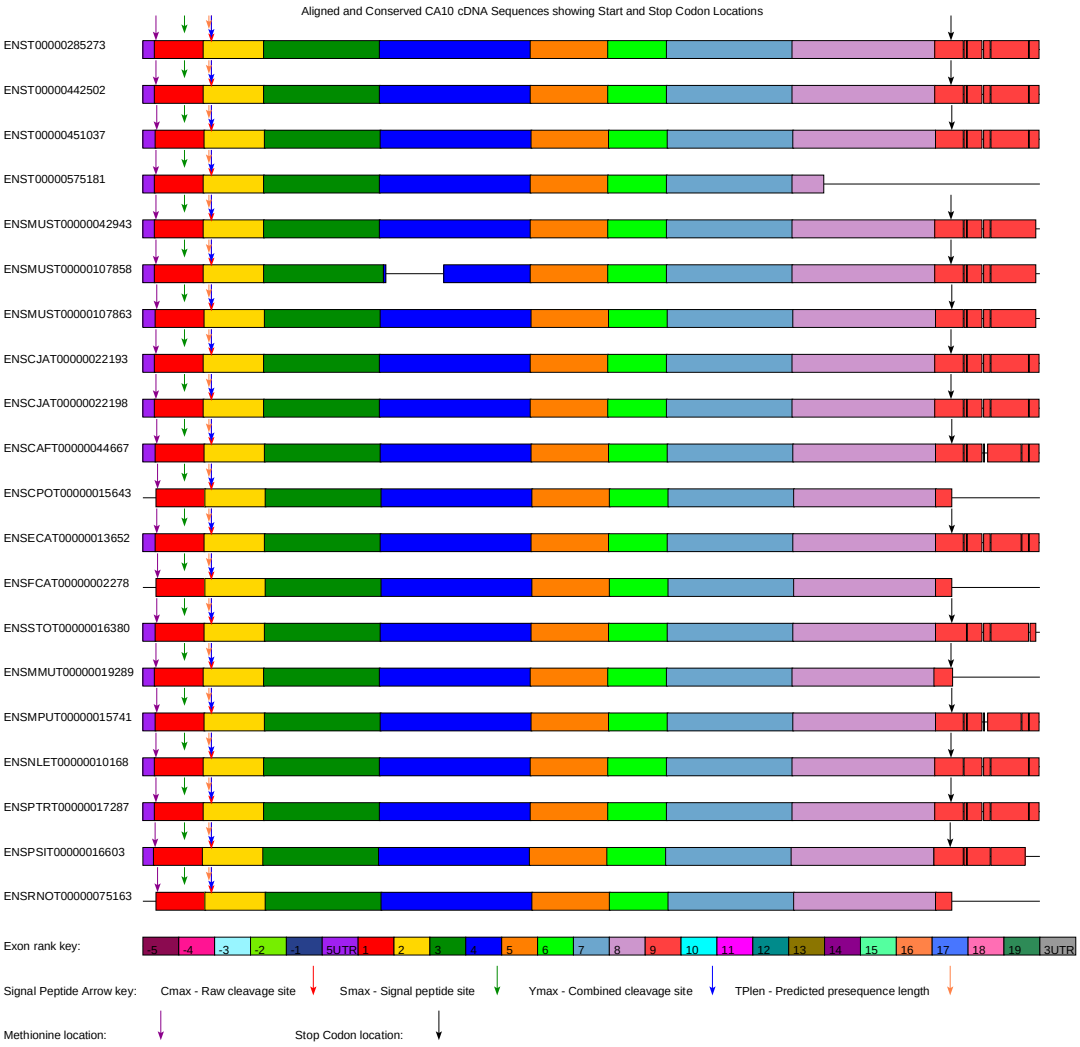

Figure 30 : Exon MSA schematic of conserved CA10 transcripts. The PRANK MSA is zoomed in from position 2700 to 3800.

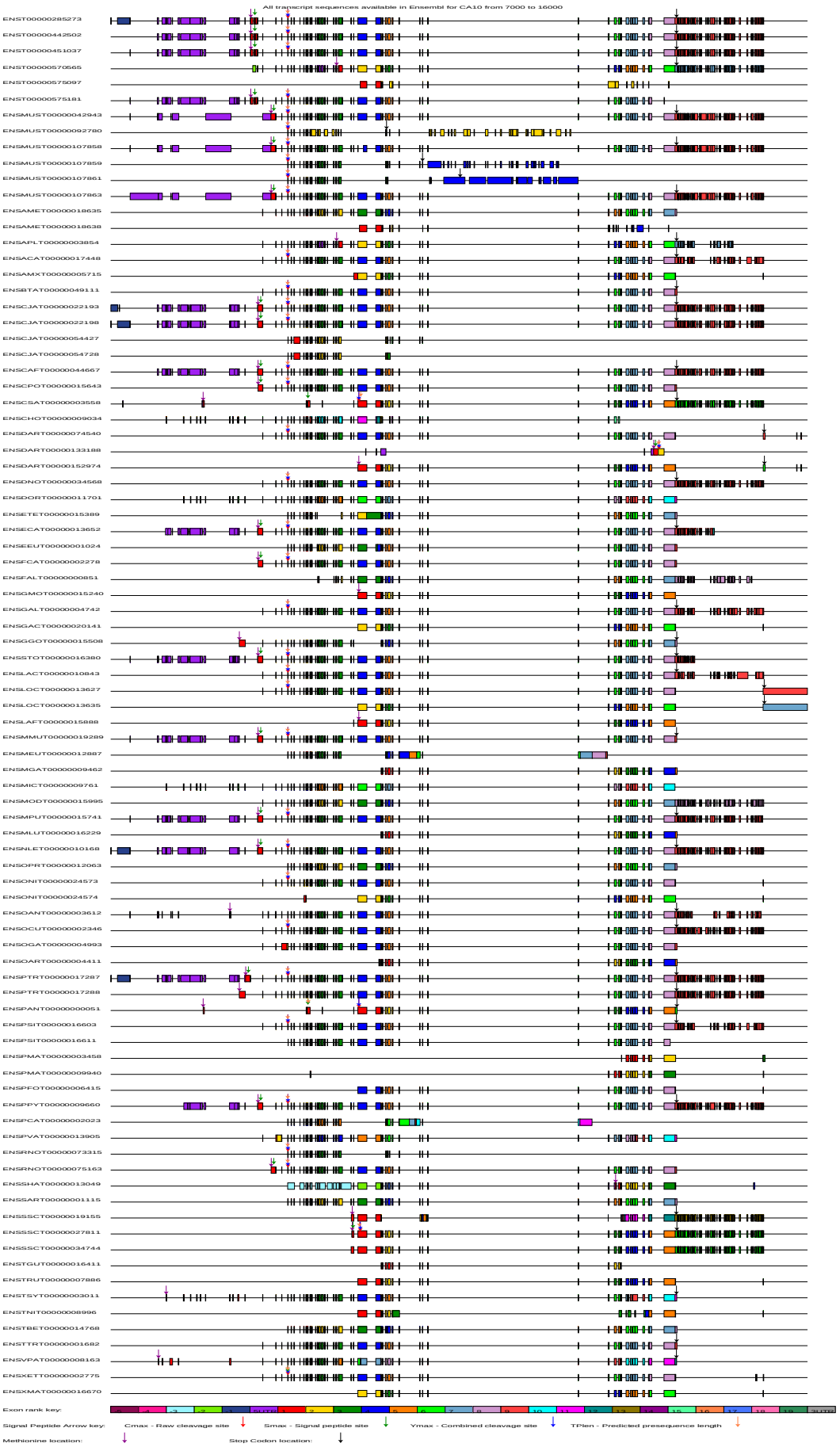

Figure 31 : Exon MSA schematic of all transcripts for CA10. Zoomed in from position 7000 to 16000.

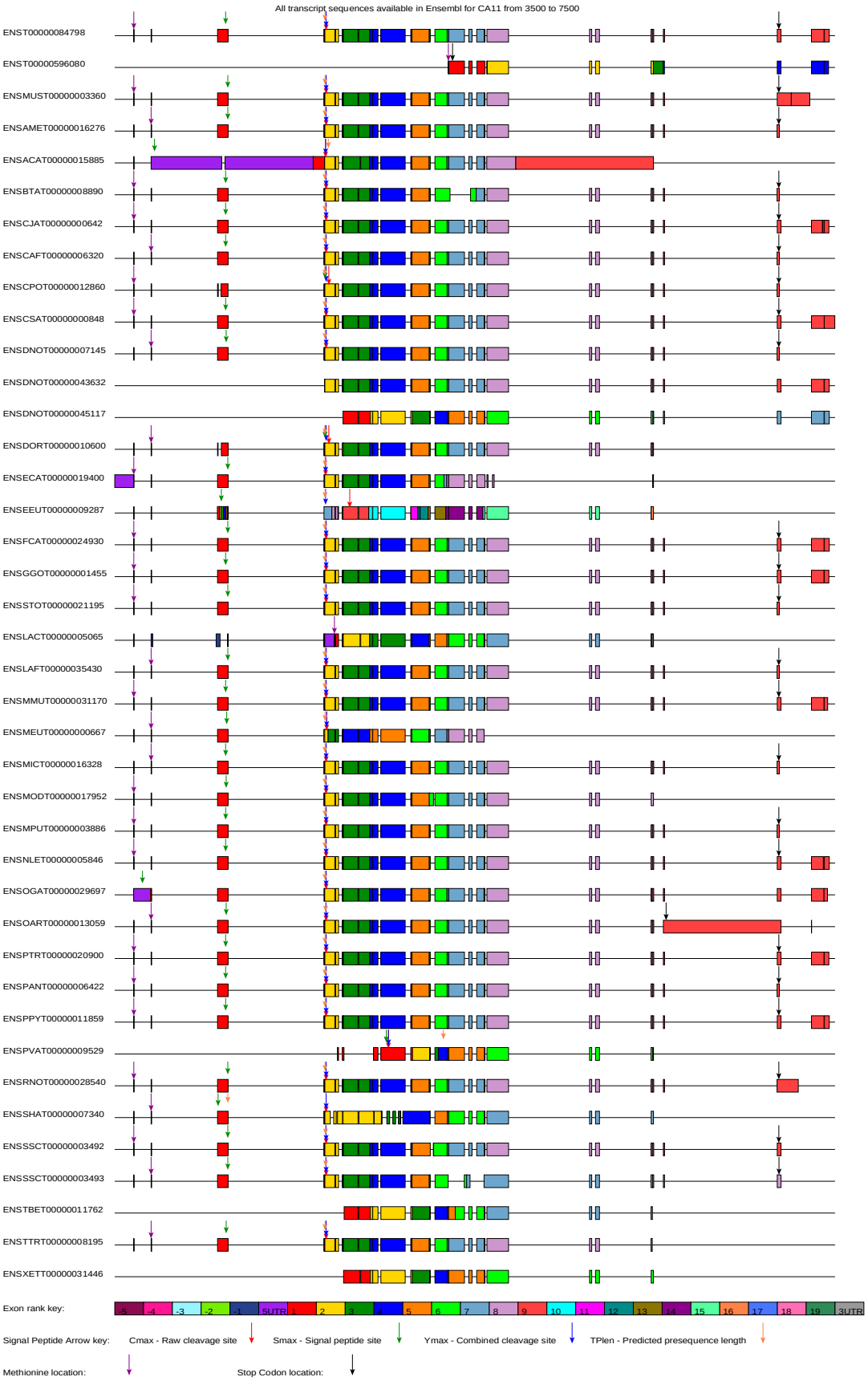

Figure 32 : Exon MSA schematic for all CA11 transcripts in Ensembl. The PRANK MSA is zoomed in from position 3500 to position 7500. Some transcripts are not included in this alignment due to Pal2Nal failing to find a matching translation of the protein sequence in the nucleotide sequence. ExonAnalysis.py skips over these transcripts, thus not including them in the schematic.

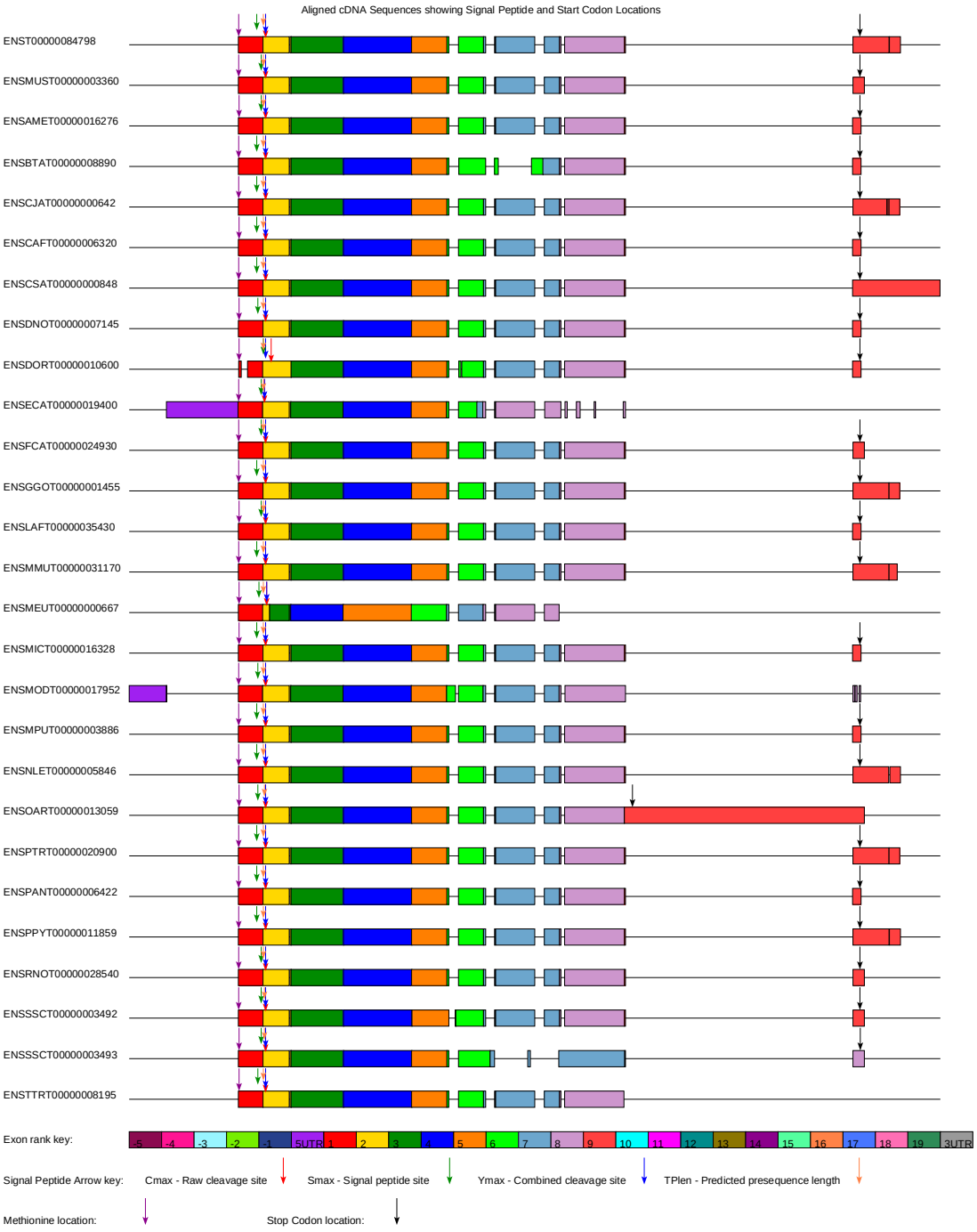

Figure 33 : Exon MSA schematic of all transcripts with start and stop codons and signal peptides. This PRANK MSA is zoomed in from position 3000 to position 5200. The location of the signal peptide is highly conserved.

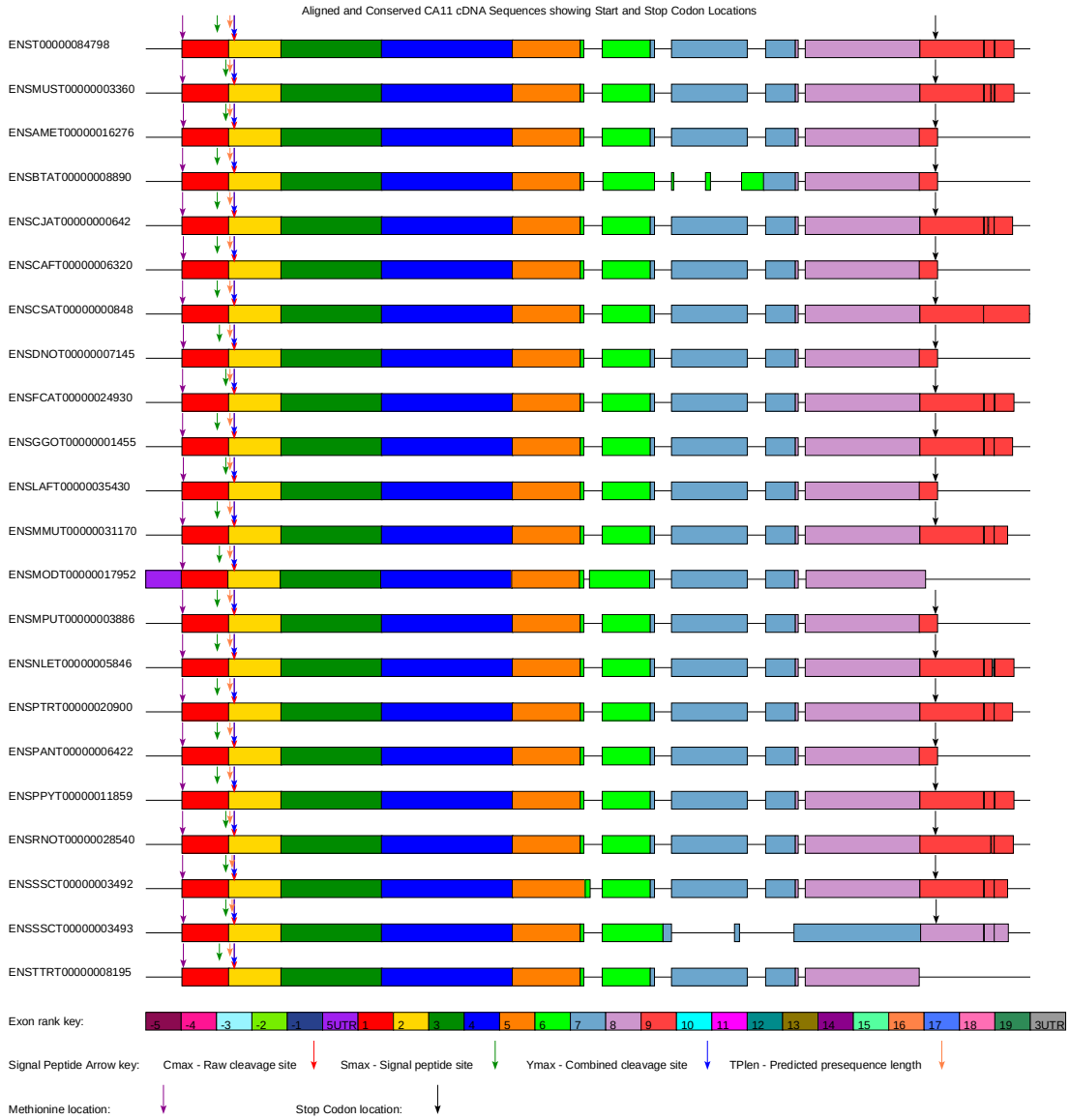

Figure 34 : Exon MSA schematic of conserved transcripts for CA11. This PRANK MSA is zoomed in from position 3050 to 4300.

4.3 The mitochondrial CAs - CA5A and CA5B

Mitochondrial targeting peptides are not as conserved as signal peptides are. Therefore the RC values for these peptides are higher. (Emanuelsson, Brunak, von Heijne, & Nielsen, 2007) Both CA5A and CA5B have the cleavage site of the targeting peptides for less than half of the transcripts mostly in exon 1 (Table 14). The detailed predicted exon locations of the mitochondrial targeting peptides are shown in Table 15 and Table 16. The exon MSA schematics of all the transcripts and the conserved transcripts are shown in Figure 36 to Figure 39.

Table 14 : Mitochondrial targeting peptide statistics for CA5A and CA5B.

CA	Number of transcripts in isoform	Exon containing cleavage site	Percentage
CA5A	14		52%
	11	1	41%
	2	2	7%
CA5B	49		68%
	23	1	32%

Table 15 : Mitochondrial targeting peptide predictions for CA5A

TranscriptId for CA5A	Exon containing predicted Tplen site	Loc	RC
ENST00000309893	1	M	5
ENSMUST00000057653	1	M	2
ENSBTAT00000013397	1	M	2
ENSCJAT00000016857	1	M	4
ENSCAFT00000031659	1	M	4
ENSDNOT00000030913	1	M	1
ENSMODT00000005505	1	M	2
ENSMLUT00000013364	1	M	2
ENSOANT00000005926	2	M	5
ENSPTRT00000015568	1	M	4
ENSPANT00000000511	1	M	3
ENSRNOT00000025848	1	M	2
ENSTTRT00000001377	2	M	2

Table 16 : TargetP results for CA5B where a mitochondrial targeting peptide has been predicted.

TranscriptId for CA5B	Exon containing predicted Tplen site	Loc	RC
ENST00000318636	1	M	2
ENSMUST00000033739	1	M	2
ENSBTAT00000027210	1	M	1
ENSCJAT00000008356	1	M	2
ENSCAFT00000019334	1	M	1
ENSCPOT00000004268	1	M	1
ENSCSAT00000013000	1	M	2

TranscriptId for CA5B	Exon containing predicted Tplen site	Loc	RC
ENSECAT00000000515	1	M	2
ENSFCAT00000029639	1	M	1
ENSGGOT00000004910	1	M	2
ENSLAFT00000009323	1	M	2
ENSMMUT00000019728	1	M	2
ENSMMUT00000019729	1	M	2
ENSMPUT00000002719	1	M	1
ENSNLET00000012110	1	M	2
ENSOGAT00000011341	1	M	2
ENSOART00000013810	1	M	1
ENSPTRT00000055072	1	M	2
ENSPANT00000002414	1	M	2
ENSPPYT00000023497	1	M	1
ENSRNOT00000047354	1	M	1
ENSSART00000000113	1	M	2
ENSSSCT00000013280	1	M	1

Just like the cytoplasmic CA isoforms, the mitochondrial CA isoforms also contain short exons that are less than 11 residues long (Table 17). Looking at the detailed list of the transcripts with short exons in Table 18, some of the transcripts are potentially mispredicted since nearly every exon within the transcript is short (ENSPCAT00000000634 and ENSDORT00000001236, both in CA5B).

Table 17 : A summary of the short exons that are at most 11 residues long found in the mitochondrial CA group. The last three columns (reading from left to right) give the number of transcripts where the 2nd last exon is short, and finally a count of transcripts where at least one of the last three exons is short and at least one of the 3rd, 4th or 5th exons are short (Early Short Exons).

	Count of Short Exons	Protein Coding	% with Short Exon(s)	Last Exon Short	2nd last exon short	Last 3 exon(s) short	Early Short Exons
Mitochondrial Transcripts	14	99	14%	5	3	15	8
CA5A	3	27	11%	3	1	4	0
CA5B	11	72	15%	2	2	11	8

Table 18 : An exhaustive list of transcripts that have exons that are at most 33 bases (11 residues) long.

CA	EnsTranscriptId	Short Exon Num	Last Exon Num	Bases	Species	Start Codon Location	SignalP	Loc	RC
CA5A	ENSDORT00000014169	8	8	9	Dipodomys ordii		N	M	2
CA5A	ENSOANT00000005926	8	9	19	Ornithorhynchus anatinus	295	N	M	5
CA5A	ENSOANT00000005926	9	9	24	Ornithorhynchus anatinus	295	N	M	5
CA5A	ENSTTRT00000001377	9	9	27	Tursiops truncatus	0	N	M	2
CA5B	ENST00000498004	-3	2	29	Homo sapiens	258	N	M	2
CA5B	ENSAMXT00000004577	4	11	27	Astyanax mexicanus		N	_	2
CA5B	ENSAMXT00000004577	8	11	20	Astyanax mexicanus		N	_	2
CA5B	ENSAMXT00000004577	9	11	6	Astyanax mexicanus		N	_	2
CA5B	ENSCJAT00000008356	9	9	20	Callithrix jacchus	54	N	M	2
CA5B	ENSDORT00000001236	1	14	30	Dipodomys ordii		N	_	2
CA5B	ENSDORT00000001236	3	14	14	Dipodomys ordii		N	_	2
CA5B	ENSDORT00000001236	4	14	3	Dipodomys ordii		N	_	2
CA5B	ENSDORT00000001236	5	14	6	Dipodomys ordii		N	_	2
CA5B	ENSDORT00000001236	6	14	9	Dipodomys ordii		N	_	2
CA5B	ENSDORT00000001236	7	14	15	Dipodomys ordii		N	_	2
CA5B	ENSDORT00000001236	8	14	9	Dipodomys ordii		N	_	2
CA5B	ENSDORT00000001236	12	14	9	Dipodomys ordii		N	_	2
CA5B	ENSGGOT00000031322	3	7	4	Gorilla gorilla		N	_	2
CA5B	ENSGGOT00000031322	4	7	23	Gorilla gorilla		N	_	2
CA5B	ENSLAFT00000009323	5	8	7	Loxodonta africana	0	Y	M	2
CA5B	ENSMPUT00000002719	6	10	21	Mustela putorius furo	460	N	M	1
CA5B	ENSMPUT00000002719	9	10	25	Mustela putorius furo	460	N	M	1
CA5B	ENSORLT00000015777	1	9	9	Oryzias latipes		N	S	5
CA5B	ENSPMAT00000010325	4	8	14	Petromyzon marinus		N	_	1
CA5B	ENSPCAT00000000634	7	18	3	Procavia capensis		N	_	2
CA5B	ENSPCAT00000000634	8	18	3	Procavia capensis		N	_	2
CA5B	ENSPCAT00000000634	9	18	15	Procavia capensis		N	_	2
CA5B	ENSPCAT00000000634	10	18	3	Procavia capensis		N	_	2
CA5B	ENSPCAT00000000634	11	18	9	Procavia capensis		N	_	2
CA5B	ENSPCAT00000000634	12	18	9	Procavia capensis		N	_	2
CA5B	ENSPCAT00000000634	13	18	12	Procavia capensis		N	_	2
CA5B	ENSPCAT00000000634	14	18	9	Procavia capensis		N	_	2
CA5B	ENSPCAT00000000634	15	18	3	Procavia capensis		N	_	2
CA5B	ENSPCAT00000000634	16	18	13	Procavia capensis		N	_	2
CA5B	ENSPCAT00000000634	17	18	5	Procavia capensis		N	_	2
CA5B	ENSPCAT00000000634	18	18	12	Procavia capensis		N	_	2
CA5B	ENSSHAT00000015991	4	14	25	Sarcophilus harrisii		N	_	5
CA5B	ENSSHAT00000015991	5	14	19	Sarcophilus harrisii		N	_	5
CA5B	ENSSHAT00000015991	6	14	8	Sarcophilus harrisii		N	_	5
CA5B	ENSSHAT00000015991	7	14	10	Sarcophilus harrisii		N	_	5
CA5B	ENSSHAT00000015991	9	14	21	Sarcophilus harrisii		N	_	5

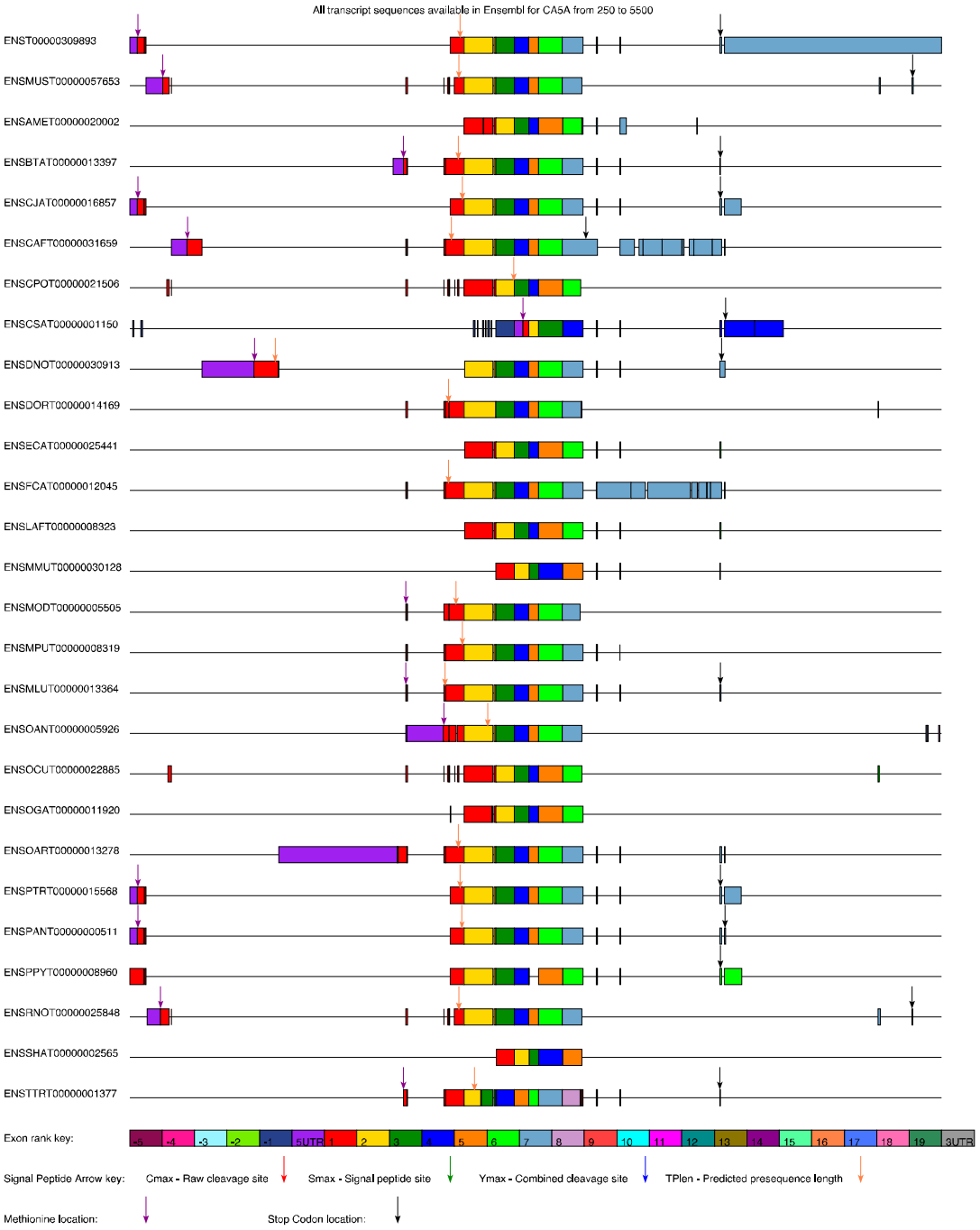

Figure 35 : Exon MSA schematic of all the Ensembl transcripts for CA5A. The PRANK MSA is zoomed in from position 250 to 5500 to capture the alignment between the start and stop codons of most of the transcripts.

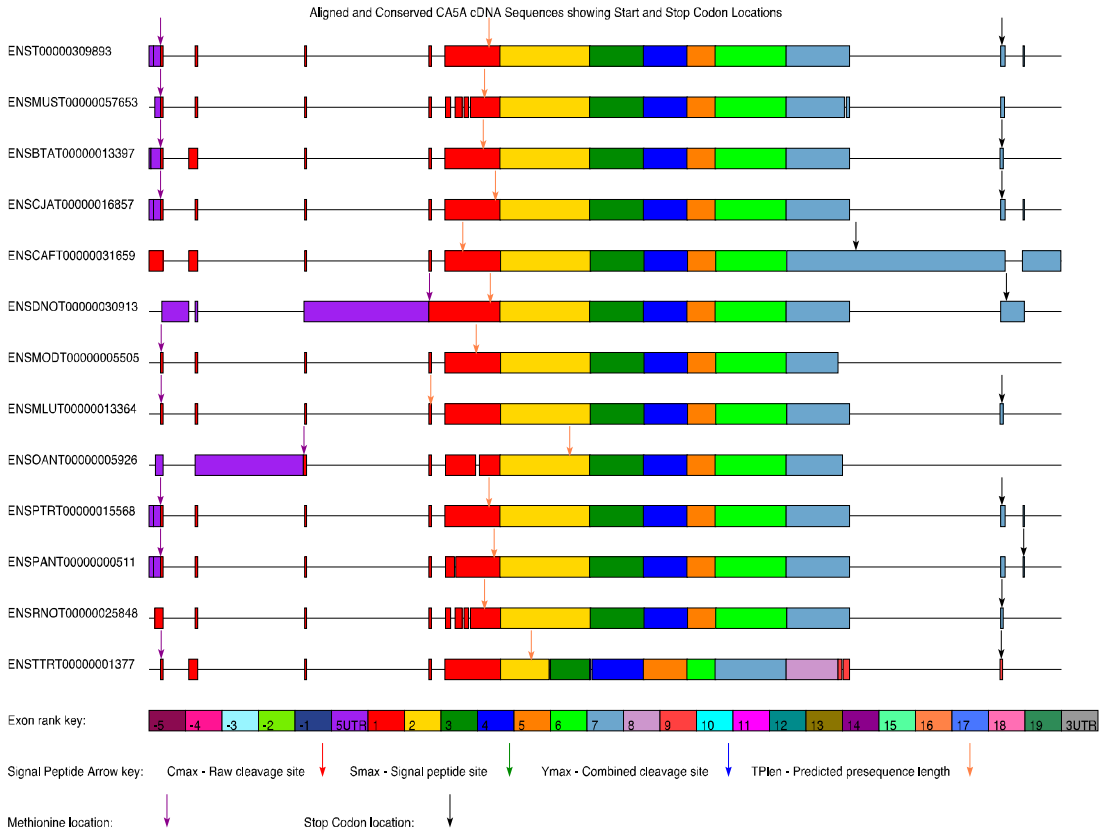

Figure 36 : Exon MSA schematic of CA5A where all the transcripts have a start codon and a predicted mitochondrial targeting peptide. The PRANK MSA is zoomed in from position 300 to position 2300.

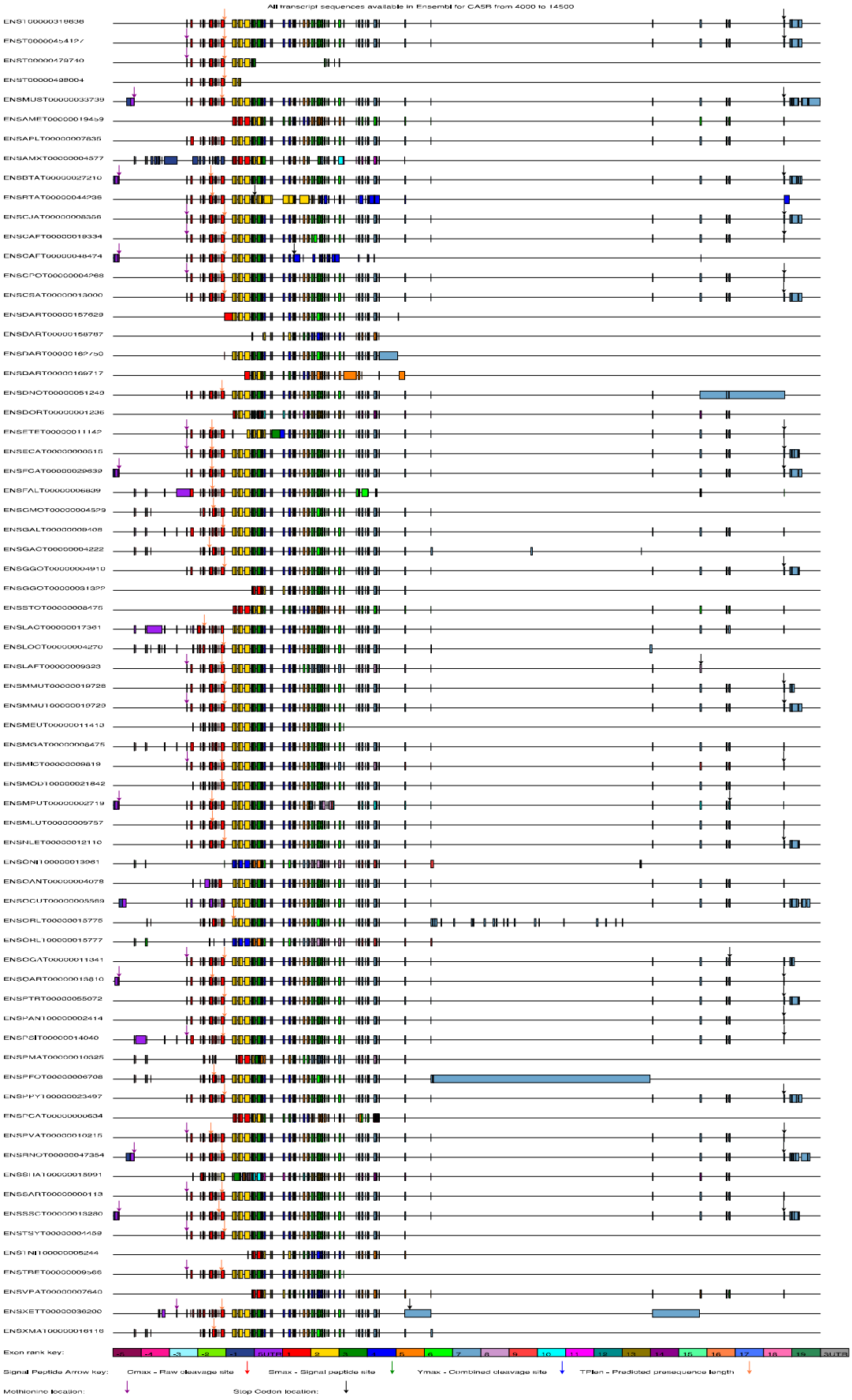

Figure 37 : Exon MSA schematic of CA5B of all the transcripts in Ensembl. The PRANK MSA is zoomed in from position 4000 to 14500.

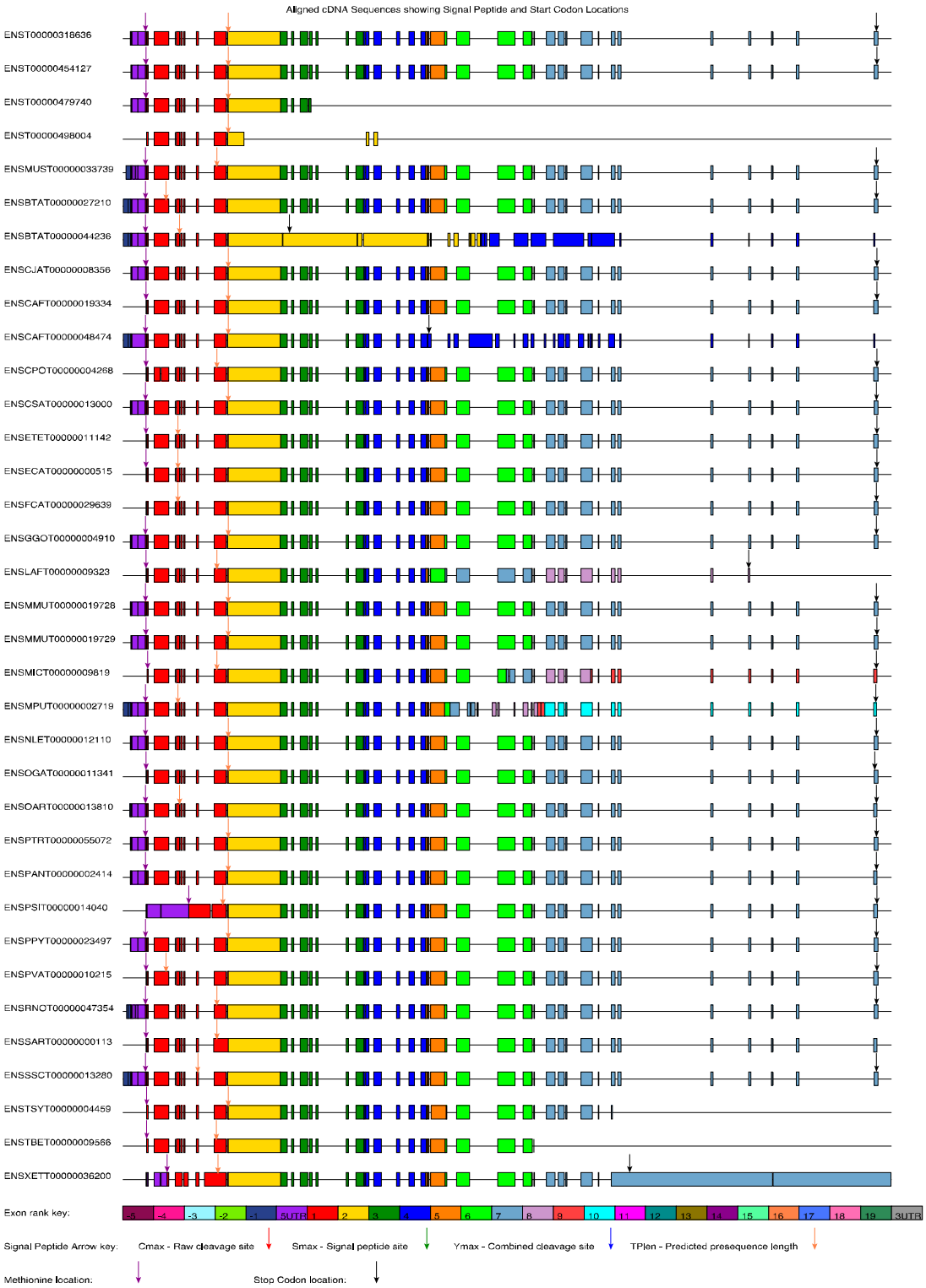

Figure 38 : Exon schematic of CA5B where all the transcripts have a start codon and a predicted mitochondrial targeting peptide. Zoomed in from position 1900 to position 4800 in the PRANK MSA.

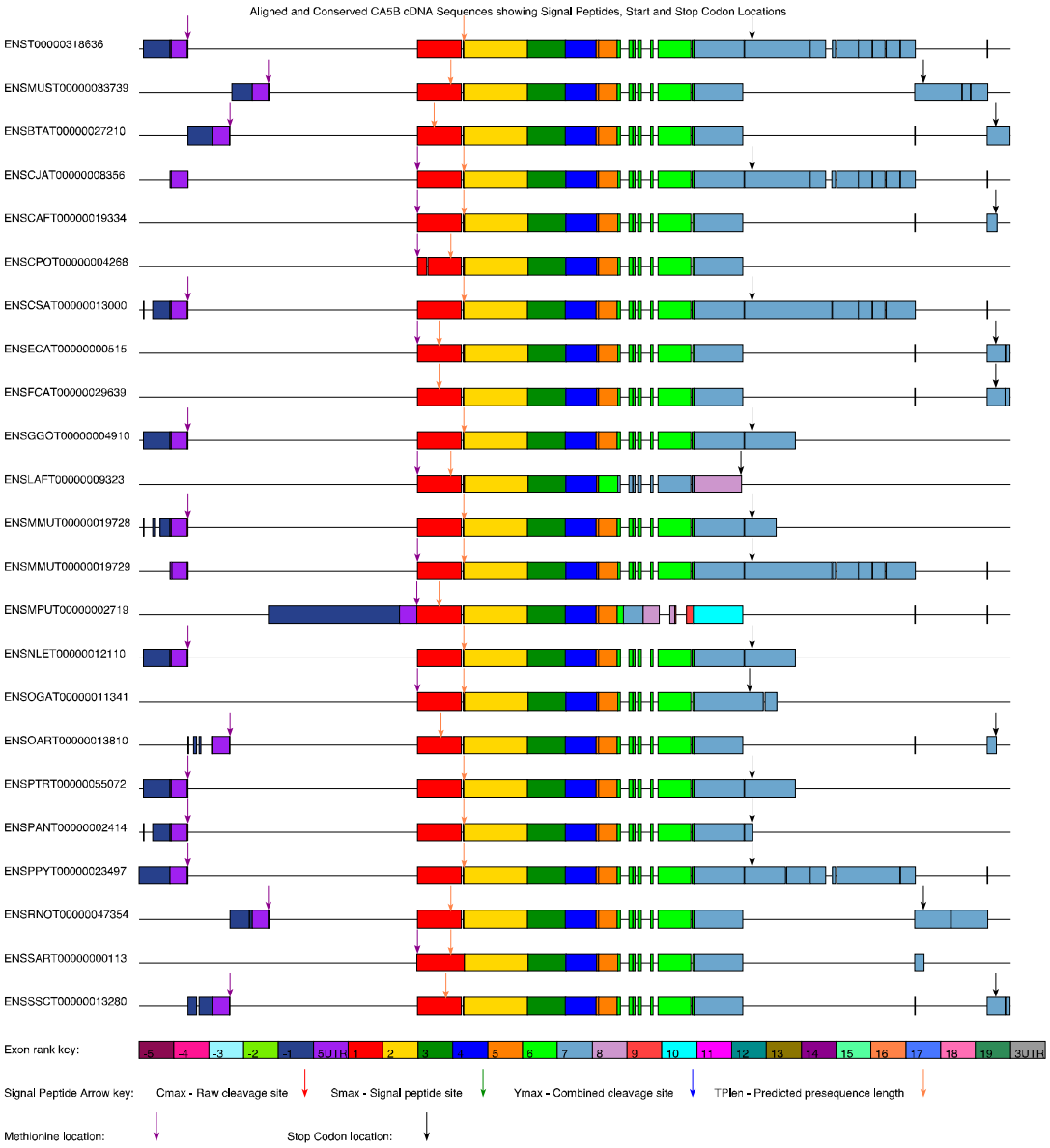

Figure 39 : Exon MSA schematic for conserved transcripts for CA5B. The PRANK MSA was zoomed in from position 1100 to 3800.

4.4 The membrane associated CAs - CA4, CA9, CA12, CA14 and CA15

SignalP has predicted that most of the membrane associated signal peptides would lie within exon 1 or exon 2, as is shown in Table 19. (Emanuelsson, Brunak, von Heijne, & Nielsen, 2007) The detailed predicted exon locations of the signal peptides are shown in Table 22 to Table 26. The various exon MSA schematics of the membrane associated CAs is in Figure 40 to Figure 54. In addition Table 20 and Table 21 the transcripts with short exons (at most 11 residues long).

Table 19 : Signal peptide statistics for the membrane associated CAs.

CA	Number of transcripts in isoform	Exon containing cleavage site	Percentage
CA4	11	1	16%
	48	2	72%
	5	3	8%
	1	4	2%
	1	5	2%
CA9	23		34%
	43	1	64%
	1	2	1%
	1	3	1%
CA12	30		39%
	38	1	50%
	8	2	11%
CA14	25		34%
	13	1	18%
	33	2	45%
	2	3	3%
CA15	17		42%
	12	1	29%
	12	2	29%

Table 20 : A summary of the number of transcripts with exons that are at most 11 residues long. The last three columns (reading from left to right) refer to the number of transcripts with the 2nd last exon that is short, the number of transcripts where at least one of the last three exons is short and finally the number of transcripts where at least one of the 3rd, 4th or 5th exons are short.

	Count of Short exon transcripts	Protein Coding transcripts	% containing Short Exon	Last Exon Short	2nd last exon	At least one of last 3 exon(s) short	Early exon(s) short
Membr-ane Transcripts	227	375	61%	15	19	152	35
CA4	35	117	30%	5	5	17	12
CA9	57	68	84%	4	1	54	10
CA12	56	76	74%	1	6	21	2
CA14	65	73	89%	3	7	57	9
CA15	14	41	34%	2	0	3	2

Table 21 : A list of the short exons of at most 11 residues long found in the membrane-associated group of CAs.

CA	EnsTranscriptId	Short Exon Num	Last Exon Num	Bases	Species	Start Codon Loctn	SignalP	Loc	RC
CA4	ENST00000587265	1	2	20	Homo sapiens		N	_	2
CA4	ENSMUST00000150596	6	6	29	Mus musculus		N	M	4
CA4	ENSAMET00000008981	7	8	30	Ailuropoda melanoleuca		N	_	1
CA4	ENSAMXT00000006159	4	11	22	Astyanax mexicanus		Y	S	1
CA4	ENSAMXT00000006159	6	11	17	Astyanax mexicanus		Y	S	1
CA4	ENSAMXT00000006159	7	11	21	Astyanax mexicanus		Y	S	1
CA4	ENSCJAT00000001999	2	8	12	Callithrix jacchus		N	_	1
CA4	ENSCJAT00000001999	6	8	11	Callithrix jacchus		N	_	1
CA4	ENSETET00000017555	1	6	19	Echinops telfairi		N	_	2
CA4	ENSEEUT00000013433	3	9	3	Erinaceus europaeus		N	_	1
CA4	ENSEEUT00000013433	4	9	3	Erinaceus europaeus		N	_	1
CA4	ENSEEUT00000013433	5	9	9	Erinaceus europaeus		N	_	1
CA4	ENSGMOT00000006568	6	8	4	Gadus morhua	9	N	M	5
CA4	ENSGMOT00000006568	7	8	8	Gadus morhua	9	N	M	5
CA4	ENSGACT00000014885	1	11	23	Gasterosteus aculeatus	3	Y	S	2
CA4	ENSGACT00000014885	2	11	13	Gasterosteus aculeatus	3	Y	S	2
CA4	ENSGACT00000014885	3	11	13	Gasterosteus aculeatus	3	Y	S	2
CA4	ENSGACT00000014885	4	11	20	Gasterosteus aculeatus	3	Y	S	2
CA4	ENSGACT00000014885	11	11	31	Gasterosteus aculeatus	3	Y	S	2
CA4	ENSLOCT00000003670	2	9	11	Lepisosteus oculatus		N	S	1
CA4	ENSLOCT00000006442	2	10	27	Lepisosteus oculatus		Y	S	3
CA4	ENSLOCT00000006442	7	10	28	Lepisosteus oculatus		Y	S	3
CA4	ENSMGAT00000008696	1	8	29	Meleagris gallopavo		N	S	5
CA4	ENSNLET00000000432	2	9	32	Nomascus leucogenys	99	Y	S	2
CA4	ENSPMAT00000005932	7	9	22	Petromyzon marinus	0	N	_	2
CA4	ENSPFOT00000025039	2	9	9	Poecilia formosa		N	_	3
CA4	ENSPCAT00000005106	10	13	9	Procavia capensis	0	Y	S	1
CA4	ENSPCAT00000005106	11	13	18	Procavia capensis	0	Y	S	1
CA4	ENSPCAT00000005106	12	13	6	Procavia capensis	0	Y	S	1
CA4	ENSPCAT00000005106	13	13	9	Procavia capensis	0	Y	S	1
CA4	ENSSHAT00000003693	1	9	22	Sarcophilus harrisii		N	_	2
CA4	ENSTRUT00000015657	1	9	20	Takifugu rubripes		Y	S	1
CA4	ENSTRUT00000015658	2	9	20	Takifugu rubripes	6	Y	S	1
CA4	ENSTRUT00000015658	5	9	15	Takifugu rubripes	6	Y	S	1
CA4	ENSTRUT00000015660	5	6	32	Takifugu rubripes		N	_	1
CA4	ENSTRUT00000017798	2	9	25	Takifugu rubripes		Y	S	1

CA	EnsTranscriptId	Short Exon Num	Last Exon Num	Bases	Species	Start Codon Loctn	SignalP	Loc	RC
CA4	ENSTRUT00000017800	3	9	3	Takifugu rubripes	0	Y	S	1
CA4	ENSTRUT00000017948	2	7	30	Takifugu rubripes		N	_	1
CA4	ENSTRUT00000017949	1	8	28	Takifugu rubripes		N	_	1
CA4	ENSTRUT00000032459	2	9	6	Takifugu rubripes		Y	S	1
CA4	ENSTRUT00000032460	2	7	24	Takifugu rubripes		Y	S	1
CA4	ENSTRUT00000032462	1	8	30	Takifugu rubripes		Y	S	3
CA4	ENSTRUT00000032463	1	8	21	Takifugu rubripes		Y	M	5
CA4	ENSTRUT00000032463	2	8	28	Takifugu rubripes		Y	M	5
CA4	ENSTNIT00000011580	2	9	3	Tetraodon nigroviridis	0	Y	S	1
CA4	ENSTBET00000014115	4	10	23	Tupaia belangeri	0	Y	S	2
CA4	ENSTBET00000014115	10	10	3	Tupaia belangeri	0	Y	S	2
CA4	ENSTTRT00000008779	3	11	25	Tursiops truncatus		N	_	2
CA4	ENSTTRT00000008779	9	11	6	Tursiops truncatus		N	_	2
CA4	ENSXETT00000029575	2	9	26	Xenopus tropicalis		N	_	3
CA4	ENSXETT00000029575	3	9	20	Xenopus tropicalis		N	_	3
CA4	ENSXMAT00000005597	2	9	30	Xiphophorus maculatus	12	N	S	4
CA4	ENSXMAT00000007891	10	10	28	Xiphophorus maculatus		Y	S	1
CA4	ENSXMAT00000012003	2	11	21	Xiphophorus maculatus		Y	S	1
CA4	ENSXMAT00000012003	3	11	12	Xiphophorus maculatus		Y	S	1
CA9	ENST00000378357	2	11	30	Homo sapiens	104	Y	S	1
CA9	ENST00000378357	9	11	27	Homo sapiens	104	Y	S	1
CA9	ENST00000493245	5	7	27	Homo sapiens				
CA9	ENST00000617161	2	9	30	Homo sapiens	42	Y	S	1
CA9	ENSMUST00000030183	2	11	30	Mus musculus	90	Y	S	2
CA9	ENSMUST00000030183	9	11	27	Mus musculus	90	Y	S	2
CA9	ENSMUST00000124114	2	9	30	Mus musculus				
CA9	ENSMUST00000124114	7	9	27	Mus musculus				
CA9	ENSMUST00000128232	2	4	30	Mus musculus				
CA9	ENSMUST00000138073	2	9	30	Mus musculus		N	_	1
CA9	ENSMUST00000138073	7	9	27	Mus musculus		N	_	1
CA9	ENSMUST00000138073	9	9	15	Mus musculus		N	_	1
CA9	ENSAMET00000010466	2	11	30	Ailuropoda melanoleuca	0	Y	S	2
CA9	ENSAMET00000010466	9	11	27	Ailuropoda melanoleuca	0	Y	S	2
CA9	ENSACAT00000009548	11	14	24	Anolis carolinensis		N	_	2
CA9	ENSAMXT00000010898	1	12	6	Astyanax mexicanus	0	N	_	1
CA9	ENSAMXT00000010898	12	12	9	Astyanax mexicanus	0	N	_	1
CA9	ENSBTAT00000015174	2	11	30	Bos taurus	0	Y	S	2
CA9	ENSBTAT00000015174	9	11	27	Bos taurus	0	Y	S	2
CA9	ENSCJAT00000018636	2	13	3	Callithrix jacchus	104	Y	S	1
CA9	ENSCJAT00000018636	4	13	30	Callithrix jacchus	104	Y	S	1
CA9	ENSCJAT00000018636	11	13	27	Callithrix jacchus	104	Y	S	1
CA9	ENSCJAT00000059925	2	11	3	Callithrix jacchus	42	Y	S	1
CA9	ENSCJAT00000059925	4	11	30	Callithrix jacchus	42	Y	S	1
CA9	ENSCAFT00000003289	2	11	30	Canis familiaris	0	Y	S	1
CA9	ENSCAFT00000003289	9	11	27	Canis familiaris	0	Y	S	1
CA9	ENSCAFT00000048469	2	9	30	Canis familiaris	0	Y	S	1
CA9	ENSCPOT00000000403	2	11	30	Cavia porcellus	51	N	S	5
CA9	ENSCPOT00000000403	9	11	27	Cavia porcellus	51	N	S	5
CA9	ENSCSAT00000008486	2	11	30	Chlorocebus sabaeus	0	Y	S	1

CA	EnsTranscriptId	Short Exon Num	Last Exon Num	Bases	Species	Start Codon Loctn	SignalP	Loc	RC
CA9	ENSCSAT00000008486	9	11	27	Chlorocebus sabaeus	0	Y	S	1
CA9	ENSCSAVT00000011041	9	9	21	Ciona savignyi		Y	S	1
CA9	ENSDNOT00000019791	2	11	30	Dasypus novemcinctus	657	Y	S	2
CA9	ENSDNOT00000019791	9	11	27	Dasypus novemcinctus	657	Y	S	2
CA9	ENSDORT00000014587	2	11	30	Dipodomys ordii	0	Y	S	1
CA9	ENSDORT00000014587	9	11	27	Dipodomys ordii	0	Y	S	1
CA9	ENSETET00000019543	1	18	2	Echinops telfairi	0	N	S	5
CA9	ENSETET00000019543	4	18	14	Echinops telfairi	0	N	S	5
CA9	ENSETET00000019543	7	18	21	Echinops telfairi	0	N	S	5
CA9	ENSETET00000019543	9	18	30	Echinops telfairi	0	N	S	5
CA9	ENSETET00000019543	16	18	27	Echinops telfairi	0	N	S	5
CA9	ENSECAT00000010539	2	11	30	Equus caballus	0	Y	S	2
CA9	ENSECAT00000010539	9	11	27	Equus caballus	0	Y	S	2
CA9	ENSECAT00000010559	2	9	30	Equus caballus	0	Y	S	2
CA9	ENSEEUT00000004896	2	14	27	Erinaceus europaeus	0	Y	S	1
CA9	ENSEEUT00000004896	4	14	6	Erinaceus europaeus	0	Y	S	1
CA9	ENSEEUT00000004896	5	14	3	Erinaceus europaeus	0	Y	S	1
CA9	ENSEEUT00000004896	12	14	27	Erinaceus europaeus	0	Y	S	1
CA9	ENSFCAT00000012747	2	11	30	Felis catus	0	Y	S	2
CA9	ENSFCAT00000012747	9	11	27	Felis catus	0	Y	S	2
CA9	ENSFALT00000002275	2	11	24	Ficedula albicollis		Y	S	2
CA9	ENSFALT00000002275	9	11	27	Ficedula albicollis		Y	S	2
CA9	ENSGALT00000034416	2	12	30	Gallus gallus		Y	S	3
CA9	ENSGALT00000034416	9	12	24	Gallus gallus		Y	S	3
CA9	ENSGALT00000034416	12	12	17	Gallus gallus		Y	S	3
CA9	ENSSTOT00000005042	2	11	30	Ictidomys tridecemlineatus	0	Y	S	1
CA9	ENSSTOT00000005042	9	11	30	Ictidomys tridecemlineatus	0	Y	S	1
CA9	ENSLOCT00000009047	2	12	24	Lepisosteus oculatus		Y	S	1
CA9	ENSLOCT00000009047	10	12	8	Lepisosteus oculatus		Y	S	1
CA9	ENSLOCT00000009060	2	10	22	Lepisosteus oculatus		Y	S	1
CA9	ENSLAFT00000035588	2	11	30	Loxodonta africana	0	Y	S	2
CA9	ENSLAFT00000035588	9	11	27	Loxodonta africana	0	Y	S	2
CA9	ENSMMUT00000022919	2	10	30	Macaca mulatta	42	Y	S	1
CA9	ENSMMUT00000046419	2	11	30	Macaca mulatta	104	Y	S	1
CA9	ENSMMUT00000046419	9	11	27	Macaca mulatta	104	Y	S	1
CA9	ENSMGAT00000002910	7	9	24	Meleagris gallopavo		N	_	2
CA9	ENSMICT00000001058	2	7	22	Microcebus murinus	0	Y	S	1
CA9	ENSMICT00000001058	3	7	8	Microcebus murinus	0	Y	S	1
CA9	ENSMODT00000009473	2	11	30	Monodelphis domestica	0	Y	S	1
CA9	ENSMODT00000009473	9	11	27	Monodelphis domestica	0	Y	S	1

CA	EnsTranscriptId	Short Exon Num	Last Exon Num	Bases	Species	Start Codon Loctn	SignalP	Loc	RC
CA9	ENSMPUT00000005310	9	11	27	Mustela putorius furo	0	Y	S	2
CA9	ENSMLUT00000011457	2	11	30	Myotis lucifugus	0	Y	S	3
CA9	ENSMLUT00000011457	9	11	27	Myotis lucifugus	0	Y	S	3
CA9	ENSNLET00000005534	2	11	30	Nomascus leucogenys	0	Y	S	2
CA9	ENSNLET00000005534	9	11	27	Nomascus leucogenys	0	Y	S	2
CA9	ENSOPRT00000012743	2	15	10	Ochotona princeps	0	Y	S	2
CA9	ENSOPRT00000012743	4	15	15	Ochotona princeps	0	Y	S	2
CA9	ENSOPRT00000012743	6	15	30	Ochotona princeps	0	Y	S	2
CA9	ENSOPRT00000012743	13	15	27	Ochotona princeps	0	Y	S	2
CA9	ENSOCUT00000010190	2	11	30	Oryctolagus cuniculus	0	Y	S	2
CA9	ENSOCUT00000010190	9	11	27	Oryctolagus cuniculus	0	Y	S	2
CA9	ENSOGAT00000011794	2	11	30	Otolemur garnettii	0	Y	S	1
CA9	ENSOGAT00000011794	9	11	27	Otolemur garnettii	0	Y	S	1
CA9	ENSOART00000012948	2	11	30	Ovis aries	51	N	M	5
CA9	ENSOART00000012948	9	11	27	Ovis aries	51	N	M	5
CA9	ENSPTRT00000038697	2	11	30	Pan troglodytes	104	Y	S	1
CA9	ENSPTRT00000038697	9	11	27	Pan troglodytes	104	Y	S	1
CA9	ENSPANT00000026695	2	11	30	Papio anubis	246	Y	S	1
CA9	ENSPANT00000026695	9	11	27	Papio anubis	246	Y	S	1
CA9	ENSPPYT00000022255	2	11	30	Pongo abelii	124	Y	S	2
CA9	ENSPPYT00000022255	9	11	27	Pongo abelii	124	Y	S	2
CA9	ENSPCAT00000013359	2	12	30	Procavia capensis	0	Y	S	1
CA9	ENSPCAT00000013359	10	12	27	Procavia capensis	0	Y	S	1
CA9	ENSPVAT00000002086	2	12	30	Pteropus vampyrus	0	Y	S	1
CA9	ENSPVAT00000002086	10	12	27	Pteropus vampyrus	0	Y	S	1
CA9	ENSRNOT00000023060	2	11	30	Rattus norvegicus	49	Y	S	1
CA9	ENSRNOT00000023060	9	11	27	Rattus norvegicus	49	Y	S	1
CA9	ENSSHAT00000017258	5	11	26	Sarcophilus harrisii	56	N	M	4
CA9	ENSSHAT00000017259	4	12	32	Sarcophilus harrisii	0	Y	S	2
CA9	ENSSHAT00000017259	10	12	27	Sarcophilus harrisii	0	Y	S	2
CA9	ENSSART00000009228	5	10	18	Sorex araneus		N	M	5
CA9	ENSSART00000009228	8	10	27	Sorex araneus		N	M	5
CA9	ENSSSCT00000005853	2	11	30	Sus scrofa	0	Y	S	1
CA9	ENSSSCT00000005853	9	11	27	Sus scrofa	0	Y	S	1
CA9	ENSTGUT00000001956	7	9	24	Taeniopygia guttata		N	_	2
CA9	ENSTSYT00000004272	2	15	30	Tarsius syrichta	0	Y	S	2
CA9	ENSTSYT00000004272	5	15	27	Tarsius syrichta	0	Y	S	2
CA9	ENSTSYT00000004272	7	15	6	Tarsius syrichta	0	Y	S	2
CA9	ENSTSYT00000004272	8	15	15	Tarsius syrichta	0	Y	S	2
CA9	ENSTSYT00000004272	13	15	27	Tarsius syrichta	0	Y	S	2
CA9	ENSTBET00000008828	2	12	30	Tupaia belangeri	0	Y	S	2
CA9	ENSTBET00000008828	9	12	1	Tupaia belangeri	0	Y	S	2

CA	EnsTranscriptId	Short Exon Num	Last Exon Num	Bases	Species	Start Codon Loctn	SignalP	Loc	RC
CA9	ENSTBET00000008828	10	12	26	Tupaia belangeri	0	Y	S	2
CA9	ENSTTRT00000011923	2	14	30	Tursiops truncatus	0	Y	S	1
CA9	ENSTTRT00000011923	9	14	26	Tursiops truncatus	0	Y	S	1
CA9	ENSTTRT00000011923	10	14	3	Tursiops truncatus	0	Y	S	1
CA9	ENSTTRT00000011923	11	14	1	Tursiops truncatus	0	Y	S	1
CA9	ENSTTRT00000011923	12	14	2	Tursiops truncatus	0	Y	S	1
CA9	ENSTTRT00000011923	13	14	29	Tursiops truncatus	0	Y	S	1
CA12	ENST00000178638	2	11	21	Homo sapiens	441	Y	S	2
CA12	ENST00000344366	2	10	21	Homo sapiens	115	Y	S	2
CA12	ENST00000422263	2	9	21	Homo sapiens	126	Y	S	2
CA12	ENSMUST00000071889	2	11	21	Mus musculus	163	Y	S	3
CA12	ENSMUST00000071889	9	11	30	Mus musculus	163	Y	S	3
CA12	ENSMUST00000085420	2	10	21	Mus musculus	119	Y	S	3
CA12	ENSMUST00000123195	2	4	21	Mus musculus				
CA12	ENSMUST00000134829	1	8	12	Mus musculus		N	_	2
CA12	ENSMUST00000134829	6	8	30	Mus musculus		N	_	2
CA12	ENSMUST00000134829	8	8	30	Mus musculus		N	_	2
CA12	ENSMUST00000152011	2	8	21	Mus musculus				
CA12	ENSAMET00000002552	2	11	21	Ailuropoda melanoleuca		Y	S	3
CA12	ENSAPLT00000003569	2	10	29	Anas platyrhynchos	12	Y	S	2
CA12	ENSACAT00000012115	2	9	21	Anolis carolinensis	0	Y	S	1
CA12	ENSAMXT00000018445	2	11	21	Astyanax mexicanus	0	Y	S	1
CA12	ENSAMXT00000018445	9	11	24	Astyanax mexicanus	0	Y	S	1
CA12	ENSCAFT00000026900	2	8	21	Canis familiaris	501	N	M	5
CA12	ENSCSAT00000010432	2	11	21	Chlorocebus sabaeus	130	Y	S	2
CA12	ENSCHOT00000013519	1	10	19	Choloepus hoffmanni		N	_	1
CA12	ENSDART00000157235	6	9	30	Danio rerio		N	M	5
CA12	ENSDART00000157235	7	9	24	Danio rerio		N	M	5
CA12	ENSDNOT00000031582	2	8	21	Dasypus novemcinctus	0	Y	S	1
CA12	ENSETET00000004699	2	11	21	Echinops telfairi	0	Y	S	1
CA12	ENSECAT00000018420	2	11	21	Equus caballus	9	Y	S	1
CA12	ENSGALT00000039628	2	10	21	Gallus gallus		Y	S	2
CA12	ENSGACT00000006096	3	8	6	Gasterosteus aculeatus		Y	S	1
CA12	ENSGGOT00000002107	2	11	21	Gorilla gorilla	156	Y	S	2
CA12	ENSSTOT00000016054	-1	9	22	Ictidomys tridecemlineatus		N	_	2
CA12	ENSSTOT00000016054	7	9	19	Ictidomys tridecemlineatus		N	_	2
CA12	ENSLOCT00000018359	2	10	21	Lepisosteus oculatus		Y	S	2
CA12	ENSLOCT00000018361	2	10	21	Lepisosteus oculatus		Y	S	2
CA12	ENSMMUT00000030038	2	11	21	Macaca mulatta	130	Y	S	1
CA12	ENSMMUT00000030040	2	10	21	Macaca mulatta	115	Y	S	1
CA12	ENSMEUT00000003974	11	12	18	Macropus eugenii		N	_	3
CA12	ENSMGAT00000004565	2	11	21	Meleagris gallopavo	0	N	S	2
CA12	ENSMGAT00000004565	9	11	17	Meleagris gallopavo	0	N	S	2
CA12	ENSMICT00000007123	2	12	21	Microcebus murinus	0	Y	S	2

CA	EnsTranscriptId	Short Exon Num	Last Exon Num	Bases	Species	Start Codon Loctn	SignalP	Loc	RC
CA12	ENSMICT00000007123	11	12	1	Microcebus murinus	0	Y	S	2
CA12	ENSMODT00000014066	2	11	21	Monodelphis domestica	398	Y	S	1
CA12	ENSMPUT00000000641	2	11	21	Mustela putorius furo		Y	S	1
CA12	ENSNLET00000015645	2	11	21	Nomascus leucogenys	130	Y	S	2
CA12	ENSOPRT00000001031	2	10	21	Ochotona princeps	0	Y	S	1
CA12	ENSONIT00000019345	2	10	21	Oreochromis niloticus		Y	S	2
CA12	ENSONIT00000019345	3	10	12	Oreochromis niloticus		Y	S	2
CA12	ENSOANT00000017667	2	19	21	Ornithorhynchus anatinus	366	Y	S	1
CA12	ENSOANT00000017667	8	19	19	Ornithorhynchus anatinus	366	Y	S	1
CA12	ENSOANT00000017667	9	19	20	Ornithorhynchus anatinus	366	Y	S	1
CA12	ENSOANT00000017667	11	19	18	Ornithorhynchus anatinus	366	Y	S	1
CA12	ENSOANT00000017667	14	19	25	Ornithorhynchus anatinus	366	Y	S	1
CA12	ENSOANT00000017667	16	19	7	Ornithorhynchus anatinus	366	Y	S	1
CA12	ENSOANT00000017667	17	19	18	Ornithorhynchus anatinus	366	Y	S	1
CA12	ENSOCUT00000004634	2	11	21	Oryctolagus cuniculus		Y	S	2
CA12	ENSOART00000022644	2	12	21	Ovis aries	0	Y	S	1
CA12	ENSOART00000022644	9	12	24	Ovis aries	0	Y	S	1
CA12	ENSOART00000022644	10	12	18	Ovis aries	0	Y	S	1
CA12	ENSOART00000022645	2	11	21	Ovis aries	0	Y	S	1
CA12	ENSPTRT00000045069	2	11	21	Pan troglodytes	156	Y	S	2
CA12	ENSPANT00000025834	2	11	21	Papio anubis	0	Y	S	1
CA12	ENSPPYT00000007723	2	11	21	Pongo abelii	354	Y	S	2
CA12	ENSPCAT00000005325	2	12	21	Procavia capensis	0	Y	S	1
CA12	ENSPVAT00000014854	2	11	21	Pteropus vampyrus	0	Y	S	1
CA12	ENSRNOT00000023952	2	11	21	Rattus norvegicus	175	Y	S	1
CA12	ENSRNOT00000023952	9	11	30	Rattus norvegicus	175	Y	S	1
CA12	ENSSHAT00000020799	2	13	21	Sarcophilus harrisii	139	Y	S	1
CA12	ENSSHAT00000020799	6	13	15	Sarcophilus harrisii	139	Y	S	1
CA12	ENSSHAT00000020799	7	13	17	Sarcophilus harrisii	139	Y	S	1
CA12	ENSSHAT00000020799	8	13	32	Sarcophilus harrisii	139	Y	S	1
CA12	ENSSHAT00000020800	2	11	21	Sarcophilus harrisii	139	Y	S	1
CA12	ENSTRUT00000033691	8	9	10	Takifugu rubripes		Y	S	5
CA12	ENSTRUT00000033692	1	10	31	Takifugu rubripes		N	_	1
CA12	ENSTRUT00000033692	8	10	25	Takifugu rubripes		N	_	1
CA12	ENSTRUT00000033693	1	8	31	Takifugu rubripes		N	_	1
CA12	ENSTRUT00000033694	7	8	10	Takifugu rubripes		N	_	1
CA12	ENSTBET00000009348	2	11	21	Tupaia belangeri	0	Y	S	2
CA12	ENSTTRT00000001445	2	11	21	Tursiops truncatus	0	Y	S	1

CA	EnsTranscriptId	Short Exon Num	Last Exon Num	Bases	Species	Start Codon Loctn	SignalP	Loc	RC
CA12	ENSVPAT00000004603	9	10	22	Vicugna pacos		N	_	2
CA12	ENSXETT00000021422	2	10	21	Xenopus tropicalis	6	Y	S	1
CA12	ENSXMAT00000003404	2	10	31	Xiphophorus maculatus		Y	S	2
CA12	ENSXMAT00000003404	9	10	24	Xiphophorus maculatus		Y	S	2
CA14	ENST00000369111	2	11	21	Homo sapiens	970	Y	S	2
CA14	ENST00000369111	9	11	21	Homo sapiens	970	Y	S	2
CA14	ENST00000483993	2	9	21	Homo sapiens		Y	S	1
CA14	ENST00000607082	3	4	18	Homo sapiens		N	_	5
CA14	ENST00000607652	7	8	21	Homo sapiens				
CA14	ENSMUST00000036181	2	11	21	Mus musculus	317	Y	S	2
CA14	ENSMUST00000036181	9	11	21	Mus musculus	317	Y	S	2
CA14	ENSMUST00000147962	1	6	21	Mus musculus	381	N	_	1
CA14	ENSMUST00000147962	2	6	22	Mus musculus	381	N	_	1
CA14	ENSAMET00000001976	2	11	21	Ailuropoda melanoleuca	0	Y	S	1
CA14	ENSAMET00000001976	9	11	21	Ailuropoda melanoleuca	0	Y	S	1
CA14	ENSACAT00000008081	5	6	31	Anolis carolinensis		N	M	5
CA14	ENSBTAT00000028634	2	11	21	Bos taurus		Y	S	1
CA14	ENSBTAT00000028634	9	11	21	Bos taurus		Y	S	1
CA14	ENSCJAT00000018939	2	9	21	Callithrix jacchus	75	Y	S	2
CA14	ENSCJAT00000018948	2	10	21	Callithrix jacchus	358	Y	S	2
CA14	ENSCJAT00000018959	2	11	21	Callithrix jacchus	75	Y	S	2
CA14	ENSCJAT00000018959	9	11	21	Callithrix jacchus	75	Y	S	2
CA14	ENSCJAT00000061608	2	9	23	Callithrix jacchus	258	Y	S	1
CA14	ENSCAFT00000018682	2	11	24	Canis familiaris	0	Y	S	1
CA14	ENSCAFT00000018682	9	11	21	Canis familiaris	0	Y	S	1
CA14	ENSCAFT00000035998	2	10	24	Canis familiaris		N	M	2
CA14	ENSCPOT00000006143	2	11	21	Cavia porcellus	0	Y	S	2
CA14	ENSCPOT00000006143	9	11	21	Cavia porcellus	0	Y	S	2
CA14	ENSCSAT00000016199	2	11	21	Chlorocebus sabaeus	316	Y	S	2
CA14	ENSCSAT00000016199	9	11	21	Chlorocebus sabaeus	316	Y	S	2
CA14	ENSCHOT00000011384	1	10	19	Choloepus hoffmanni		N	_	2
CA14	ENSCHOT00000011384	8	10	21	Choloepus hoffmanni		N	_	2
CA14	ENSDART00000088276	2	11	21	Danio rerio		Y	S	2
CA14	ENSDART00000088276	8	11	24	Danio rerio		Y	S	2
CA14	ENSDART00000149574	2	12	21	Danio rerio		Y	S	2
CA14	ENSDART00000149574	9	12	24	Danio rerio		Y	S	2
CA14	ENSDNOT00000007178	2	12	21	Dasypus novemcinctus	0	Y	S	3
CA14	ENSDNOT00000007178	10	12	21	Dasypus novemcinctus	0	Y	S	3
CA14	ENSDORT00000011692	2	11	21	Dipodomys ordii	0	Y	S	2
CA14	ENSDORT00000011692	9	11	21	Dipodomys ordii	0	Y	S	2
CA14	ENSETET00000012943	2	8	21	Echinops telfairi	0	Y	S	1
CA14	ENSECAT00000003315	2	10	21	Equus caballus	91	Y	S	1
CA14	ENSEEUT00000013276	1	12	19	Erinaceus europaeus		N	_	1
CA14	ENSEEUT00000013276	2	12	32	Erinaceus europaeus		N	_	1
CA14	ENSEEUT00000013276	4	12	6	Erinaceus europaeus		N	_	1

CA	EnsTranscriptId	Short Exon Num	Last Exon Num	Bases	Species	Start Codon Loctn	SignalP	Loc	RC
CA14	ENSEEUT00000013276	10	12	21	Erinaceus europaeus		N	_	1
CA14	ENSFCAT00000013938	2	11	21	Felis catus	0	Y	S	1
CA14	ENSFCAT00000013938	9	11	21	Felis catus	0	Y	S	1
CA14	ENSGMOT00000009400	7	8	25	Gadus morhua		N	_	1
CA14	ENSGALT00000045610	1	10	13	Gallus gallus		N	_	2
CA14	ENSGALT00000045610	8	10	21	Gallus gallus		N	_	2
CA14	ENSGACT00000015647	2	8	24	Gasterosteus aculeatus		Y	S	1
CA14	ENSGGOT00000017174	2	11	21	Gorilla gorilla	357	Y	S	2
CA14	ENSGGOT00000017174	9	11	21	Gorilla gorilla	357	Y	S	2
CA14	ENSSTOT00000020015	2	10	21	Ictidomys tridecemlineatus	801	Y	S	2
CA14	ENSLACT00000010151	8	10	24	Latimeria chalumnae	0	N	S	1
CA14	ENSLOCT00000007788	2	11	21	Lepisosteus oculatus		Y	S	1
CA14	ENSLOCT00000007788	9	11	24	Lepisosteus oculatus		Y	S	1
CA14	ENSLAFT00000004700	2	11	21	Loxodonta africana	0	Y	S	2
CA14	ENSLAFT00000004700	9	11	21	Loxodonta africana	0	Y	S	2
CA14	ENSMMUT00000030102	2	11	21	Macaca mulatta	361	Y	S	2
CA14	ENSMMUT00000030102	9	11	21	Macaca mulatta	361	Y	S	2
CA14	ENSMEUT00000006822	1	11	19	Macropus eugenii		N	_	1
CA14	ENSMEUT00000006822	6	11	15	Macropus eugenii		N	_	1
CA14	ENSMEUT00000006822	9	11	21	Macropus eugenii		N	_	1
CA14	ENSMGAT00000000706	9	10	21	Meleagris gallopavo		N	M	3
CA14	ENSMICT00000015263	2	11	21	Microcebus murinus	0	Y	S	1
CA14	ENSMICT00000015263	9	11	21	Microcebus murinus	0	Y	S	1
CA14	ENSMODT00000023828	-1	11	8	Monodelphis domestica	31	Y	S	2
CA14	ENSMODT00000023828	2	11	21	Monodelphis domestica	31	Y	S	2
CA14	ENSMODT00000023828	9	11	21	Monodelphis domestica	31	Y	S	2
CA14	ENSMPUT00000005379	2	11	21	Mustela putorius furo	46	Y	S	1
CA14	ENSMPUT00000005379	9	11	21	Mustela putorius furo	46	Y	S	1
CA14	ENSMLUT00000013683	2	12	21	Myotis lucifugus	3	Y	S	1
CA14	ENSMLUT00000013683	3	12	14	Myotis lucifugus	3	Y	S	1
CA14	ENSMLUT00000013683	10	12	21	Myotis lucifugus	3	Y	S	1
CA14	ENSNLET00000012516	2	10	21	Nomascus leucogenys	345	Y	S	2
CA14	ENSOPRT00000010072	2	11	21	Ochotona princeps	0	Y	S	2
CA14	ENSOPRT00000010072	9	11	21	Ochotona princeps	0	Y	S	2
CA14	ENSONIT00000008059	2	13	24	Oreochromis niloticus		Y	S	1
CA14	ENSONIT00000008059	9	13	24	Oreochromis niloticus		Y	S	1
CA14	ENSOCUT00000009030	2	11	21	Oryctolagus cuniculus	0	Y	S	1

CA	EnsTranscriptId	Short Exon Num	Last Exon Num	Bases	Species	Start Codon Loctn	SignalP	Loc	RC
CA14	ENSOCUT00000009030	9	11	21	Oryctolagus cuniculus	0	Y	S	1
CA14	ENSORLT00000012895	7	10	24	Oryzias latipes		N	_	2
CA14	ENSOGAT00000004845	2	11	21	Otolemur garnettii	0	Y	S	1
CA14	ENSOGAT00000004845	9	11	21	Otolemur garnettii	0	Y	S	1
CA14	ENSOART00000022632	2	12	30	Ovis aries		N	M	3
CA14	ENSOART00000022632	3	12	21	Ovis aries		N	M	3
CA14	ENSOART00000022632	10	12	21	Ovis aries		N	M	3
CA14	ENSPTRT00000002314	2	11	21	Pan troglodytes	353	Y	S	2
CA14	ENSPTRT00000002314	9	11	21	Pan troglodytes	353	Y	S	2
CA14	ENSPANT00000025708	2	11	21	Papio anubis	246	Y	S	2
CA14	ENSPANT00000025708	9	11	21	Papio anubis	246	Y	S	2
CA14	ENSPFOT00000013838	2	10	18	Poecilia formosa		Y	S	5
CA14	ENSPFOT00000013838	3	10	24	Poecilia formosa		Y	S	5
CA14	ENSPPYT00000001094	2	11	21	Pongo abelii	297	Y	S	2
CA14	ENSPPYT00000001094	9	11	21	Pongo abelii	297	Y	S	2
CA14	ENSPCAT00000007272	2	11	21	Procavia capensis	0	Y	S	3
CA14	ENSPCAT00000007272	9	11	21	Procavia capensis	0	Y	S	3
CA14	ENSPCAT00000007272	11	11	31	Procavia capensis	0	Y	S	3
CA14	ENSPVAT00000013464	2	11	24	Pteropus vampyrus	0	Y	S	1
CA14	ENSPVAT00000013464	9	11	21	Pteropus vampyrus	0	Y	S	1
CA14	ENSRNOT00000025523	2	11	21	Rattus norvegicus	39	Y	S	2
CA14	ENSRNOT00000025523	9	11	21	Rattus norvegicus	39	Y	S	2
CA14	ENSSHAT00000010532	2	11	21	Sarcophilus harrisii	0	Y	S	2
CA14	ENSSHAT00000010532	9	11	21	Sarcophilus harrisii	0	Y	S	2
CA14	ENSSART00000011016	1	17	17	Sorex araneus	0	Y	S	2
CA14	ENSSART00000011016	3	17	21	Sorex araneus	0	Y	S	2
CA14	ENSSART00000011016	9	17	3	Sorex araneus	0	Y	S	2
CA14	ENSSART00000011016	10	17	6	Sorex araneus	0	Y	S	2
CA14	ENSSART00000011016	11	17	16	Sorex araneus	0	Y	S	2
CA14	ENSSART00000011016	12	17	26	Sorex araneus	0	Y	S	2
CA14	ENSSART00000011016	14	17	21	Sorex araneus	0	Y	S	2
CA14	ENSSART00000011016	15	17	5	Sorex araneus	0	Y	S	2
CA14	ENSSSCT00000028733	2	11	21	Sus scrofa		Y	S	1
CA14	ENSSSCT00000028733	9	11	21	Sus scrofa		Y	S	1
CA14	ENSTSYT00000012163	2	12	4	Tarsius syrichta	0	Y	S	2
CA14	ENSTSYT00000012163	3	12	21	Tarsius syrichta	0	Y	S	2
CA14	ENSTSYT00000012163	10	12	21	Tarsius syrichta	0	Y	S	2
CA14	ENSTBET00000002646	2	9	21	Tupaia belangeri	0	Y	S	1
CA14	ENSTBET00000002646	3	9	20	Tupaia belangeri	0	Y	S	1
CA14	ENSTBET00000002646	5	9	17	Tupaia belangeri	0	Y	S	1
CA14	ENSTTRT00000005149	2	11	21	Tursiops truncatus	0	Y	S	1
CA14	ENSTTRT00000005149	10	11	21	Tursiops truncatus	0	Y	S	1
CA14	ENSVPAT00000005420	1	13	19	Vicugna pacos		N	_	1
CA14	ENSVPAT00000005420	8	13	21	Vicugna pacos		N	_	1
CA14	ENSVPAT00000005420	10	13	4	Vicugna pacos		N	_	1
CA14	ENSVPAT00000005420	11	13	30	Vicugna pacos		N	_	1
CA14	ENSVPAT00000005420	12	13	12	Vicugna pacos		N	_	1
CA14	ENSVPAT00000005420	13	13	9	Vicugna pacos		N	_	1
CA14	ENSXETT00000035852	2	11	17	Xenopus tropicalis	0	Y	S	3
CA14	ENSXETT00000035852	9	11	18	Xenopus tropicalis	0	Y	S	3
CA14	ENSXETT00000035855	2	11	24	Xenopus tropicalis	0	Y	S	3

CA	EnsTranscriptId	Short Exon Num	Last Exon Num	Bases	Species	Start Codon Loctn	SignalP	Loc	RC
CA14	ENSXETT00000035855	9	11	18	Xenopus tropicalis	0	Y	S	3
CA14	ENSXMAT00000016389	1	13	26	Xiphophorus maculatus		N	S	5
CA14	ENSXMAT00000016389	3	13	8	Xiphophorus maculatus		N	S	5
CA14	ENSXMAT00000016389	10	13	24	Xiphophorus maculatus		N	S	5
CA14	ENSXMAT00000016389	13	13	21	Xiphophorus maculatus		N	S	5
CA15	ENSAMXT00000021275	2	6	27	Astyanax mexicanus	805	Y	S	1
CA15	ENSDART00000144919	2	8	27	Danio rerio	41	Y	S	1
CA15	ENSGACT00000022492	2	8	30	Gasterosteus aculeatus	125	Y	S	3
CA15	ENSLACT00000005015	1	8	28	Latimeria chalumnae		N	_	3
CA15	ENSLOCT00000005989	9	9	13	Lepisosteus oculatus		Y	S	1
CA15	ENSORLT00000008287	-1	8	26	Oryzias latipes	26	Y	S	1
CA15	ENSORLT00000008287	2	8	27	Oryzias latipes	26	Y	S	1
CA15	ENSORLT00000008296	2	11	30	Oryzias latipes		Y	S	1
CA15	ENSORLT00000008296	6	11	21	Oryzias latipes		Y	S	1
CA15	ENSPSIT00000006594	4	9	12	Pelodiscus sinensis	18	Y	S	1
CA15	ENSPFOT00000019918	2	8	30	Poecilia formosa		Y	S	1
CA15	ENSPPYT00000023245	5	11	12	Pongo abelii		N	_	2
CA15	ENSTRUT00000023733	6	6	27	Takifugu rubripes		N	_	4
CA15	ENSTNIT00000012140	2	4	30	Tetraodon nigroviridis		Y	_	4
CA15	ENSXETT00000048444	1	8	28	Xenopus tropicalis		N	_	1
CA15	ENSXMAT00000014162	1	8	29	Xiphophorus maculatus		N	_	5

Table 22 : SignalP and TargetP results for CA4.

TranscriptId	Cmax_Pos	Ymax_Pos	Smax_Pos	Tplen	Loc	RC
ENST00000300900	2	2	1	2	S	3
ENSMUST00000103194	2	2	1	2	S	1
ENSAPLT00000007809	1	1	1	1	S	2
ENSACAT00000015205	2	1	1	1	S	1
ENSAMXT00000004075	2	2	1	2	M	5
ENSAMXT00000006159	2	2	1	2	S	1
ENSAMXT00000016346	1	1	1	1	S	1
ENSBTAT00000023909	2	2	1	2	S	2
ENSCJAT00000002004	2	2	1	1	S	4
ENSCJAT00000021898	2	2	1	2	S	5
ENSCAFT00000028280	2	2	1	2	S	4
ENSCPOT00000020509	2	2	1	2	S	1
ENSCSAT00000004047	2	2	1	1	S	3
ENSDART00000064012	1	1	1	1	S	1
ENSDART00000065362	2	2	1	2	S	1
ENSDNOT00000042392	2	2	1	2	S	1
ENSECAT00000024391	2	2	1	2	S	1
ENSFCAT00000009074	2	2	1	2	S	2
ENSFALT00000005604	2	2	1	1	S	1
ENSGMOT00000015251	2	2	1	2	S	1
ENSGACT00000014885	4	4	5	4	S	2

TranscriptId	Cmax_Pos	Ymax_Pos	Smax_Pos	Tplen	Loc	RC
ENSGACT00000014889	2	2	1	1	S	2
ENSGACT00000026794	2	2	1	1	S	1
ENSGGOT00000011551	2	2	1	2	S	3
ENSSTOT00000010433	2	2	1	1	S	3
ENSLACT00000019812	2	2	1	1	S	1
ENSLOCT00000003662	2	2	1	2	S	2
ENSLOCT00000006442	2	2	1	2	S	3
ENSLAFT00000007673	2	2	1	2	S	3
ENSMMUT00000011085	2	2	1	1	S	2
ENSMICT00000013377	1	1	1	1	S	2
ENSNLET00000000432	2	2	1	2	S	2
ENSONIT00000010931	2	2	1	2	S	1
ENSONIT00000016074	1	1	1	1	S	1
ENSORLT00000014110	1	1	1	1	S	1
ENSOGAT00000003441	2	2	1	1	S	2
ENSPTRT00000017382	2	2	1	2	S	3
ENSPANT00000002527	2	2	1	1	S	2
ENSPSIT00000011410	2	2	1	1	S	2
ENSPPYT00000009945	2	2	1	2	S	2
ENSPCAT00000005106	2	2	1	2	S	1
ENSPVAT00000015002	2	2	1	2	S	2
ENSRNOT00000003908	2	1	1	2	S	1
ENSSHAT00000003692	2	2	1	2	S	1
ENSSSCT00000032087	2	2	1	2	S	2
ENSTGUT00000008316	2	2	1	1	S	1
ENSTRUT00000015656	3	3	2	2	S	3
ENSTRUT00000015657	3	3	2	2	S	1
ENSTRUT00000015658	2	2	1	1	S	1
ENSTRUT00000015659	1	1	1	1	S	1
ENSTRUT00000017798	3	3	1	2	S	1
ENSTRUT00000017799	2	2	1	1	S	1
ENSTRUT00000017800	2	2	1	1	S	1
ENSTRUT00000017944	2	2	1	2	S	2
ENSTRUT00000032459	3	3	1	3	S	1
ENSTRUT00000032460	2	2	1	3	S	1
ENSTRUT00000032461	1	1	1	1	S	2
ENSTRUT00000032462	2	2	2	2	S	3
ENSTRUT00000032463	3	3	2	5	M	5
ENSTNIT00000000097	1	1	1	1	S	4
ENSTNIT00000011580	2	2	1	1	S	1
ENSTNIT00000019455	2	2	1	2	S	1
ENSTNIT00000019557	2	2	1	1	S	1
ENSTBET00000014115	2	2	1	2	S	2
ENSXMAT00000007891	2	2	1	2	S	1
ENSXMAT00000012003	5	5	4	5	S	1

Table 23 : SignalP and TargetP results for CA9

TranscriptId	Cmax_Pos	Ymax_Pos	Smax_Pos	TPlen	Loc	RC
ENST00000378357	1	1	1	1	S	1
ENST00000617161	1	1	1	1	S	1
ENSMUST00000030183	1	1	1	1	S	2
ENSAMET00000010466	1	1	1	1	S	2
ENSBTAT00000015174	1	1	1	1	S	2
ENSCJAT00000018636	1	1	1	1	S	1
ENSCJAT00000059925	1	1	1	1	S	1
ENSCAFT00000003289	1	1	1	1	S	1
ENSCAFT00000048469	1	1	1	1	S	1
ENSCSAT00000008486	1	1	1	1	S	1
ENSDART00000164902	1	1	1	1	S	1
ENSDNOT00000019791	1	1	1	1	S	2
ENSDORT00000014587	1	1	1	1	S	1
ENSECAT00000010539	1	1	1	1	S	2
ENSECAT00000010559	1	1	1	1	S	2
ENSEEUT00000004896	1	1	1	1	S	1
ENSFCAT00000012747	1	1	1	1	S	2
ENSFALT00000002275	1	1	1	1	S	2
ENSGALT00000034416	1	1	1	1	S	3
ENSSTOT00000005042	1	1	1	1	S	1
ENSLACT00000026638	1	1	1	1	S	1
ENSLOCT00000009047	1	1	1	1	S	1
ENSLOCT00000009060	2	2	1	2	S	1
ENSLAFT00000035588	1	1	1	1	S	2
ENSMMUT00000022919	1	1	1	1	S	1
ENSMMUT00000046419	1	1	1	1	S	1
ENSMICT00000001058	1	1	1	1	S	1
ENSMODT00000009473	1	1	1	1	S	1
ENSMPUT00000005310	1	1	1	1	S	2
ENSMLUT00000011457	1	1	1	1	S	3
ENSNLET00000005534	1	1	1	1	S	2
ENSOPRT00000012743	3	3	3	3	S	2
ENSOCUT00000010190	1	1	1	1	S	2
ENSOGAT00000011794	1	1	1	1	S	1
ENSPTRT00000038697	1	1	1	1	S	1
ENSPANT00000026695	1	1	1	1	S	1
ENSPPYT00000022255	1	1	1	1	S	2
ENSPCAT00000013359	1	1	1	1	S	1
ENSPVAT00000002086	1	1	1	1	S	1
ENSRNOT00000023060	1	1	1	1	S	1
ENSSHAT00000017259	1	1	1	1	S	2
ENSSSCT00000005853	1	1	1	1	S	1
ENSTSYT00000004272	1	1	1	1	S	2
ENSTBET00000008828	1	1	1	1	S	2
ENSTTRT00000011923	1	1	1	1	S	1

Table 24 : SignalP and TargetP results for CA12

TranscriptId	Cmax_Pos	Ymax_Pos	Smax_Pos	TPlen	Loc	RC
ENST00000178638	1	1	1	1	S	2
ENST00000344366	1	1	1	1	S	2
ENST00000422263	1	1	1	1	S	2
ENSMUST00000071889	1	1	1	1	S	3
ENSMUST00000085420	1	1	1	1	S	3
ENSAMET00000002552	1	1	1	1	S	3
ENSAPLT00000003569	1	1	1	1	S	2
ENSACAT00000012115	1	1	1	1	S	1
ENSAMXT00000018445	2	2	1	2	S	1
ENSCSAT00000010432	1	1	1	1	S	2
ENSDNOT00000031582	1	1	1	1	S	1
ENSETET00000004699	1	1	1	1	S	1
ENSECAT00000018420	1	1	1	1	S	1
ENSGMOT00000009148	2	2	1	1	S	3
ENSGALT00000039628	1	1	1	1	S	2
ENSGACT00000006096	2	2	1	2	S	1
ENSGGOT00000002107	1	1	1	1	S	2
ENSLACT00000013411	2	2	1	2	S	1
ENSLOCT00000018359	2	2	1	1	S	2
ENSLOCT00000018361	2	2	1	1	S	2
ENSMMUT00000030038	1	1	1	1	S	1
ENSMMUT00000030040	1	1	1	1	S	1
ENSMICT00000007123	1	1	1	1	S	2
ENSMODT00000014066	1	1	1	1	S	1
ENSMPUT00000000641	1	1	1	1	S	1
ENSNLET00000015645	1	1	1	1	S	2
ENSOPRT00000001031	1	1	1	1	S	1
ENSONIT00000019345	2	2	1	2	S	2
ENSOANT00000017667	1	1	1	1	S	1
ENSOCUT00000004634	1	1	1	1	S	2
ENSOART00000022644	1	1	1	1	S	1
ENSOART00000022645	1	1	1	1	S	1
ENSPTRT00000045069	1	1	1	1	S	2
ENSPANT00000025834	1	1	1	1	S	1
ENSPPYT00000007723	1	1	1	1	S	2
ENSPCAT00000005325	1	1	1	1	S	1
ENSPVAT00000014854	1	1	1	1	S	1
ENSRNOT00000023952	1	1	1	1	S	1
ENSSHAT00000020799	1	1	1	1	S	1
ENSSHAT00000020800	1	1	1	1	S	1
ENSTRUT00000033690	1	1	1	1	S	3
ENSTRUT00000033691	1	1	1	1	S	5
ENSTBET00000009348	1	1	1	1	S	2
ENSTTRT00000001445	1	1	1	1	S	1
ENSXETT00000021422	1	1	1	1	S	1
ENSXMAT00000003404	2	2	1	2	S	2

Table 25 : SignalP and TargetP results for CA14

TranscriptId	Cmax_Pos	Ymax_Pos	Smax_Pos	TPlen	Loc	RC
ENST00000369111	2	2	1	2	S	2
ENSMUST00000036181	1	1	1	1	S	2
ENSAMET00000001976	2	2	1	2	S	1
ENSBTAT00000028634	2	2	1	2	S	1
ENSCJAT00000018939	2	2	1	2	S	2
ENSCJAT00000018948	2	2	1	2	S	2
ENSCJAT00000018959	2	2	1	2	S	2
ENSCJAT00000061608	2	2	1	2	S	1
ENSCAFT00000018679	1	1	1	1	S	1
ENSCAFT00000018682	2	2	1	2	S	1
ENSCPOT00000006143	1	1	1	1	S	2
ENSCSAT00000016199	2	2	1	2	S	2
ENSDART00000149574	1	1	1	1	S	2
ENSDNOT00000007178	2	2	1	2	S	3
ENSDORT00000011692	1	1	1	1	S	2
ENSETET00000012943	2	2	1	2	S	1
ENSECAT00000003315	2	2	1	2	S	1
ENSFCAT00000013938	2	2	1	2	S	1
ENSGACT00000015647	2	2	1	1	S	1
ENSGGOT00000017174	2	2	1	2	S	2
ENSSTOT00000020015	2	2	1	2	S	2
ENSLOCT00000007788	1	1	1	1	S	1
ENSLAFT00000004700	2	2	1	1	S	2
ENSMMUT00000030102	2	2	1	2	S	2
ENSMICT00000015263	2	2	1	2	S	1
ENSMODT00000023828	1	1	1	1	S	2
ENSMPUT00000005379	2	2	1	2	S	1
ENSMLUT00000013683	2	2	1	2	S	1
ENSNLET00000012516	2	2	1	2	S	2
ENSOPRT00000010072	2	2	1	2	S	2
ENSONIT00000008059	1	1	1	1	S	1
ENSOCUT00000009030	2	2	1	2	S	1
ENSOGAT00000004845	2	2	1	2	S	1
ENSPTRT00000002314	2	2	1	2	S	2
ENSPANT00000025708	2	2	1	2	S	2
ENSPFOT00000013838	3	3	1	2	S	5
ENSPPYT00000001094	2	2	1	2	S	2
ENSPCAT00000007272	1	1	1	1	S	3
ENSPVAT00000013464	2	2	1	2	S	1
ENSRNOT00000025523	1	1	1	1	S	2
ENSSHAT00000010532	1	1	1	1	S	2
ENSSART00000011016	2	2	2	2	S	2
ENSSSCT00000028733	1	1	1	1	S	1
ENSTSYT00000012163	3	3	1	3	S	2
ENSTBET00000002646	2	2	1	2	S	1
ENSTTRT00000005149	1	1	1	1	S	1
ENSXETT00000035852	2	2	1	1	S	3
ENSXETT00000035855	2	2	1	2	S	3

Table 26 : SignalP and TargetP results for CA15.

TranscriptId	Cmax_Pos	Ymax_Pos	Smax_Pos	TPlen	Loc	RC
ENSMUST00000118960	2	2	1	1	S	1
ENSACAT00000002736	2	2	1	2	S	1
ENSAMXT00000021275	1	1	1	1	S	1
ENSCAFT00000042949	1	1	1	1	S	1
ENSDART00000144919	1	1	1	1	S	1
ENSDNOT00000048563	1	1	1	1	S	1
ENSFCAT00000031379	1	1	1	1	S	2
ENSFALT00000008201	1	1	1	1	S	1
ENSGALT00000009489	2	2	1	2	S	1
ENSGACT00000022492	2	2	1	2	S	3
ENSGACT00000022494	1	1	1	1	S	3
ENSSTOT00000023103	1	1	1	1	S	1
ENSSTOT00000024755	2	2	1	2	S	2
ENSLOCT00000005971	2	2	1	1	S	1
ENSLOCT00000005989	2	1	1	1	S	1
ENSLAFT00000027652	1	1	1	1	S	1
ENSMGAT00000008102	2	2	1	2	S	1
ENSMODT00000012800	1	1	1	1	S	1
ENSORLT00000008287	1	1	8	1	S	1
ENSORLT00000008296	2	2	1	2	S	1
ENSPSIT00000006584	2	2	1	2	S	1
ENSPSIT00000006594	2	2	1	2	S	1
ENSPFOT00000019918	2	2	1	2	S	1
ENSRNOT00000000312	2	2	1	1	S	1

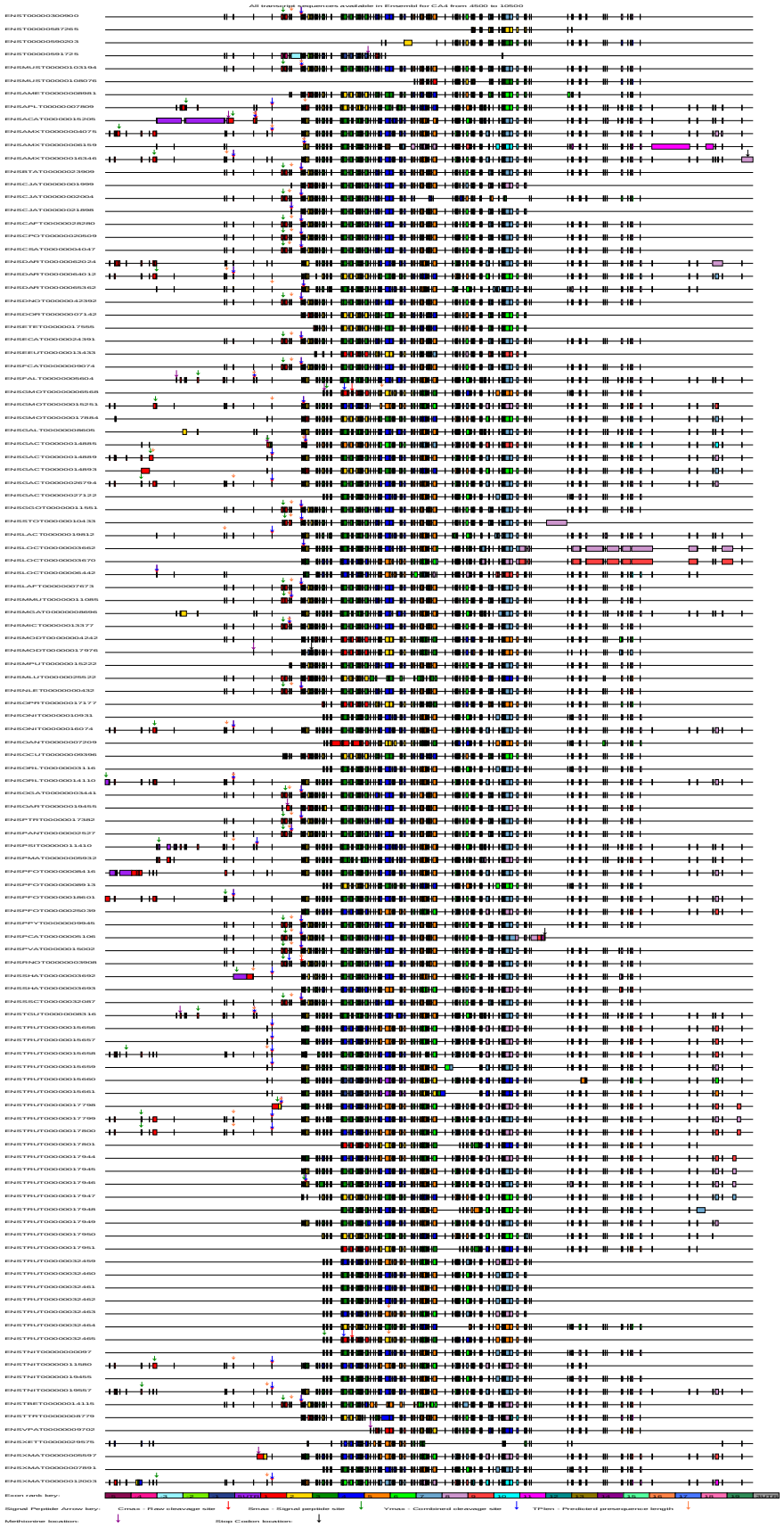

Figure 40 : Exon MSA schematic of all Ensembl CA4 transcripts zoomed in on the PRANK MSA from position 4500 to 10500. The variation here makes it very hard to read the entire exon MSA schematic. The signal peptides generally are located at the boundary of exon one and two. The catalytic domain is less conserved than other CAs, but only slightly. The public Dropbox folder (https://www.dropbox.com/l/sh/xYQqDdI4TGAi0TTfaJr9Hu) has many close up images.

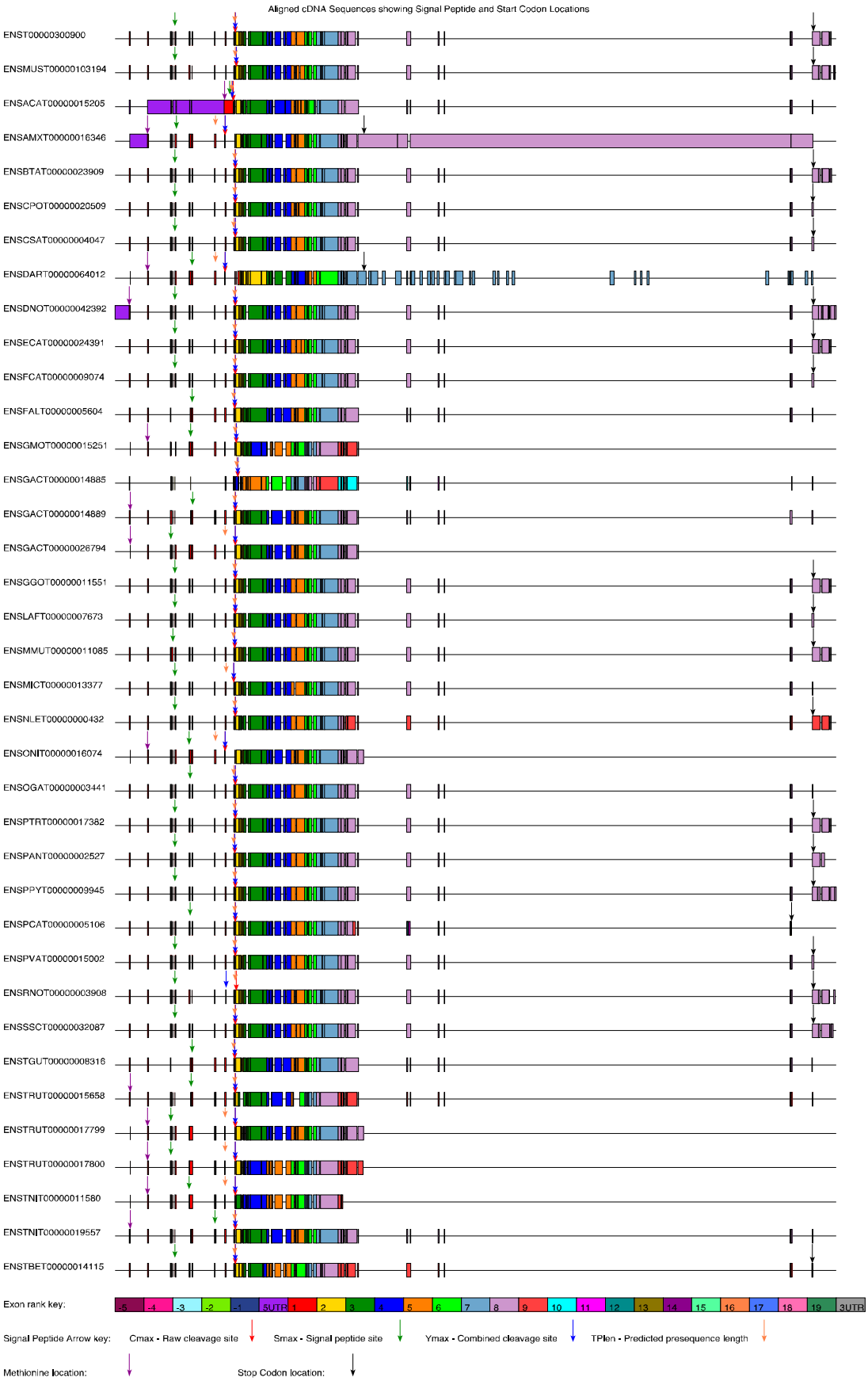

Figure 41 : Exon MSA schematic for CA4 where all the transcripts have a start codon and a signal peptide. Zoomed PRANK MSA from position 2050 to 7800 to include most start and stop codons and the locations of the predicted signal peptide indicators.

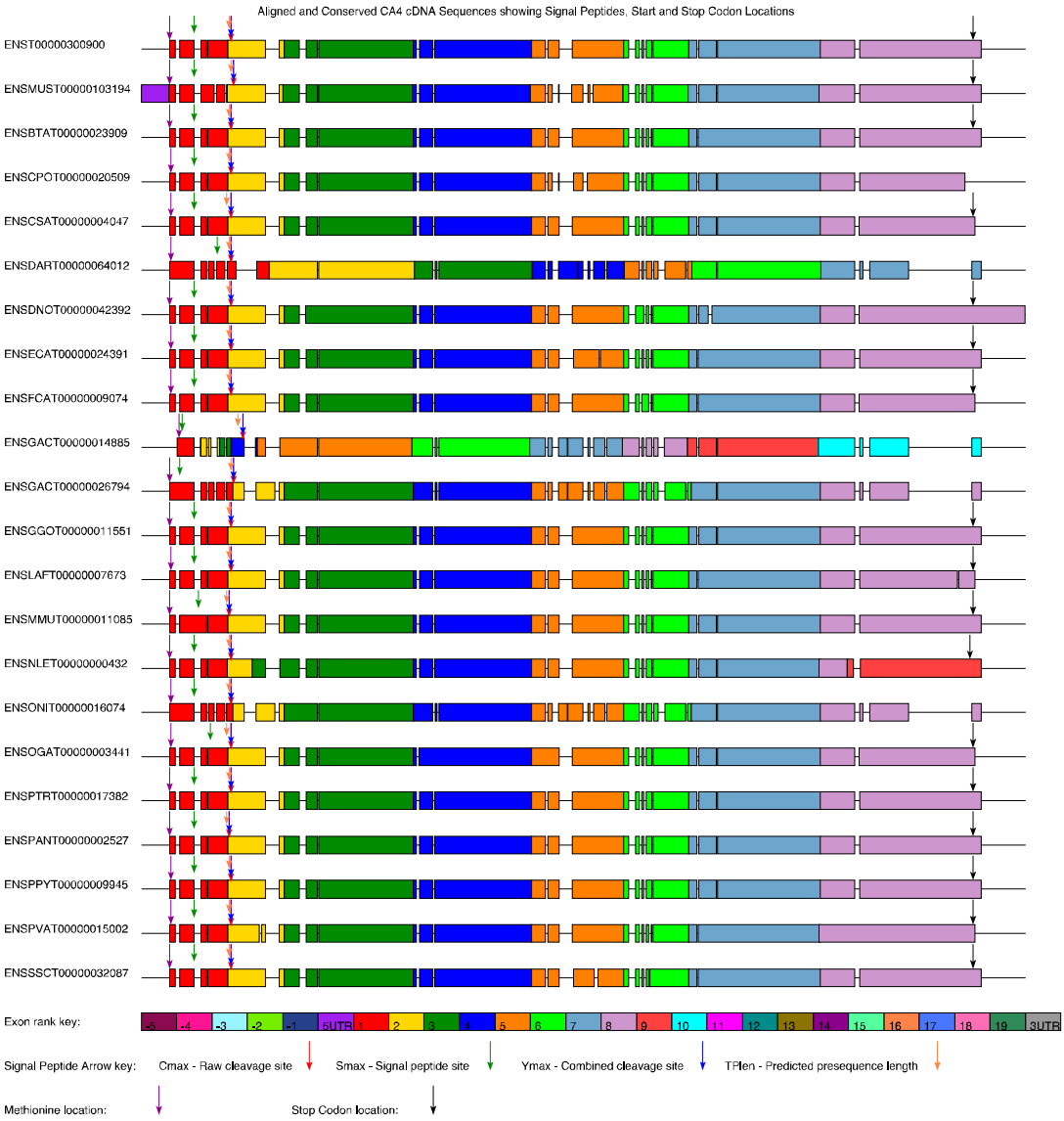

Figure 42 : Exon MSA schematic for conserved transcripts of CA4. The PRANK MSA was zoomed in from position 1560 to position 2700.

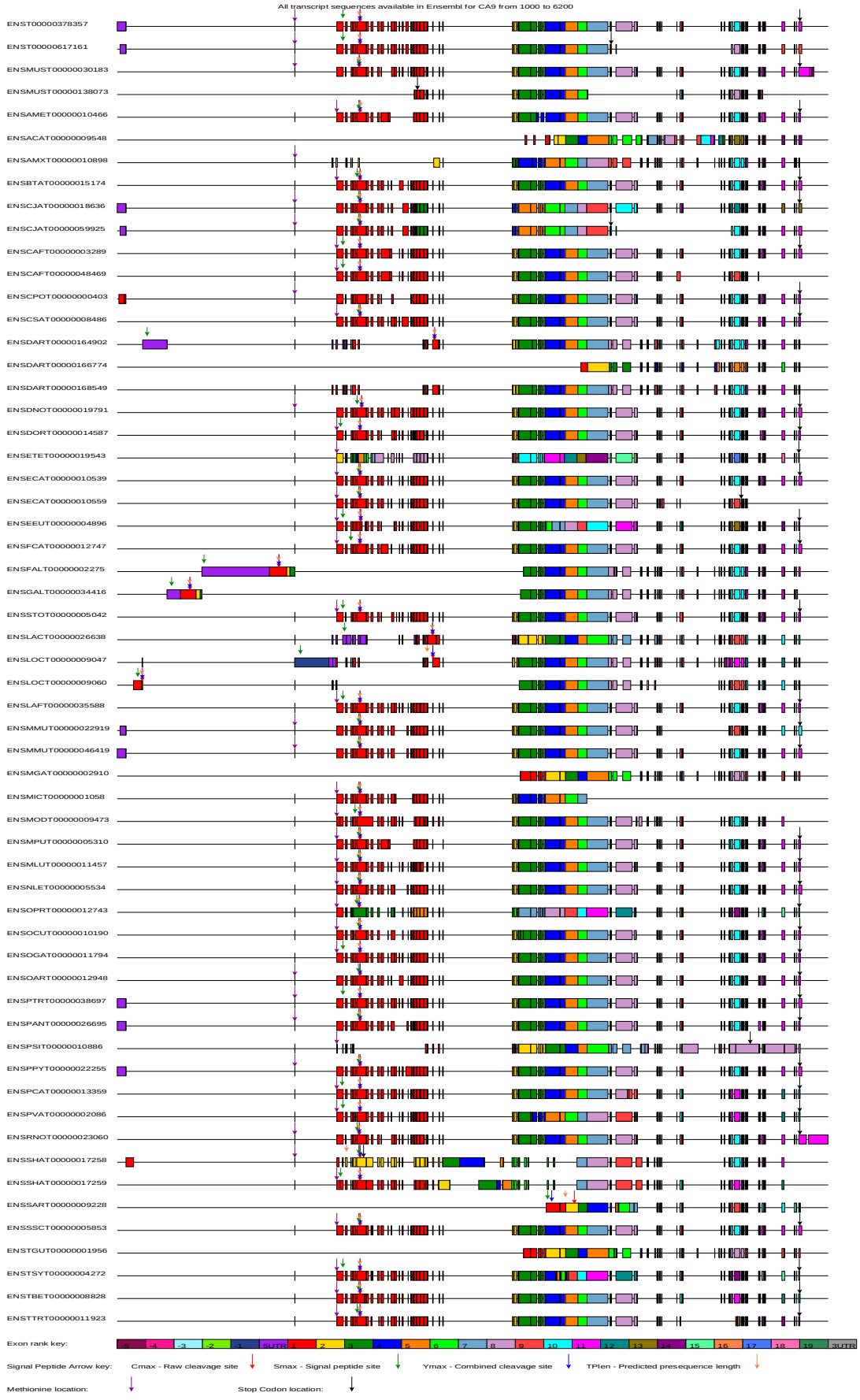

Figure 43 : Exon MSA schematic of all transcripts in Ensembl. This schematic highlights the extra residues found at the start of mammalian sequences. The PRANK MSA was zoomed in from position 1000 to 6200.

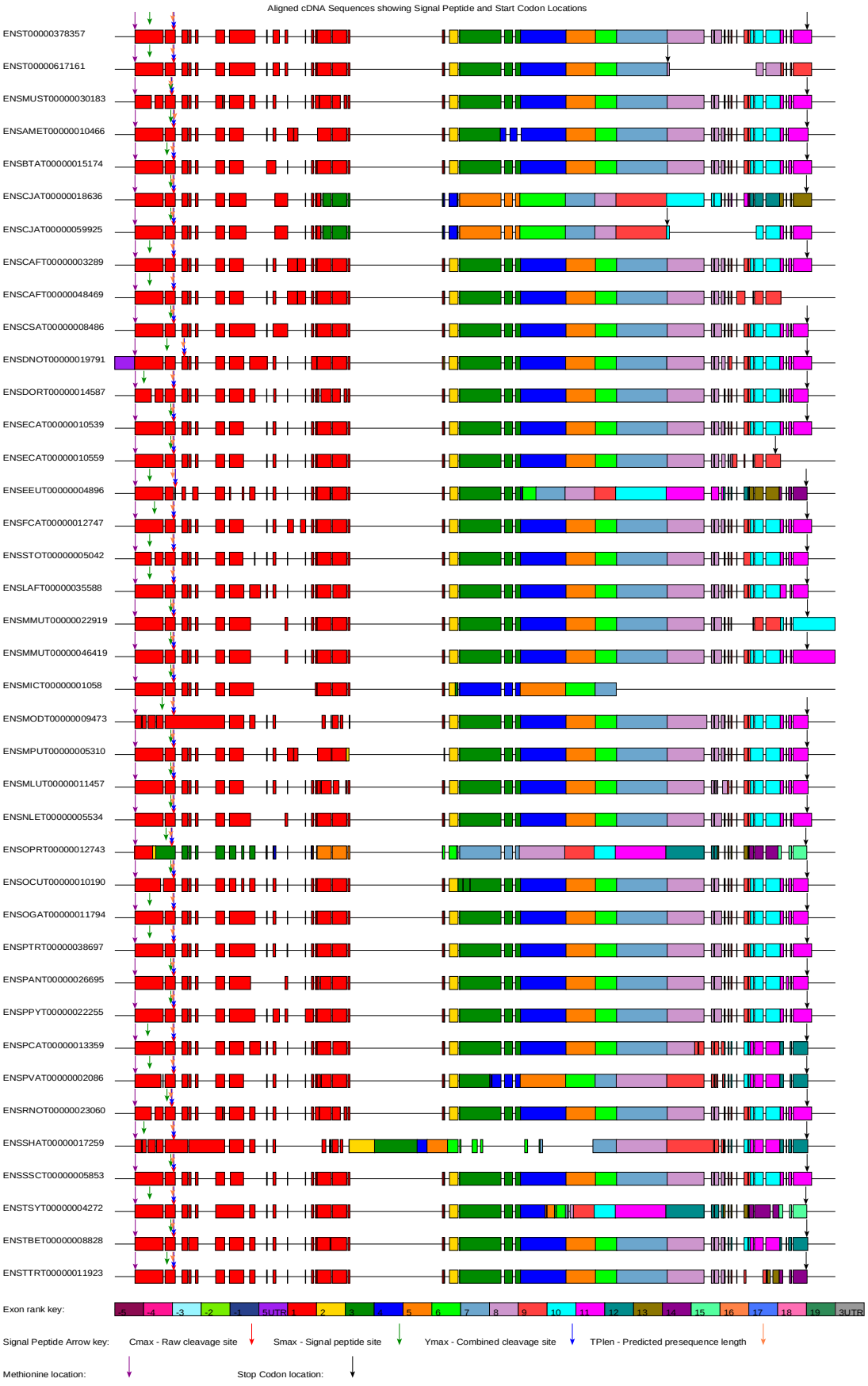

Figure 44 : Exon MSA schematic of transcripts of CA9 that have a start codon and predicted signal peptides. Only the mammalian sequences fulfil these conditions. The PRANK MSA was zoomed in from position 950 to position 3200.

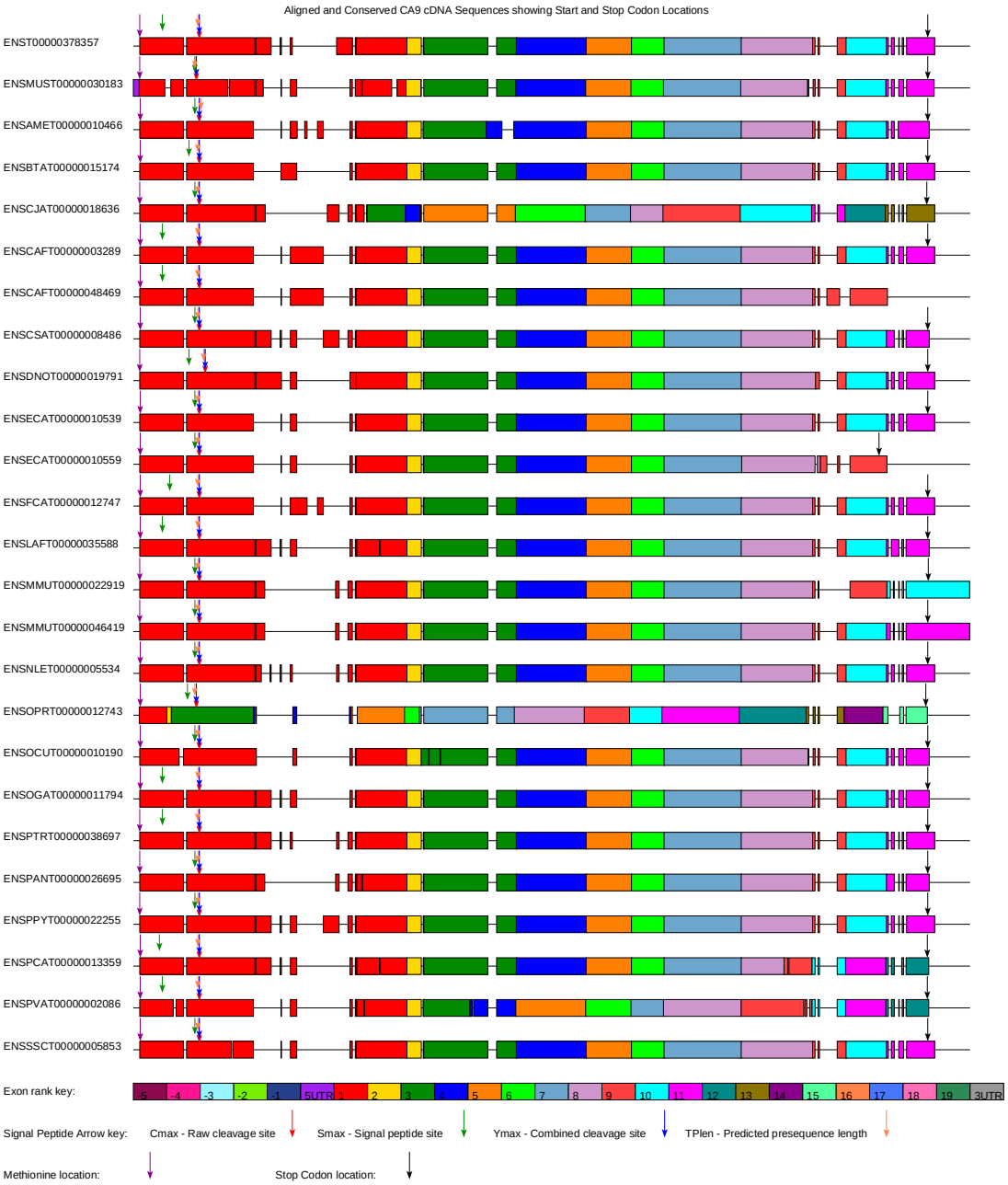

Figure 45 : Exon MSA schematic of conserved transcripts of CA9. The PRANK MSA is zoomed in from position 1000 to 2700. Due to the restrictions within the programming for assessing conserved sequences, only mammalian sequences are shown.

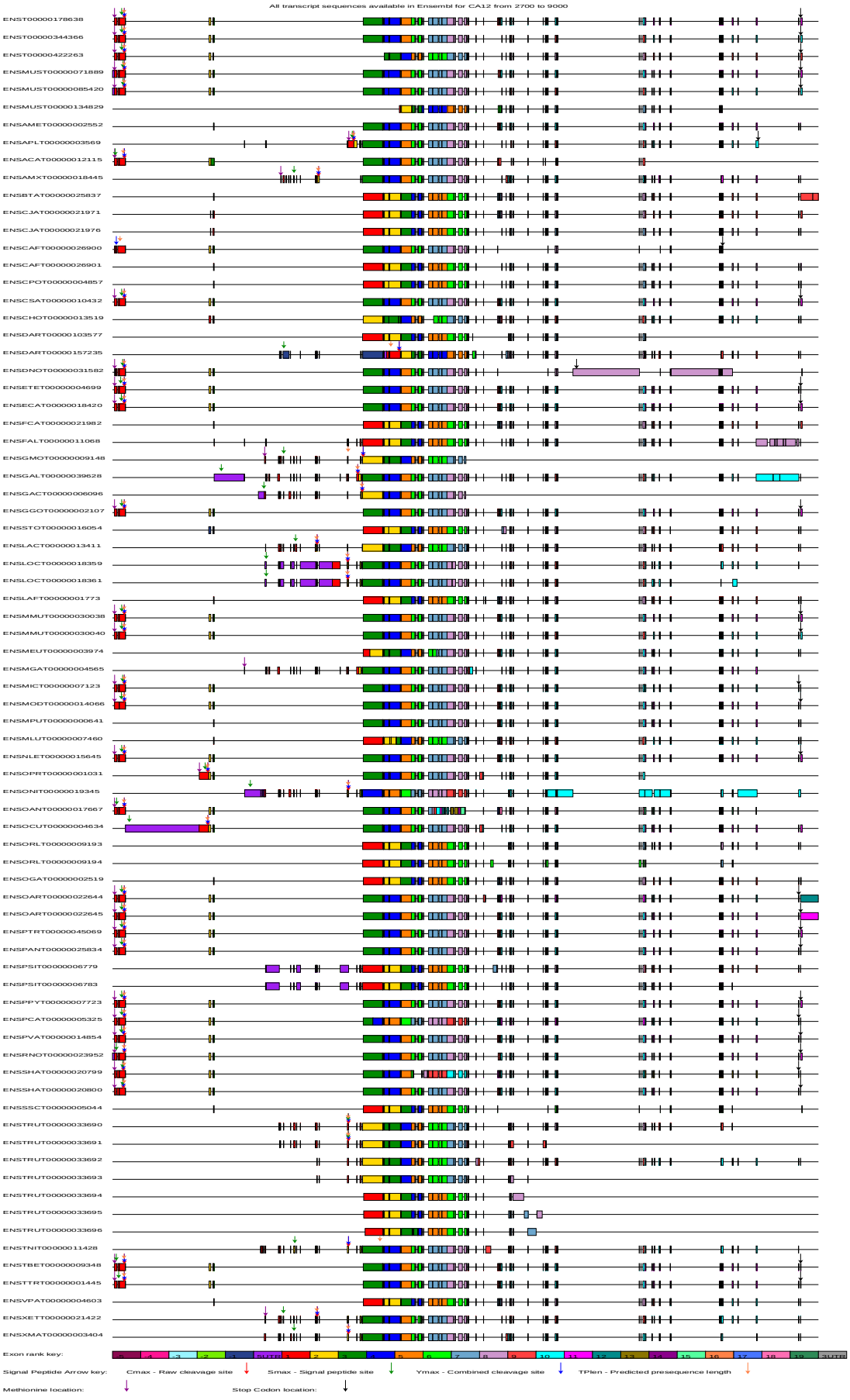

Figure 46 : Exon MSA schematic of all CA12 transcripts. The PRANK MSA has been zoomed in from position 2700 to position 9000.

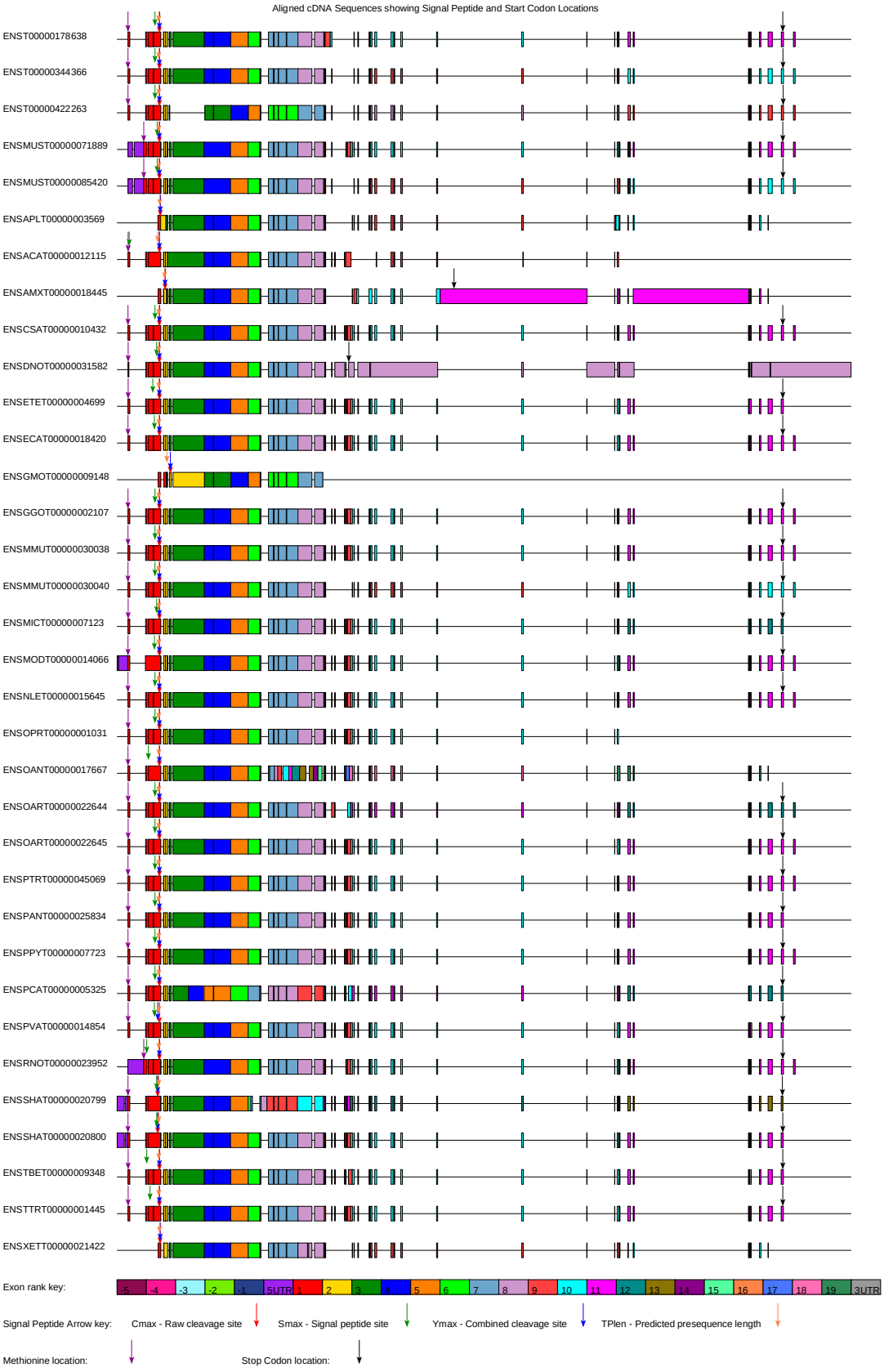

Figure 47 : Exon MSA schematic of CA12 transcripts where all transcripts have a start codon and predicted signal peptides. The PRANK MSA is zoomed in from position 1000 to 5000.

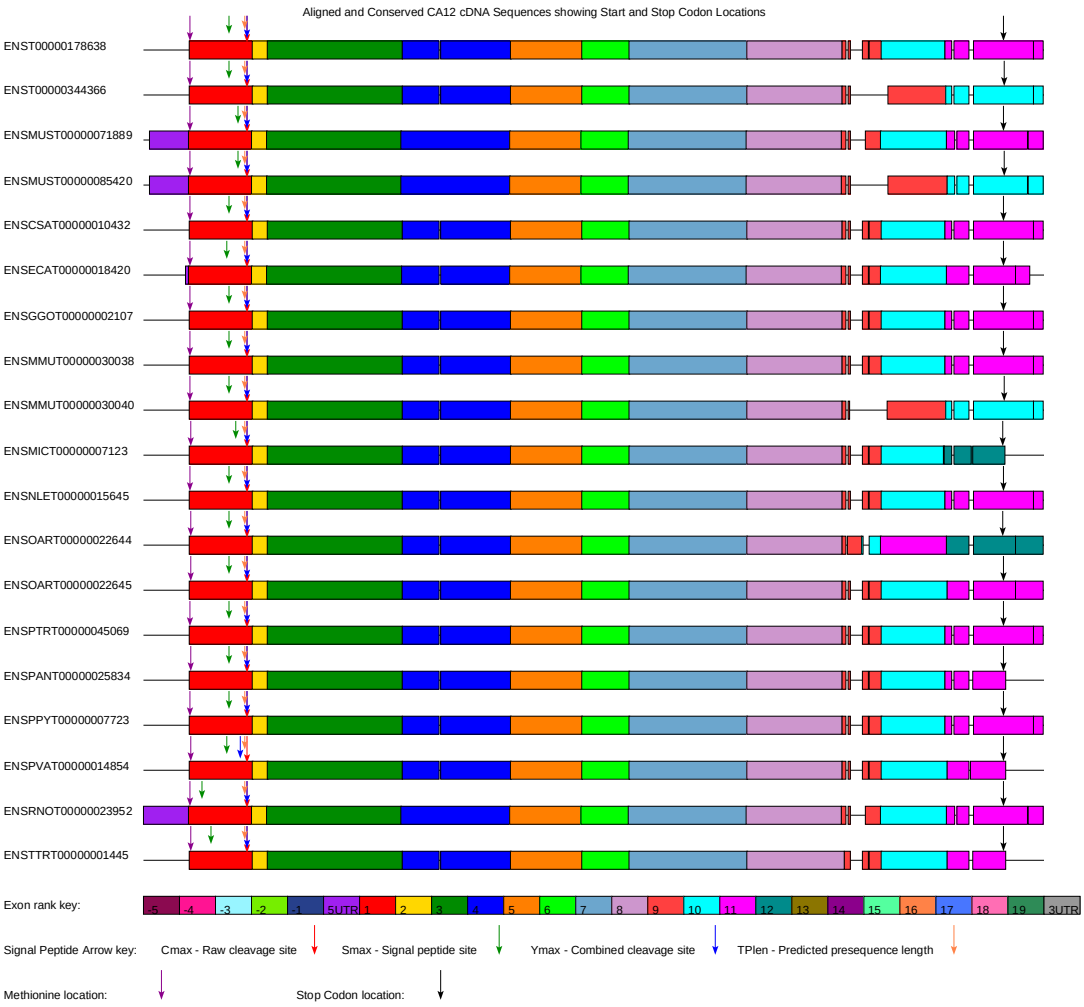

Figure 48 : Exon MSA schematic for CA12 of conserved transcripts where the PRANK MSA has zoomed in from position 800 to position 2000.

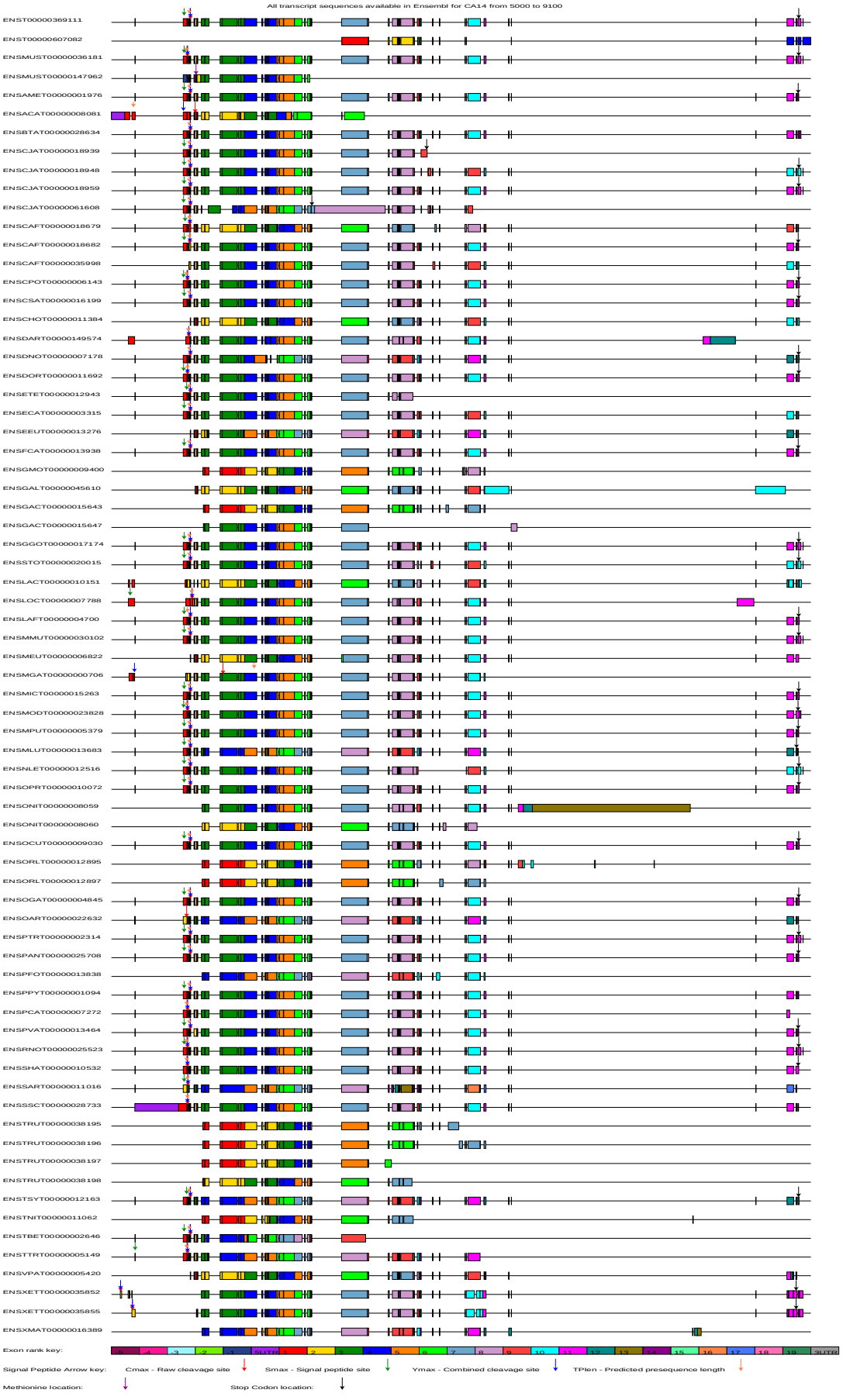

Figure 49 : Exon MSA schematic of all CA14 transcripts. The PRANK MSA is zoomed in from position 5000 to 9100.

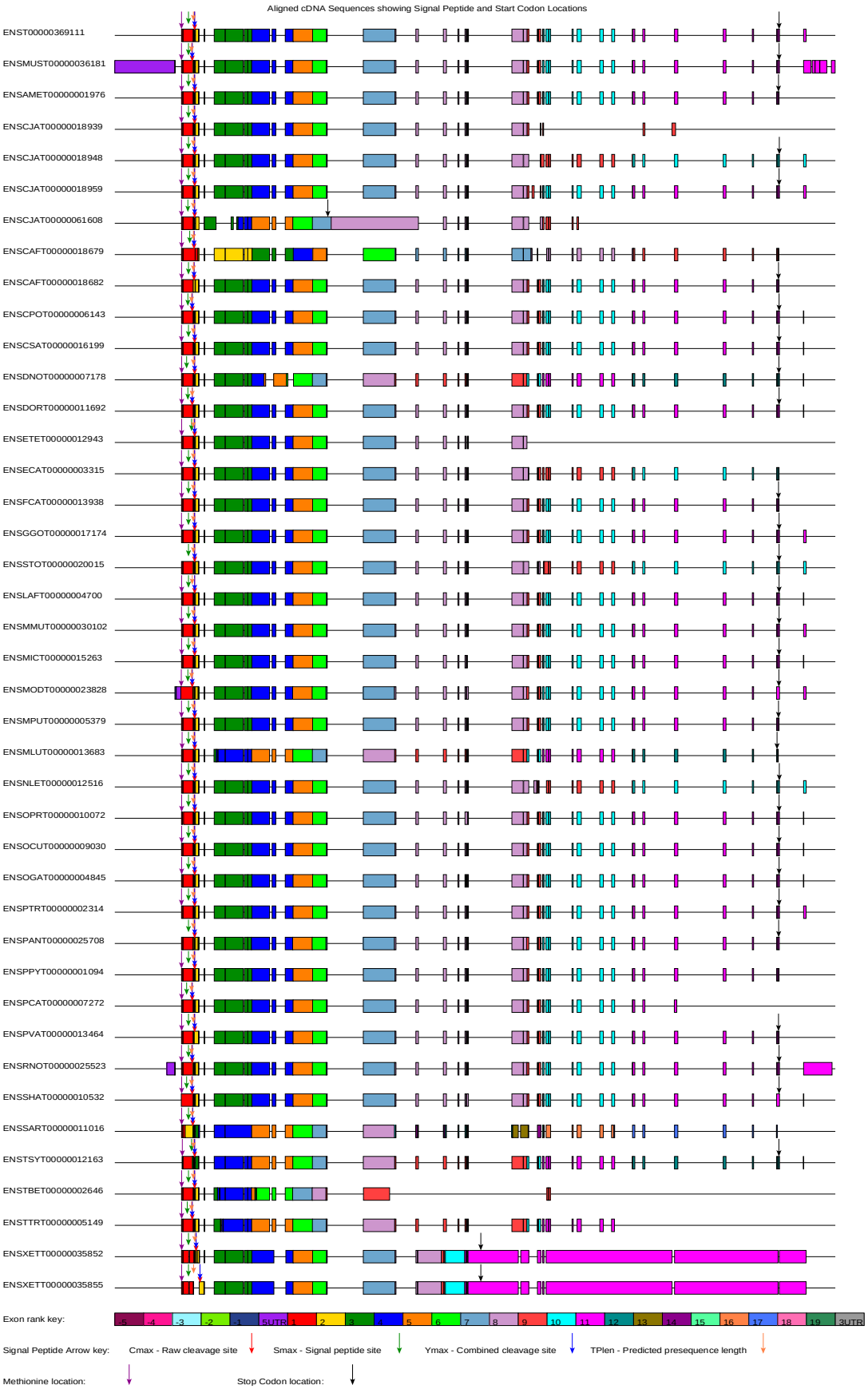

Figure 50 : Exon MSA schematic of CA14 transcripts with a start codon and predicted signal peptide. The PRANK MSA is zoomed in from position 2500 to 6000.

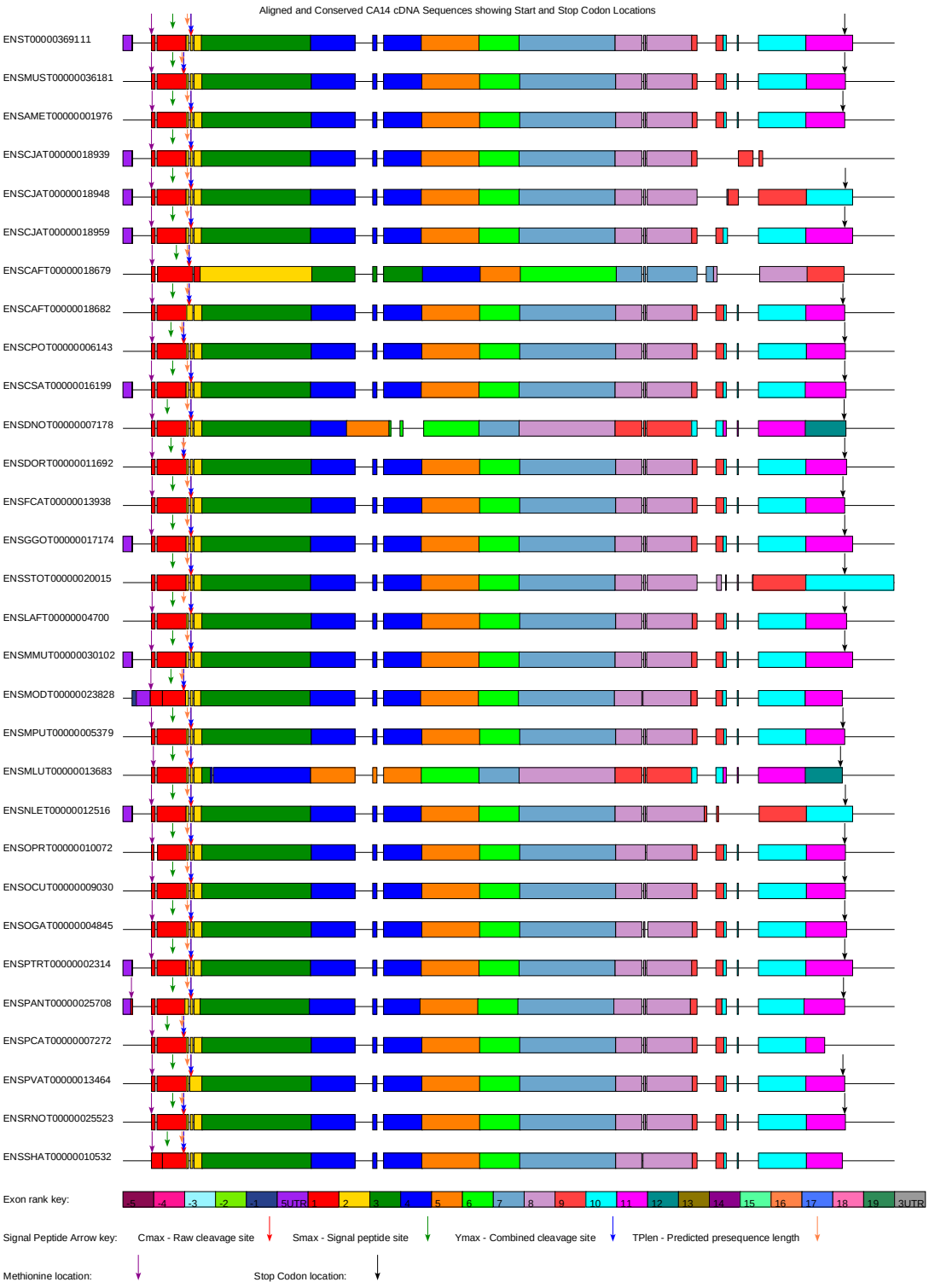

Figure 51 : Exon MSA schematic for conserved CA14 transcripts. The PRANK MSA is zoomed in from position 2600 to position 3870.

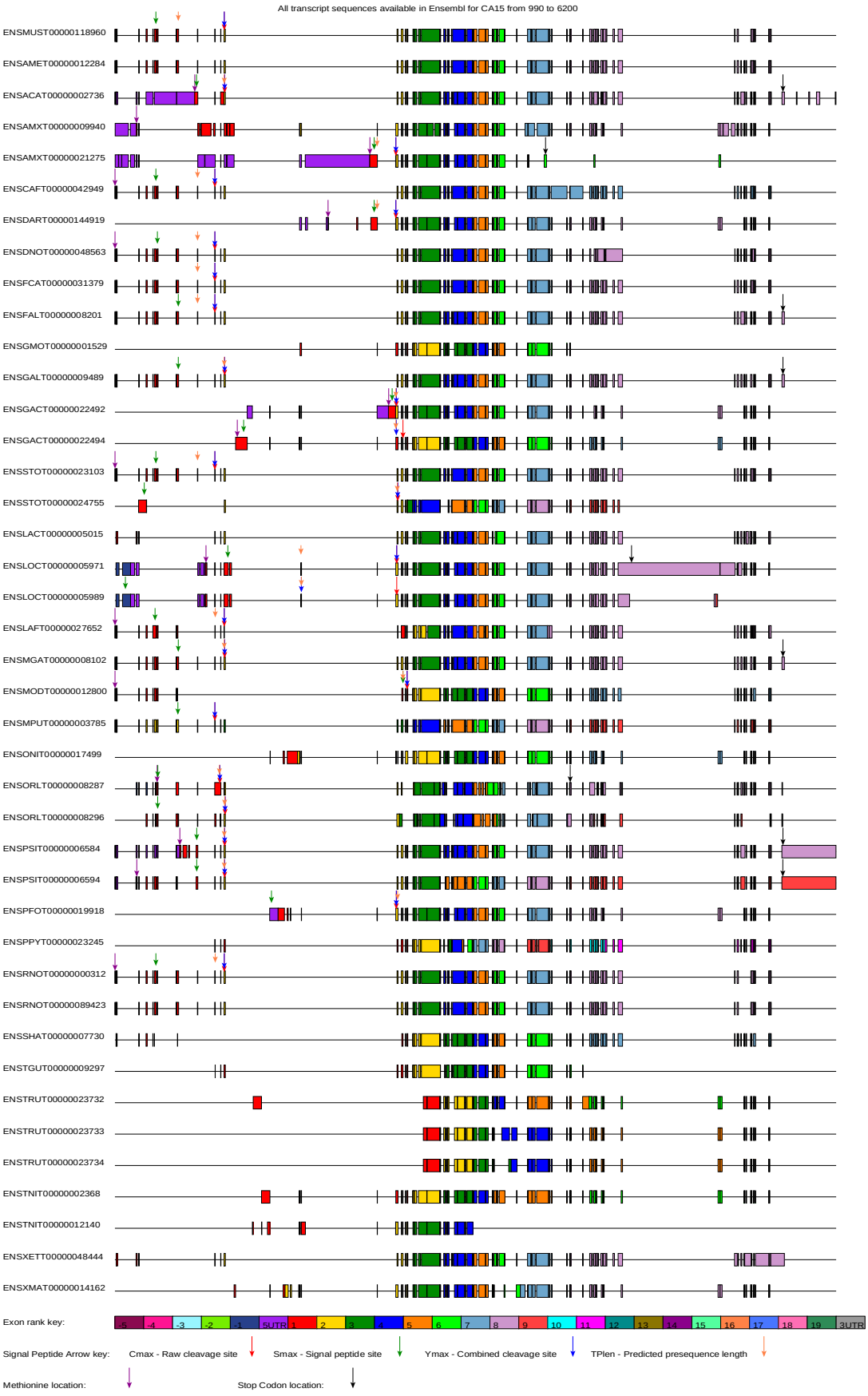

Figure 52 : Exon MSA schematic of all transcripts with a predicted signal peptide and a start codon. The PRANK MSA has been zoomed in from position 990 to position 6200. This CA has more variation in exon 1 than other isoforms in the membrane associated group.

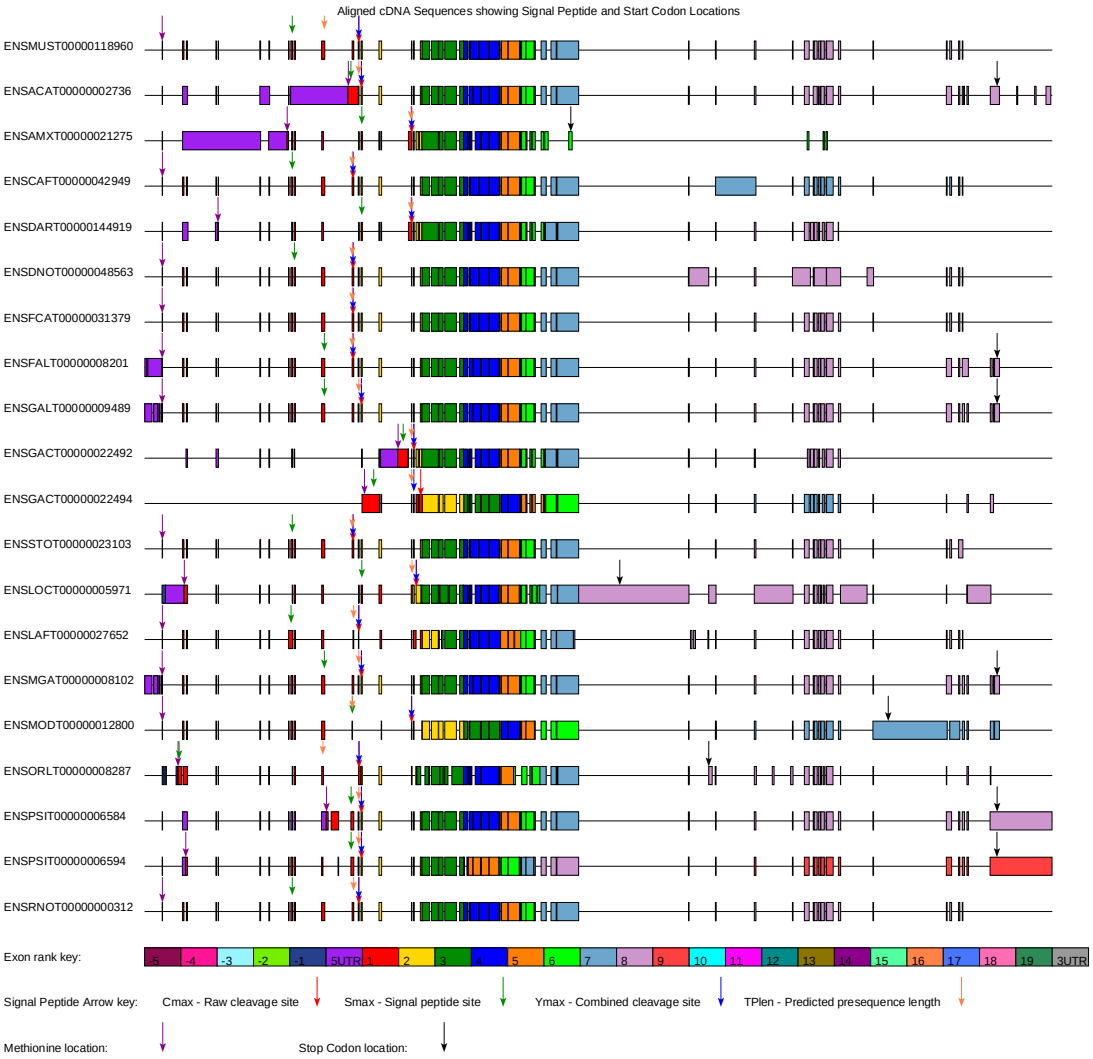

Figure 53 : Exon MSA schematic for CA15 transcripts start codons and predicted signal peptides. The PRANK MSA has been zoomed in from position 1000 to position 5500.

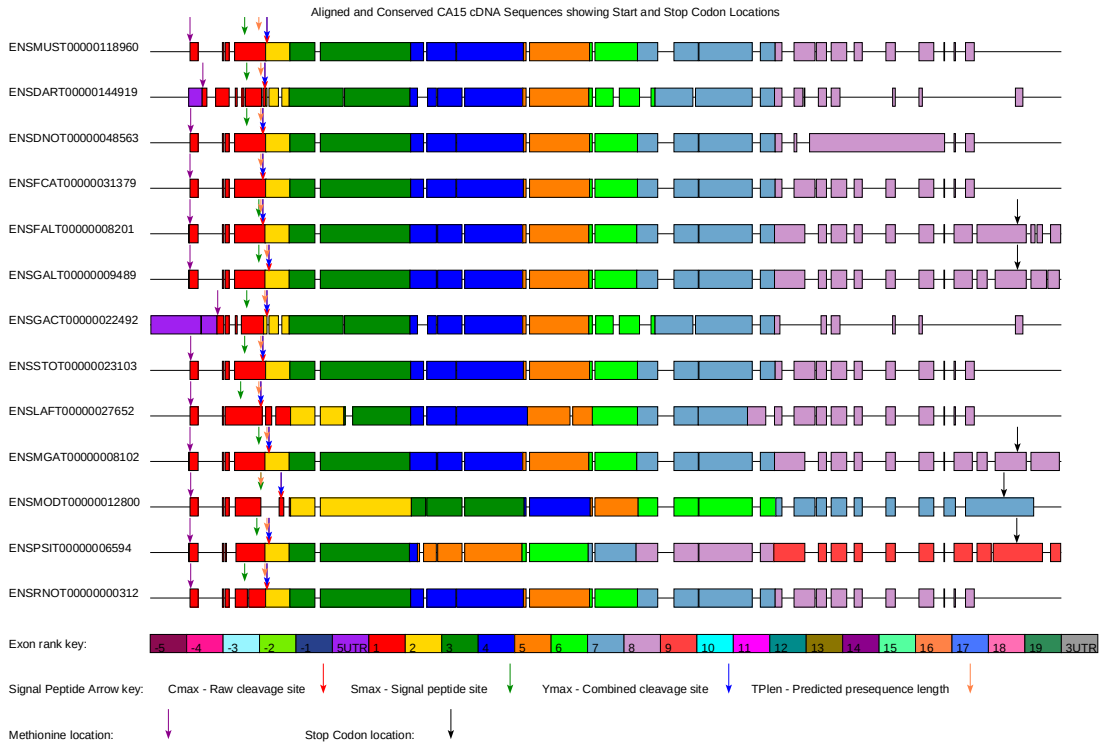

Figure 54 : Exon MSA schematic of conserved transcripts for CA15. The PRANK MSA has been zoomed in from position 650 to 2000.

4.5 The secreted CA - CA6

The only CA that is secreted within the alpha CAs is CA6. There are 81 protein coding CA6 transcripts in CAbase, of which 66 are secreted and have a predicted signal peptide cleavage site. Almost all of the secreted proteins have the cleavage site at the border of exons one and two as is shown in Table 27. Due to the pentraxin domains at the 3' end of the non-mammalian transcripts, the exon MSA schematics (Figure 55 to Figure 57) show greater variance and hence produce more 'choppy' images.

Table 27 : Signal peptide statistics of CA6

CA	Number of transcripts in isoform	Exon containing cleavage site	Percentage
CA6	15		19%
	65	1	80%
	1	2	1%

Table 28 : A summary of the short exon (less than 12 residues long) containing transcripts in CA6. The last three columns (from left to right) refer to the number of transcripts where the 2nd last exon is short, the next column refers to the number of instances where at least one of the last three exons is short and the final column is the count of transcripts where at least one of the 3rd, 4th or 5th exons is at most 11 residues long.

	Transcripts with Short exon(s)	Protein Coding	% Containing Short Exon	Last Exon Short	2nd last exon short	One or more of last 3 exons short	Early exon short
CA6	26	81	32%	5	15	32	4

Table 29 : An exhaustive list of transcripts containing exons that are at most 11 residues long.

CA	EnsTranscriptId	Short Exon	Last Exon	Bases	Species	Start Codon Loctn	Signal P	Loc	RC
CA6	ENSMUST00000126449	7	7	28	Mus musculus				
CA6	ENSAMXT00000025504	9	10	30	Astyanax mexicanus	140	Y	S	1
CA6	ENSCJAT00000000681	3	5	5	Callithrix jacchus	0	Y	S	1
CA6	ENSDART00000079007	9	10	24	Danio rerio	168	Y	S	1
CA6	ENSDNOT00000010137	9	9	5	Dasypus novemcinctus	0	Y	S	2
CA6	ENSETET00000016330	5	11	4	Echinops telfairi	0	Y	S	2
CA6	ENSETET00000016330	10	11	9	Echinops telfairi	0	Y	S	2
CA6	ENSECAT00000001390	2	8	11	Equus caballus	0	Y	S	1
CA6	ENSEEUT00000005563	8	10	11	Erinaceus europaeus	0	Y	S	1
CA6	ENSEEUT00000005563	9	10	15	Erinaceus europaeus	0	Y	S	1
CA6	ENSGMOT00000004259	9	10	2	Gadus morhua		N	_	2
CA6	ENSGACT00000009936	2	10	16	Gasterosteus aculeatus		N	_	4
CA6	ENSGACT00000009946	5	10	2	Gasterosteus aculeatus		N	S	3
CA6	ENSGACT00000009946	9	10	7	Gasterosteus aculeatus		N	S	3
CA6	ENSGGOT00000007553	7	8	14	Gorilla gorilla	0	Y	S	1
CA6	ENSGGOT00000032783	5	9	32	Gorilla gorilla		Y	S	2
CA6	ENSGGOT00000032783	6	9	14	Gorilla gorilla		Y	S	2
CA6	ENSSTOT00000012034	8	10	10	Ictidomys tridecemlineatus	0	Y	S	1
CA6	ENSSTOT00000012034	9	10	10	Ictidomys tridecemlineatus	0	Y	S	1
CA6	ENSMMUT00000047603	8	9	11	Macaca mulatta	0	Y	S	1
CA6	ENSMICT00000016230	8	13	11	Microcebus murinus	0	Y	S	1
CA6	ENSMICT00000016230	9	13	6	Microcebus murinus	0	Y	S	1
CA6	ENSMICT00000016230	10	13	3	Microcebus murinus	0	Y	S	1
CA6	ENSMICT00000016230	12	13	23	Microcebus murinus	0	Y	S	1
CA6	ENSMICT00000016230	13	13	6	Microcebus murinus	0	Y	S	1
CA6	ENSNLET00000011786	8	9	14	Nomascus leucogenys	0	Y	S	1
CA6	ENSOPRT00000012293	2	9	11	Ochotona princeps	0	Y	S	1
CA6	ENSOPRT00000012293	3	9	6	Ochotona princeps	0	Y	S	1
CA6	ENSOGAT00000005428	9	9	28	Otolemur garnettii	0	Y	S	1
CA6	ENSPSIT00000009951	8	10	27	Pelodiscus sinensis		Y	S	1
CA6	ENSPSIT00000009951	9	10	7	Pelodiscus sinensis		Y	S	1
CA6	ENSPFOT00000022522	8	10	21	Poecilia formosa	631	Y	S	1
CA6	ENSPPYT00000002291	7	15	9	Pongo abelii	0	Y	S	1
CA6	ENSPPYT00000002291	8	15	6	Pongo abelii	0	Y	S	1
CA6	ENSPPYT00000002291	9	15	13	Pongo abelii	0	Y	S	1
CA6	ENSPPYT00000002291	10	15	18	Pongo abelii	0	Y	S	1
CA6	ENSPPYT00000002291	11	15	7	Pongo abelii	0	Y	S	1
CA6	ENSPPYT00000002291	12	15	8	Pongo abelii	0	Y	S	1
CA6	ENSPPYT00000002291	14	15	21	Pongo abelii	0	Y	S	1
CA6	ENSTRUT00000022027	2	10	17	Takifugu rubripes		N	S	1
CA6	ENSTRUT00000022030	2	8	4	Takifugu rubripes		N	_	2
CA6	ENSTSYT00000002533	6	11	10	Tarsius syrichta		N	_	1
CA6	ENSTSYT00000002533	7	11	11	Tarsius syrichta		N	_	1
CA6	ENSTSYT00000002533	8	11	18	Tarsius syrichta		N	_	1
CA6	ENSTSYT00000002533	9	11	15	Tarsius syrichta		N	_	1
CA6	ENSTSYT00000002533	10	11	15	Tarsius syrichta		N	_	1
CA6	ENSTSYT00000002533	11	11	9	Tarsius syrichta		N	_	1
CA6	ENSTBET00000017045	8	12	11	Tupaia belangeri	0	Y	S	1
CA6	ENSTBET00000017045	9	12	9	Tupaia belangeri	0	Y	S	1
CA6	ENSTBET00000017045	10	12	9	Tupaia belangeri	0	Y	S	1
CA6	ENSTBET00000017045	11	12	5	Tupaia belangeri	0	Y	S	1

Table 30 : SignalP and TargetP results for CA6

TranscriptId	Cmax_Pos	Ymax_Pos	Smax_Pos	TPlen	Loc	RC
ENST00000377436	1	1	1	1	S	1
ENST00000377442	1	1	1	1	S	1
ENST00000377443	1	1	1	1	S	1
ENST00000480186	1	1	1	1	S	1
ENST00000549778	1	1	1	1	S	1
ENSMUST00000030817	1	1	1	1	S	1
ENSAMET00000004280	1	1	1	1	S	1
ENSAMXT00000025504	1	1	1	1	S	1
ENSBTAT00000012525	1	1	1	1	S	3
ENSCJAT00000000681	1	1	1	1	S	1
ENSCJAT00000000686	1	1	1	1	S	1
ENSCAFT00000031340	1	1	1	1	S	2
ENSCPOT00000015319	1	1	1	1	S	1
ENSCSAT00000017772	1	1	1	1	S	1
ENSDART00000079007	1	1	1	1	S	1
ENSDART00000132733	1	1	1	1	S	1
ENSDNOT00000010137	1	1	1	1	S	2
ENSDORT00000009199	1	1	1	1	S	1
ENSETET00000016330	1	1	1	1	S	2
ENSECAT00000000971	1	1	1	1	S	1
ENSECAT00000001390	1	1	1	1	S	1
ENSEEUT00000005563	1	1	1	1	S	1
ENSFCAT00000008240	1	1	1	1	S	1
ENSFALT00000007711	2	2	2	2	S	1
ENSGALT00000003763	1	1	1	1	S	1
ENSGGOT00000007553	1	1	1	1	S	1
ENSGGOT00000031094	1	1	1	1	S	1
ENSGGOT00000032783	1	1	1	1	S	2
ENSSTOT00000012034	1	1	1	1	S	1
ENSLACT00000017871	1	1	1	1	S	1
ENSLOCT00000008640	1	1	1	1	S	1
ENSLOCT00000008660	1	1	1	1	S	1
ENSLAFT00000021558	1	1	1	1	S	1
ENSMMUT00000009241	1	1	1	1	S	1
ENSMMUT00000027656	1	1	1	1	S	1
ENSMMUT00000047603	1	1	1	1	S	1
ENSMGAT00000004177	1	1	1	1	S	1
ENSMICT00000016230	1	1	1	1	S	1
ENSMODT00000011952	1	1	1	1	S	2
ENSMPUT00000006426	1	1	1	1	S	1
ENSMLUT00000001387	1	1	1	1	S	1
ENSNLET00000011782	1	1	1	1	S	1
ENSNLET00000011786	1	1	1	1	S	1
ENSOPRT00000012293	1	1	1	1	S	1
ENSOCUT00000012581	1	1	1	1	S	1
ENSORLT00000007105	1	1	1	1	S	1
ENSORLT00000007107	1	1	1	1	S	1
ENSOGAT00000005428	1	1	1	1	S	1
ENSOART00000010042	1	1	1	1	S	2
ENSPTRT00000000250	1	1	1	1	S	1
ENSPANT00000010196	1	1	1	1	S	1
ENSPSIT00000009937	1	1	1	1	S	1
ENSPSIT00000009951	1	1	1	1	S	1
ENSPFOT00000014289	1	1	1	1	S	2
ENSPFOT00000022522	1	1	1	1	S	1
ENSPPYT00000002291	1	1	1	1	S	1
ENSRNOT00000051309	1	1	1	1	S	1
ENSSSCT00000003762	1	1	1	1	S	1

TranscriptId	Cmax_Pos	Ymax_Pos	Smax_Pos	TPlen	Loc	RC
ENSTRUT00000022028	1	1	1	1	S	2
ENSTRUT00000022031	1	1	6	1	S	1
ENSTNIT00000008220	1	1	1	1	S	1
ENSTNIT00000018434	1	1	1	1	S	1
ENSTBET00000017045	1	1	1	1	S	1
ENSTTRT00000007079	1	1	1	1	S	1
ENSXETT00000026010	1	1	1	1	S	1
ENSXMAT00000003931	1	1	1	1	S	2

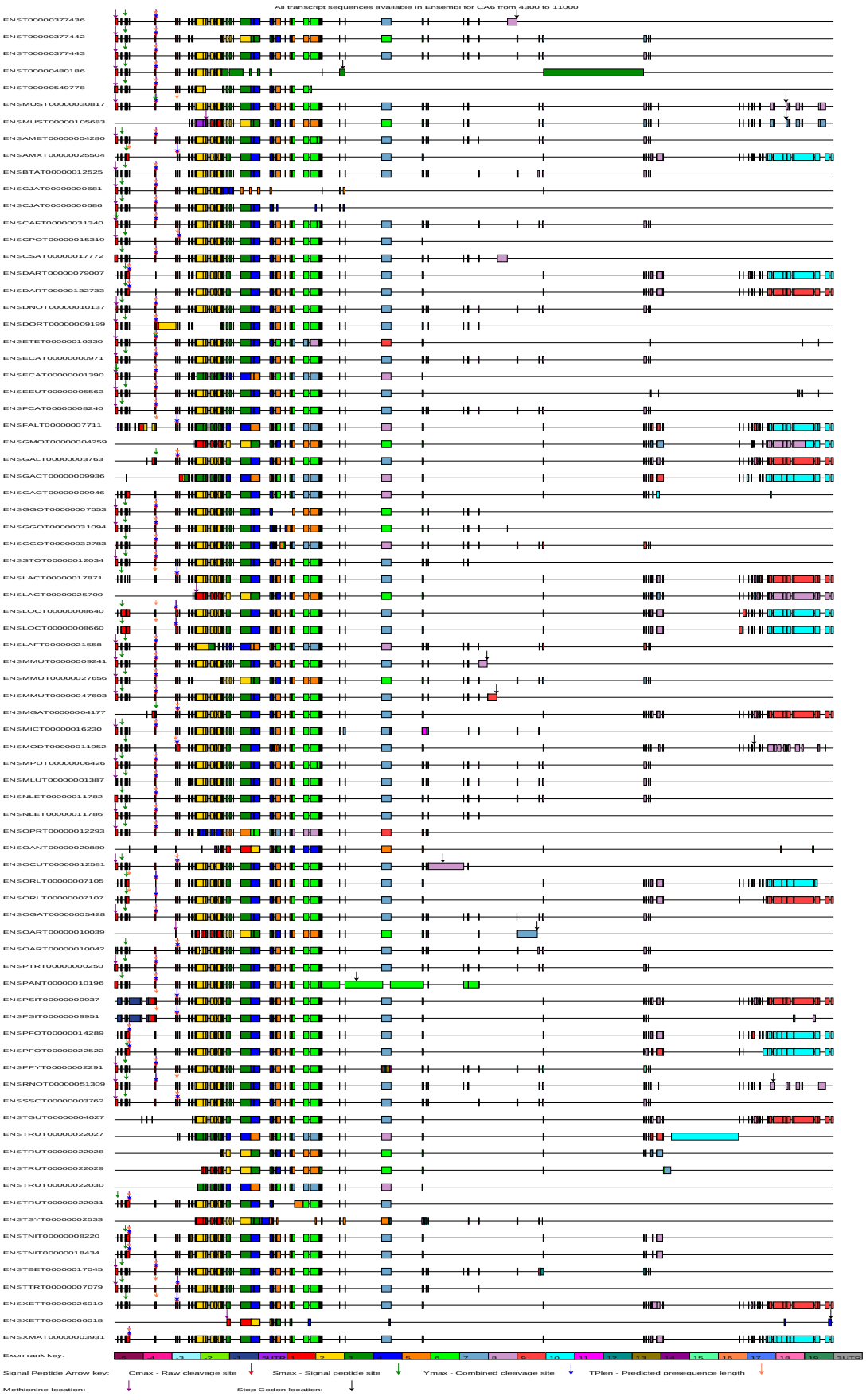

Figure 55 : Exon schematic of all CA6 transcripts. The PRANK MSA has been zoomed in from position 4300 to 11000.

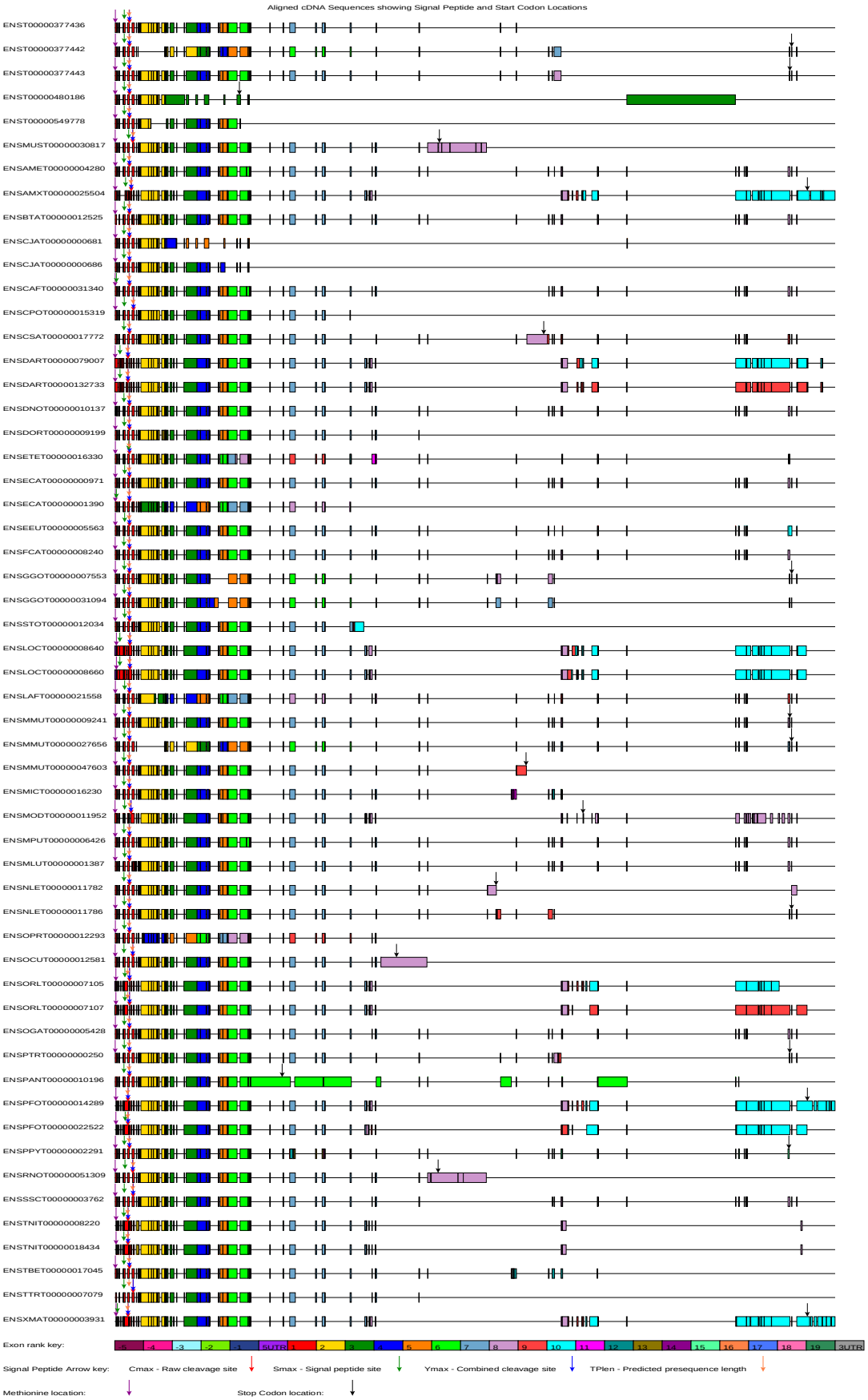

Figure 56 : Exon MSA schematic of CA6 transcripts with a signal peptide. The PRANK MSA was zoomed from position 2300 to 8300 to include as many as possible start and stop codon positions for the majority of transcripts.

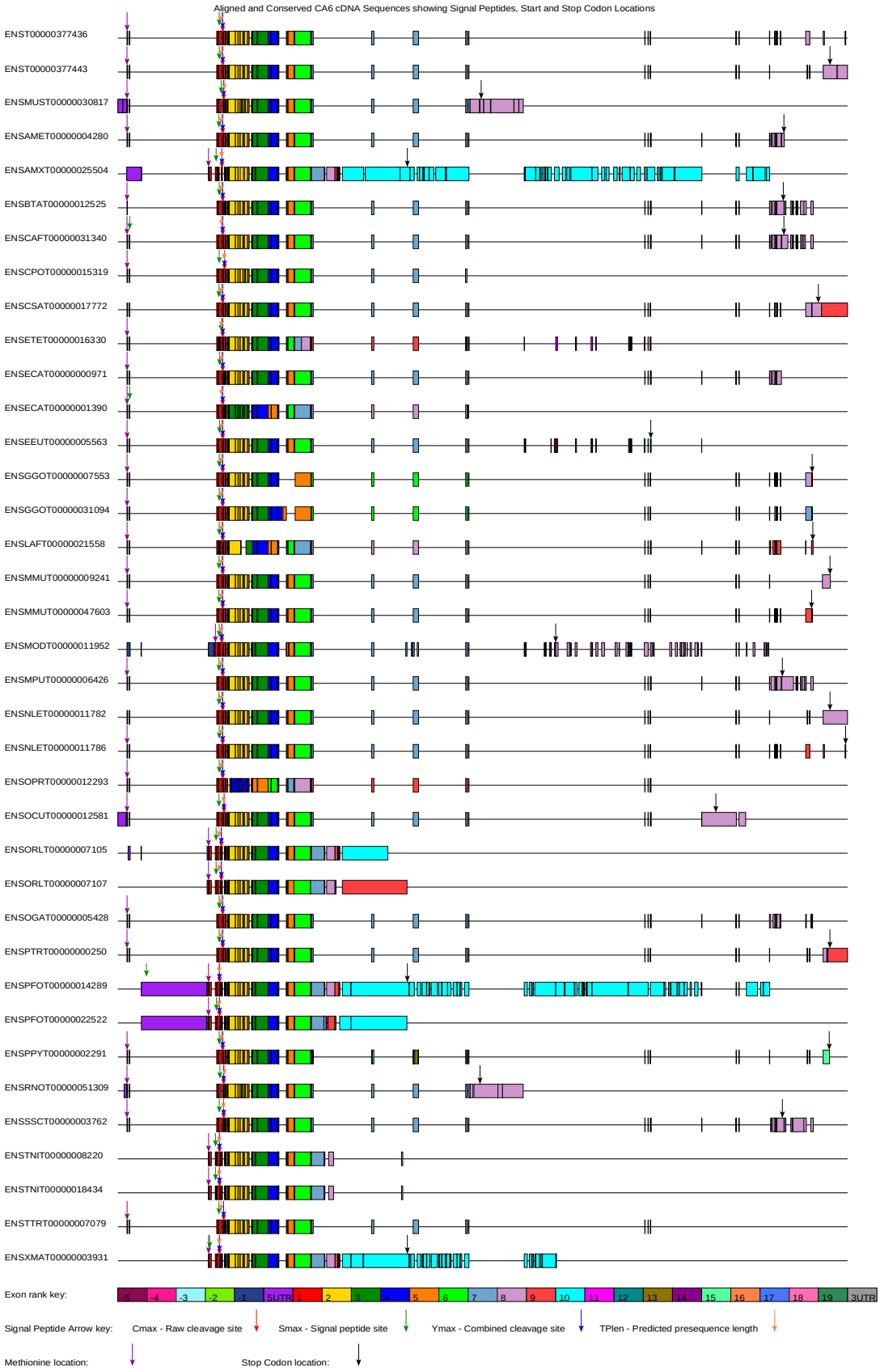

Figure 57 : Exon MSA schematic of conserved CA6 transcripts. The criteria for length conservation have been relaxed so as not to exclude the non-mammalian transcripts with the pentraxin domains. The PRANK MSA has been zoomed in from position 1000 to 8000 to capture most of the start and stop codon positions.

5 Research Goals

The aims of this research were:

- To create a database specific to the needs of carbonic anhydrase researchers.
- To discover if the location of the signal peptides and mitochondrial targeting peptides are conserved within each CA for vertebrates, and
- To enable the quality control of predicted gene models one exon at a time.

These aims were achieved by creating two pipelines:

1. The first pipeline to interrogate all the necessary online databases (Ensembl, NCBI, UniProt, CBS and PDB) to assemble and organize the data into tables for CAbase; and
2. The second pipeline to create a visual display of the aligned exons of the cDNA transcripts contained within CAbase with indicators to show the positions of start and stop codons along with the locations of the predicted signal peptide and mitochondrial targeting peptides.

The second pipeline in particular is an example of how CAbase can be used to answer research questions through querying a database. Furthermore it facilitates the quality control of the predicted genes one exon at a time and it allows users to see evolutionary events such as exon swapping or the conservation of short exons within some CAs.

6 Discussion

This chapter considers the two tools developed for this thesis; both were implemented using a scripted pipeline written in Python. The first pipeline created a database named CAbase while the second is an example of using CAbase to create exon MSA schematics that visually show how conserved the locations of exon boundaries, exon lengths, signal peptides and mitochondrial targeting peptides are in the vertebrate alpha carbonic anhydrases.

6.1 CAbase

The first pipeline created a database named CAbase specifically to host information of the carbonic anhydrases. CAbase can significantly speed up the process of finding patterns in data since the data can be provided in a format that is useful for CA researchers. A weakness of general data sources such as Ensembl is that it attempts to provide data for all proteins, genomes etc. and as such it is not always in a convenient format for someone that is interested in a subset of that data. Furthermore, the generalised data sources use automated predictions to generate sequences where there is less supporting data from the same species, thus introducing potential errors. CAbase is a step towards remedying this situation for those inhabiting the CA research world.

CAbase is available on the Amazon Web Server (AWS) for anyone to query. Currently it is necessary for the user to have MySQL installed to use CAbase since AWS hosts the database and there has been no friendly user interface set up yet. As such it is also

necessary for users to know SQL to access the data in CAbase, though this is not as complicated a language to learn as some others.

Databases are only as useful as the data that they store - if users cannot have faith that the data that they are accessing is correct, the database is useless. This criteria of databases has been challenging to implement and has exposed a few idiosyncrasies of some of the online databases and the tools that are provided to access that data to populate CAbase, e.g. when using the Ensembl REST API to find homologs of genes Ensembl includes synonyms in the results. The synonyms and 'real' genes are returned in a seemingly random order and there is no warning provided. This situation and several others like it results in the code being very large (over 2000 lines) because of the numerous checks and filters that are needed to ensure that the data is reliable.

Sending many queries to the online databases takes a long time to execute if the user does not want to be blacklisted by those services - CAbase takes over 24 hours to populate from scratch. The code that populates CAbase, CAbaseGenerator.py, has wait times included to prevent blacklisting.

CAbase also stores some calculated values such as the location of the signal and targeting peptides. This functionality can be expanded to include any data that the future owners of CAbase deem to be valuable.

6.2 Exon MSA schematic of carbonic anhydrases

The pipeline for analysing the conservation of exons and visually displaying the locations of the cleavage sites of the signal and mitochondrial targeting peptides used Löytynoja's PRANK program to create a phylogenetically aware codon alignment of the cDNA sequences. The programs SignalP and TargetP were run to assess if the sequence had a reliably predicted signal or mitochondrial targeting peptide.

Clustal Omega was also tried as an alignment program, but it would often align the first spurious matching bases of the sequences. Meaningful alignments could be gained for parts of the sequences from Clustal Omega but only after individually adjusting the gap costs for each CA. (Sievers, et al., 2011) PRANK gave consistently meaningful results because both sequence homology and conserved structures are considered when aligning the sequences. This resulted in the codons being aligned in the correct phase.

Initially all of the transcripts for the CA of interest are aligned using PRANK. The exon MSA schematic allows a visual inspection of the aligned sequences, thus facilitating the assessments of individual exons. The pipeline then attempts to assess the quality of transcripts using a conservative approach. Transcripts are gradually removed from the 'conserved list' based initially on length, and then on a series of tightening criteria for inserts and deletions. If the number of transcripts remaining on the conserved list is less than 28, the assessment is abandoned and an exon MSA schematic is drawn. Extra conditions have been included for CAs that have unusual properties such as CA6 with the non-mammalian PTX domain and CA9 with the mammalian PG domain. Excluding potentially mispredicted transcripts from the alignments used for visualizing

the CAs removes the background noise from the information indicating how CAs are evolving.

By using the command line options available with Exon_Analysis.py, users are able to choose which type of exon MSA schematic they wish to view and on which region of the MSA they wish to focus. As an example, giving the options -prot CA1 --type full -start 1000 -end 2000 will draw all transcripts from the PRANK CA1 MSA from the alignment position 1000 to the alignment position 2000. By zooming in to a particular part of the exon MSA schematic the scaled rectangles representing the exons and gaps are expanded, thus conveying more information to the user. This program can be modified to perform the same visualization for any protein.

The exon MSA schematics for all CAs show greater variation in 5'UTR and 3'UTR when compared to the coding regions (Figure 58, Figure 59, Figure 60 and Appendix A). This can be seen in the exon MSA schematics as multiple small bars where PRANK 'spreads' the aligned areas with greater variation.

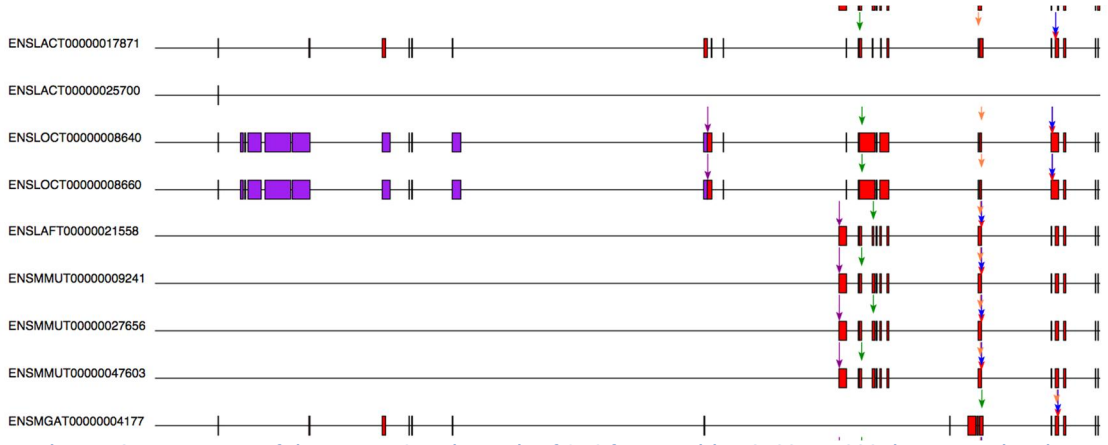

Figure 58 : An excerpt of the exon MSA schematic of CA6 from position 2500 to 5000 demonstrating the variation found in the 5'UTR regions. PRANK has spread the alignment, which is represented as bars in this image. The purple bars represent the 5'UTR and the red bars indicate the location of the first coding exon.

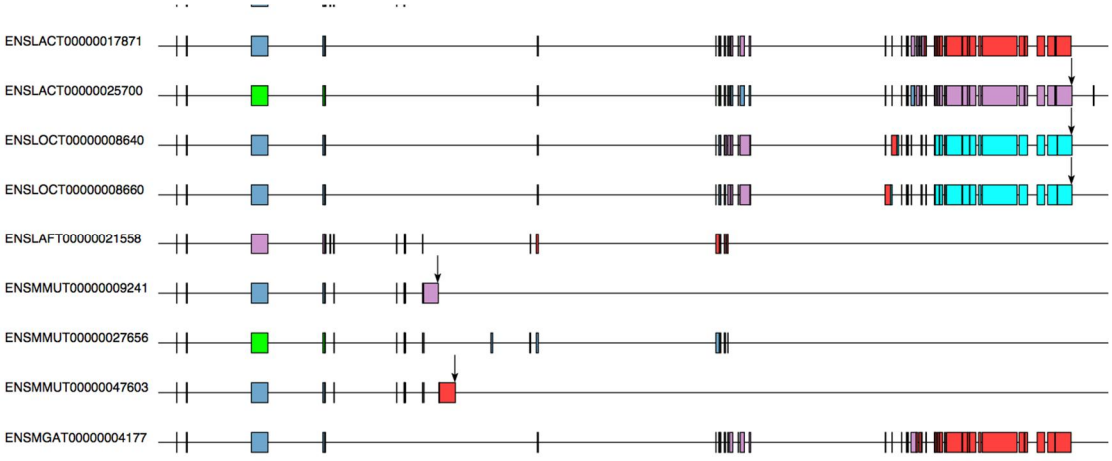

Figure 59 : An excerpt of the exon MSA schematic of CA6 from position 6300 to 11500 showing the variation found in the 3' areas and the pentraxin domains of the coelacanth, spotted gar and chicken transcripts. The black bars in the exon MSA schematic represent only one or two bases. The colour key is as follows: light green: exon 6, light blue: exon 7, light purple: exon 8, dark pink: exon 9, cyan: exon 10.

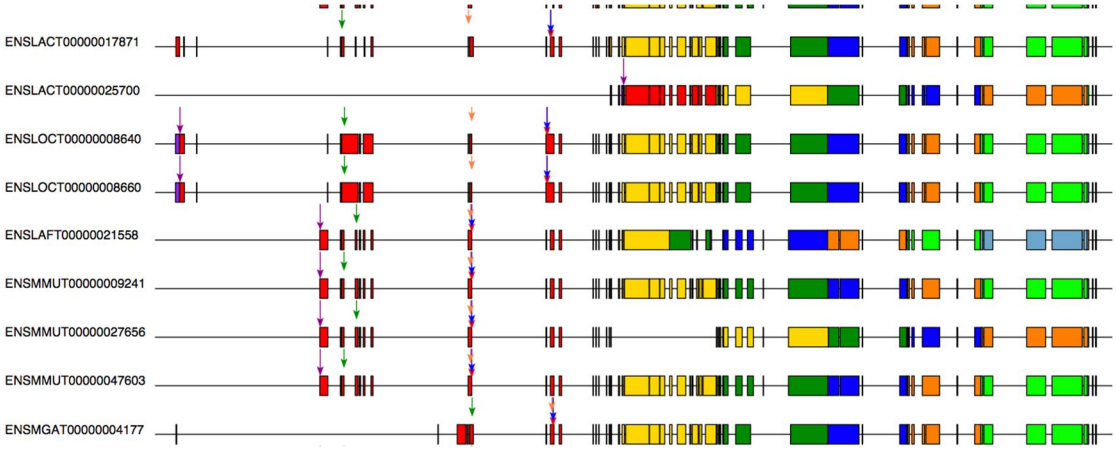

Figure 60 : An excerpt of the CA6 exon MSA schematic zoomed in from position 3900 to position 6300 demonstrating the conserved nature of the coding regions of the transcripts. The colour key is as follows: purple: 5'UTR, red: exon 1, yellow: exon 2, green: exon 3, royal blue: exon 4, orange: exon 5, light green: exon 6, light blue: exon 7.

There is very little variation within the catalytic domains, as can be seen from the proportionate sizes of the coloured rectangles representing the exons. The length of the exons and locations of the exon boundaries are also highly conserved - even for CA8, CA10 and CA11, which are all CARPs since they lack the histidine residues to bind the zinc atom to the protein. The conservation of the CA exon patterns highlights any aberrant transcripts, thus facilitating the quality control of the predicted genes one exon at a time. It also allows the user to quickly see how alternative splicing transcripts differ and to see where exons may have split.

Within the coding region some exons have split with an intron insert as can be seen below in Figure 61 for ENSAMET0000002770 in CA1. Here exon 6 (lime green) has been split and the 5' part is only 9 bases long. The 3' part of the exon is now exon 7 so the transcript has gained an exon while retaining the CDS length that it had before. This kind of exon splitting is visible in all of the CAs.

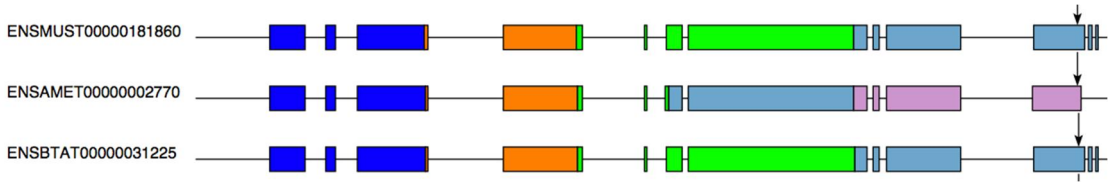

Figure 61 : ENSAMET00000002770 from CA1, has had an intron inserted in exon 6, splitting it into two exons. This is a small excerpt from a full alignment of CA1 where the MSA is zoomed in from position 5000 to 6000. The colour key is as follows: royal blue: exon 4, orange: exon 5, light green: exon 6, light blue: exon 7, light purple: exon 8.

Alternative splicing can be seen in Figure 62 in the human CA6 transcripts. At the time of accessing the data from Ensembl, the true exon number was not given in the download, only the order of transcription. Hence in ENST00000377442 exon 2 (yellow) is probably exon 3 in reality. In future versions of this software this issue could be overcome by looking at the chromosomal location of the exons.

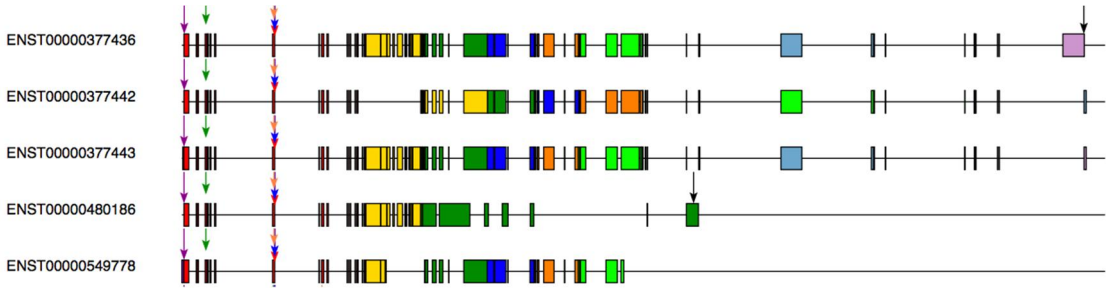

Figure 62 : Alternative splicing in human CA6 transcripts in the exon MSA schematic zoomed from position 3900 to 9000. The colour key is as follows: purple: 5'UTR, red: exon 1, yellow: exon 2, green: exon 3, royal blue: exon 4, orange: exon 5, light green: exon 6, light blue: exon 7, light purple: exon 8.

Seeing the general conservation of exon patterns in CAs, aberrant exon patterns are a clear indication of suspicious gene models, as is the case with the platypus transcript, ENSOANT00000020880. This transcript is missing the 5' sequence for the 1st exon. At the time of writing there were no EST or nucleotide sequences in NCBI that could supply the missing sequence.

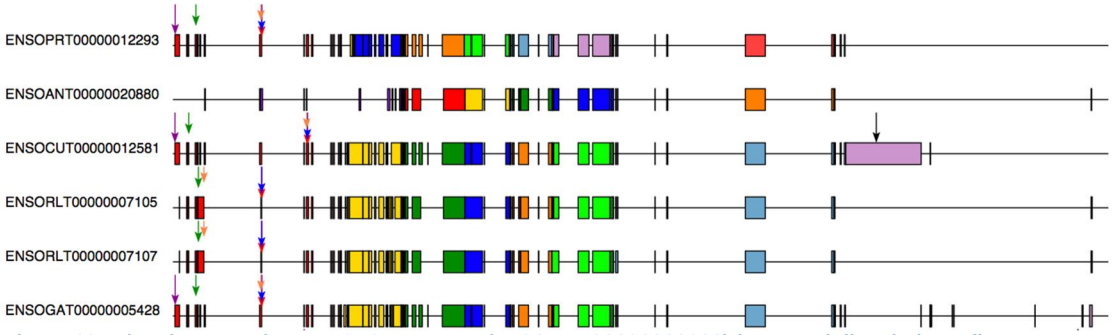

Figure 63 : The aberrant platypus CA6 sequence (ENSOANT00000020880) is potentially missing a first exon. An excerpt from the CA6 exon MSA schematic zoomed in from position 3900 to 9000. The colour key is as follows: purple: 5'UTR, red: exon 1, yellow: exon 2, green: exon 3, royal blue: exon 4, orange: exon 5, light green: exon 6, light blue: exon 7, light purple: exon 8, dark pink: exon 9.

Further investigation of the exon MSA schematics and the data within CAbase revealed that many transcripts contain exons that are short (11 residues or less). Using the following SQL query an exhaustive list was exported to Microsoft Excel and collated in Table 31. All CA groups have some short exons, with some of the short exons possibly being mispredicted while for others there is actual transcript evidence, such as for human CA12 and CA14 transcripts with TSL 1.

```
SELECT CA, EnsTranscriptId, Rank AS ExonNum, ABS(E.Start-E.End) + 1 AS
Bases, Common, G.Species, StartCodonLocation FROM Exons E, Transcripts T,
EnsSequences G, Species S WHERE G.GenomicEnsId=E.EnsParentId AND
G.GenomicEnsId=T.EnsParentId AND T.EnsTranscriptId = E.EnsTranscriptId
AND (ABS(E.Start-E.End) + 1) <33 AND G.Species=S.Scientific ORDER BY CA,
(G.Species IN ('Homo sapiens','Mus musculus')) DESC, G.Species,
T.EnsTranscriptId, Rank;
```

Furthermore, looking at the proportions of transcripts with short exons in Table 31, the membrane associated, secreted and two of the CARPs all have more than 30% of their transcripts with short exons.

Within CA9, CA12 and CA14 the most frequently listed short exons are the 2nd and the 9th exons. It is interesting to note that the 9th exon sits between the highly conserved catalytic domain coded by the exons 2 to 8 and the transmembrane region contained in exons 10 and 11. (Opavsky, et al., 1996), (Mori, et al., 1999), (Supuran, Scozzafava, & Conway, 2004). These short exons are easily visible in the conserved exon MSA schematics for these CAs. The common ancestor for these CAs, CA6, also has the 9th exon as a common short exon but does not have a common early short exon. This suggests that the short exon may have also been present in the early-jawed ancestor of the species exhibiting this phenomenon.

Table 31 : Collated list of the number of transcripts with at least on exon that was at most 11 residues long.

	Count of Transcripts containing Short Exon(s)	Count of Protein Coding Transcripts	% Short Exon Transcripts
CA1	12	65	18%
CA2	10	52	19%
CA3	11	67	16%
CA7	12	69	17%
CA13	9	53	17%
CA8	17	70	24%
CA10	41	89	46%
CA11	18	40	45%
CA5A	3	27	11%
CA5B	11	72	15%
CA4	35	117	30%
CA9	57	68	84%
CA12	56	76	74%
CA14	65	73	89%
CA15	14	41	34%
CA6	26	81	32%

Visualising alignments through the exon MSA schematics allows quick analysis of the transcripts no matter which criteria are chosen. It allows the user to see the general conservation patterns of exons in CAs, thus highlighting aberrant patterns that indicate suspicious gene models. This makes Exon_Analysis a useful quality control tool. Other features of the CAs, such as the signal and mitochondrial targeting peptides were also visualized using this method, but another application might be to look at the aligned locations of particular domains such as the GPI domain or perhaps seeing where specific amino acids are encoded. This tool could also be adapted to analyse other protein families.

6.2.1 Cytoplasmic Carbonic Anhydrases

The cytoplasmic conserved transcripts mostly have 7 exons in the coding region. The transcripts with anomalous numbers of exons were either non-protein coding or potentially mispredicted. Within human CAs there were many variant transcripts with unusual numbers of exons due to incomplete 3' or 5' regions, or retained introns.

These transcripts tended to have lower transcript support levels and lacked a CCDS code.

As expected, there are no reliably predicted signal peptides for any CA within this group. For these CAs a location was considered an indel if 90% of the other sequences in the PRANK alignment did not agree.

6.2.2 Mitochondrial Carbonic Anhydrases

Only CA5A and CA5B belong to this group. Like the cytoplasmic CAs, this group deviated recently from the ancestral CAs and was created following a duplication event. Both of these CAs have the predicted mitochondrial targeting peptide at the end of exon 1 and the CDS sequence is encoded in 7 exons. The locations of these peptides showed greater variation, but this could be due to the inherently lower reliability of predicting the locations of these peptides. (Emanuelsson, Brunak, von Heijne, & Nielsen, 2007) There were no transcripts with unusual domains within this group.

6.2.3 Secreted Carbonic Anhydrases

Of all the CAs studied in this thesis, only CA6 is secreted. The signal peptide was predicted to exist within the first exon for all except one transcript, the Flycatcher transcript, ENSFALT00000007711 where the cleavage site is at the start of exon 2. The locations of the predicted signal peptides lie within a few bases of each other in the PRANK MSA, thus demonstrating that this feature is conserved.

The criteria for finding indels was tightened to 95% of sequences in the alignment not having a gap/base in the same location instead of 90%, since the non-mammalian sequences have a pentraxin domain towards the 3' end of the sequence. The function of this domain is unknown at present.

6.2.4 Membrane Associated Carbonic Anhydrases

Five membrane associated CAs (CA IV, CA IX, CA XII, CA XIV and CA XV) were examined by the ExonAnalysis.py pipeline. All were found to have signal peptides mostly in the first exon or at the start of the second exon (Table 32). If the cleavage site was predicted for other exons, such as in CA4 with the transcript ENSGACT00000014885 where the signal peptide is in exon 4, the location of the cleavage site still aligns with the other cleavage sites predicted for exon one.

Table 32 : The number of transcripts with predicted signal peptides in the various exons for the membrane associated CAs.

	Number of transcripts with Predicted Signal Peptides in:				
CA	Exon 1	Exon 2	Exon 3	Exon 4	Exon 5
CA IV	27	34	2	1	2
CA IX	43	1	1		
CA XII	41	5			
CA XIV	16	31	1		
CA XV	15	9			

Mammalian CA9 has an extra proteoglycan domain at the start of the sequence. Therefore, like CA6, the criteria for finding CA9 indels was tightened to 95% of sequences in the alignment not having a gap/base in the same location instead of 90%. The function of this proteoglycan domain is unknown though it has been theorized by Opavsky to be involved with modulating cell interactions while Supuran suggests that it might be involved with cell adhesion and differentiation. (Opavsky, et al., 1996) (Supuran, Scozzafava, & Conway, 2004) Nearly all sequences in this CA group have the cleavage site for the signal peptide in exon one and the location lies within a few bases in the MSA between the sequences.

CA12, CA14 and CA15 do not have extra domains, consequently the criteria for selecting indels is set to 90% for other sequences not aligning with an indel base. CA15 is not expressed in primates, thus the reference sequence for this transcript comes from the mouse sequences. The locations of the predicted signal peptides for these CAs lie within a few bases of each other in the PRANK MSA, thus showing that this is also conserved.

6.2.5 Carbonic Anhydrase Related Proteins

Less than half of the CA10 and CA11 transcripts have predicted signal peptides that lie mostly at the start of exon 2. The predicted signal peptides for both isoforms all align at the same location in the exon MSA schematics. Both CA10 and CA11 have 9 exons in the coding region. The last coding exon in CA11 is often short.

Table 33 : The number of transcripts with predicted signal peptides in various exons for the CARPs.

	Number of transcripts with Predicted Signal Peptides in:				
CA	Exon 1	Exon 2	Exon 3	Exon 4	Exon 5
CA VIII					
CA X	37	2			
CA XI	8	10	3	1	1

CA8 has no predicted signal peptide and the coding region is made up of most commonly 8 exons (Figure 27). CA10 displays less conservation than CA8, especially in some mouse and zebrafish alternative transcripts (Figure 31). These transcripts are potentially mispredicted. CA11 only has one transcript (ENST00000596080) that is not homologous with the other CA11 transcripts, though it does have a low TLS of 3 (Figure 32).

6.2.6 Error Sources

The most critical error source is incorrect transfer of data from the source databases to CAbase. These potential sources of error have hopefully all been accounted for and error checking is complete. However, with new versions of API's, new data fields etc. the future owners of CAbase should periodically check for these sources of errors.

If there is some mismatch between the cDNA and protein sequences of a transcript, Exon_Analysis.py abandons drawing that particular transcript in the exon MSA schematic. This results in not all transcripts being included in the exon MSA schematics. In the future it may be desirable to include these transcripts irrespective

of this issue. For this situation the code would need to be modified to accommodate this.

The alignments of the CAs created by PRANK are very spread in the 5' and 3' regions of the non-coding sequences. To interpret these regions it is advisable to create an exon MSA schematic zoomed into the region of interest.

Selecting conserved sequences for further analysis through an automated procedure is prone to errors. Automated procedures only 'know' as much as the programmer has provided. The procedure used in this thesis could include and exclude mispredicted sequences, and the user needs to be aware of this issue. It is possible to add to the conditions of 'conservation assessment' in ExonAnalysis.py through the use of command line arguments. This could potentially remedy this situation so that the user can define the limits of what is conserved.

As Emanuelsson et al have noted in their paper, TargetP is prone to erroneously predicting mitochondrial targeting peptides since the motif is less conserved than the signal peptide motif. Therefore the predicted mitochondrial targeting peptide needs to be interpreted with some caution. (Emanuelsson, Brunak, von Heijne, & Nielsen, 2007)

6.3 Future Directions

Some potential future options for this tool include:

1. To calculate the locations of the PTX and the PG domains and draw this in the exon MSA schematics.
2. Add a few more user defined conditions for conservation so that the list of conserved transcripts are user defined
3. Allow users to only include transcripts with particular features e.g. only mammalian transcripts or perhaps only transcripts with short exons.
4. Develop ways to tag the suspect features in sequences.
5. Integrate data of various predictions and incomplete sequences to arrive at optimal gene models
6. Create reporting tools for inclusions of the improved gene models into global databases.

It would also be interesting to see the relationship between other features of the CA group in relation to their exons, such as the location of critical protein fold points or perhaps mapping the hydro-phobiscity on the exons.

After observing the entire alpha CA exon MSA schematics it was striking to see how conserved the exon lengths are within the coding regions of all of the CAs. It begs the question why is this so? Why don't more exons fuse or have intronic inserts? What mechanisms exist to set the splicing sites of the exons where they are? Why are some catalytic domains encoded in 9 exons while others require 11 exons? These questions are being addressed by researchers studying the carbonic anhydrases as well as many

other proteins and can be furthered by specialised databases and visualisation tools such as those presented here.

7 Conclusion

The goals of this thesis included the creation two pipelines: one to store the disparate CA data available on the internet in one specialized database named CAbase and one to visualize the exons in a scaled exon MSA schematic, named Exon_Analysis.py. The second tool facilitated the visual assessment of the quality of the predicted genes within Ensembl. It also showed that short exons that are at most 11 residues long are a conserved feature in CA6, CA9, CA12, CA14, CA10 and CA11. Visual inspection of the aligned locations of the predicted signal peptides demonstrated that they were conserved across species for each CA. The predicted mitochondrial targeting peptides showed less conservation, but this could be due to the lower reliability inherent in these predictions.

8 Works Cited

Aggarwal, M., Boone, C. D., Kondeti, B., & McKenna, R. (2012, October 3). Structural annotation of human carbonic anhydrases. *Journal of Enzyme Inhibition and Medicinal Chemistry*, 267-77.

Altschul, S., Gish, W., Miller, W., Myers, E., & Lipman, D. (1990). Basic local alignment search tool. *Journal of Molecular Biology*(215), 403-410.

Amazon. (2015). *Amazon Web Services*. Retrieved November 18, 2015, from Amazon Relational Database Service (RDS) Documentation: https://aws.amazon.com/documentation/rds/?icmpid=docs_menu

Amazon. (2015, October 15). *Amazon Web Services*. Retrieved October 15, 2015, from Amazon Web Services: http://aws.amazon.com/?icmpid=docs_menu

Aspatwar, A. (2014). *Distribution and Function of Carbonic Anhydrase Related Proteins (CARPs) VIII, X and XI.* Tampere: Tampere University Press.

Barker, H. R. (2013). 5. Methods. In H. R. Barker, *Development of a Protein Conservation Analysis Pipeline and apllication to Carbonic Anhydrase IV* (pp. 34-39). Tampere: Tampereen Yliopisto.

Bocchini, C., & McKusick, V. (2014, April 24). *OMIM - Online Mendelian Inheritance in Man*. Retrieved November 23, 2015, from 603179 Carbonic Anhydrase IX; CA9: http://www.omim.org/entry/603179?search=CA9&highlight=ca9

Breitling, R. (2005, October 20). *How to design a successful microarray experiment*. Retrieved October 19, 2013, from How to design a successful microarray experiment: http://www.brc.dcs.gla.ac.uk/~rb106x/microarray_tips.htm

Brinkman, R., Margaria, R., Meldrum, N. U., & Roughton, F. (1932, March). The CO2 catalyst present in blood. *Proceedings of the physiological Society*, 3-4.

Chandrashekar, J., Yarmolinsky, D., von Buchholtz, L., Oka, Y., Sly, W., Ryba, N. J., & Zuker, C. S. (2009, October 16). The Taste of Carbonation. *Science*, 443-445.

Cunningham, F., Amode, M., Barrell, D., Beal, K., Billis, K., Brent, S., . . . Flicek, P. (2015, January). Ensembl 2015. *Nucleic Acids Research, 43*, D662-9.

DePristo, M. A. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics, 43*, 491–498.

Dutta, S., & Goodsell, D. (2004, January). *RCSB BDP Protein Data Bank*. Retrieved November 21, 2015, from RSCB PDB-101 Carbonic Anhydrase: http://www.rcsb.org/pdb/101/motm.do?momID=49

Emanuelsson, O., Brunak, S., von Heijne, G., & Nielsen, H. (2007, April 19). Locating proteins in the cell using TargetP, SignalP and related tools. *Nature Protocols 2*, 953 - 971.

Emanuelsson, O., Nielsen, H., Brunak, S., & von Heijne, G. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *Journal of Molecular Biology*(300), 1005-1016.

Hassan, I., Shajee, B., Waheed, A., Ahmad, F., & Sly, W. S. (2012, April). Structure, function and applications of carbonic anhydrase isozymes. *Bioorganic & Medicinal Chemistry*, 1570-1582.

Hilvo, M., Tolvanen, M., Clark, A., Shen, B., Shah, G. N., Waheed, A., . . . Parkkila, S. (2005). Characterization of CA XV, a new GPI-anchored form of carbonic anhydrase. *Journal of Biochemistry*, 83-92.

Lowe, N., Edwards, Y., Edwards, M., & PH, B. (1991, August). Physical mapping of the human carbonic anhydrase gene cluster on chromosome 8. *Genomics*, 882-8.

Löytynoja, A., & Goldman, N. (2008, October 7). A model of evolution and structure for multiple sequence alignment. *Philosophical Transactions of the Royal Society B*, 3913-3919.

Löytynoja, A., Vilella, A., & Goldman, N. (2012). Accurate extension of multiple sequence alignments using aphylogeny-aware graph algorithm. *Bioinformatics*(28), 1684-1691.

Morgan, P., Pastorekova, S., Stuart-Tilley, A., Alper, S., & Casey, J. (2007, May). Interactions of transmembrane carbonic anhydrase, CAIX, with bicarbonate transporters. *Cell Physiology*, 738-748.

Mori, K., Ogawa, Y., Ebihara, K., Tamura, N., Tahiro, K., Kuwahara, T., . . . Nakao, K. (1999, May). Isolation and Characterization of CA XIV, a Novel Membrane-bound Carbonic Anhydrase from Mouse Kidney. *The Journal of Biological Chemistry*, 15701-15705.

Opavsky, R., Pastorekova, S., Zelnik, V., Gibadulinova, A., Stanbridge, E. J., Zavada, J., . . . Pastorek, J. (1996, February). Human MN/CA9 Gene, a Novel Member of the Carbonic Anhydrase Family: Structure and Exon to Protein Domain Rlationships. *Genomics*, 480-487.

Oracle Corporation. (2015). *MySQL*. Retrieved November 18, 2015, from MySQL Documentation: http://dev.mysql.com/doc/refman/5.7/en/introduction.html

Patrikainen, M. (2012). *Pentraxin- Carbonic Anhydrase CA VI: A novel multidomain protein.* Tampere: Tampere University Press.

Petersen, T. N., Brunak, S., von Heijne, G., & Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods*(8), 785-786.

Python Software Foundation. (n.d.). *Python* . Retrieved February 18, 2013, from Python 2.7 Documentation: https://docs.python.org/2/index.html

Sievers, F., Wilm, A., Dineen, D., TJ, G., Karplus, K., Li, W., . . . Higgins, D. (2011, January 1). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 539.

Supuran, C., Scozzafava, A., & Conway, J. (2004). *Carbonic Anhydrase: Its inhibitors and Activators* (Vol. 1). CRC Press.

Suyama, M., Torrents, D., & Bork, P. (2006, April). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, 609-612.

Swenson, E. R., Leatham, K. L., Roach, R. C., Schoene, R. B., Mills, W. J., & Hackett, P. H. (1991, December). Renal carbonic anhydrase inhibition reduces high altitude sleep periodic breathin. *Respiration Physiology*, 333-343.

Tashian, R. E. (1989, June). The Carbonic Anhydrases: Widening Perspectives on Their Evolution, Expression and Function. *BioEssays, 10*(6), 186-92.

Tolvanen, M. (2013). *Pseudogenes and Gene Duplications Tell a Story of Evolutionary Fates of Animal Alpha Carbonic Anhydrases.* Tampere: Tampere University Press.

Tolvanen, M. E., Ortutay, C., Barker, H. R., Aspatwar, A., Patrikainen, M., & Parkkila, S. (2012). Analysis of evolution of carbonic anhydrases IV and XV reveals a rich history of gene duplications and a new group of isozymes. *Bioorganic & Medicinal Chemistry*, 1503-1510.

Wikipedia. (2015, November 17). *Wikipedia, The free encyclopedia*. Retrieved November 18, 2015, from SQL: https://en.wikipedia.org/wiki/SQL

Yates, A., Beal, K., Keenan, S., McLaren, W., Pignatelli, M., Ritchie, G., . . . Flicek, P. (2015, Jan 31). The Ensembl REST API: Ensembl Data for Any Language. *Bioinformatics, 1*, 143-145.

9 Appendix A

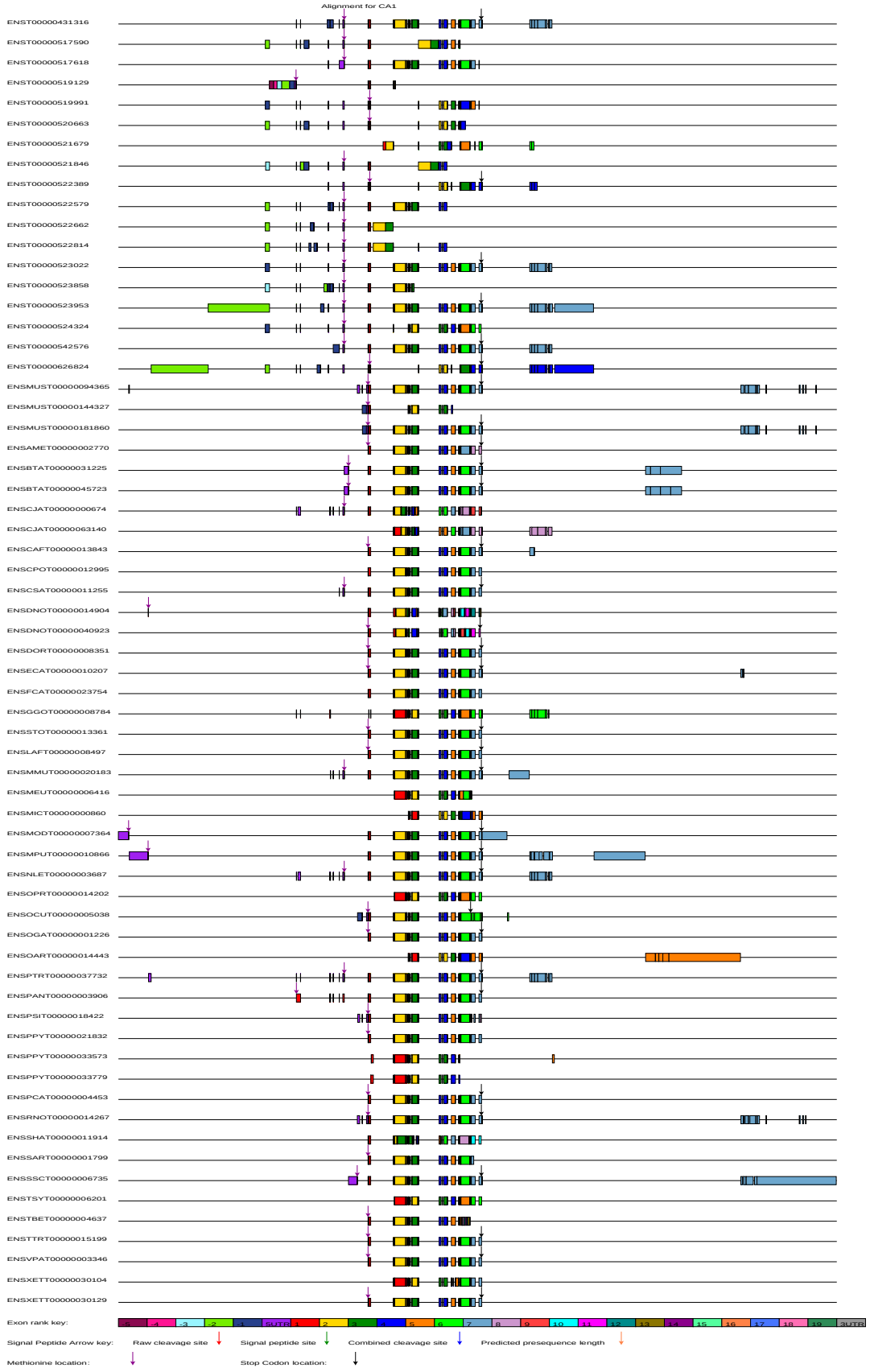

Figure 64 : The exon MSA schematic of all of the CA1 protein coding cDNA transcripts from the PRANK MSA position 0 to the end.

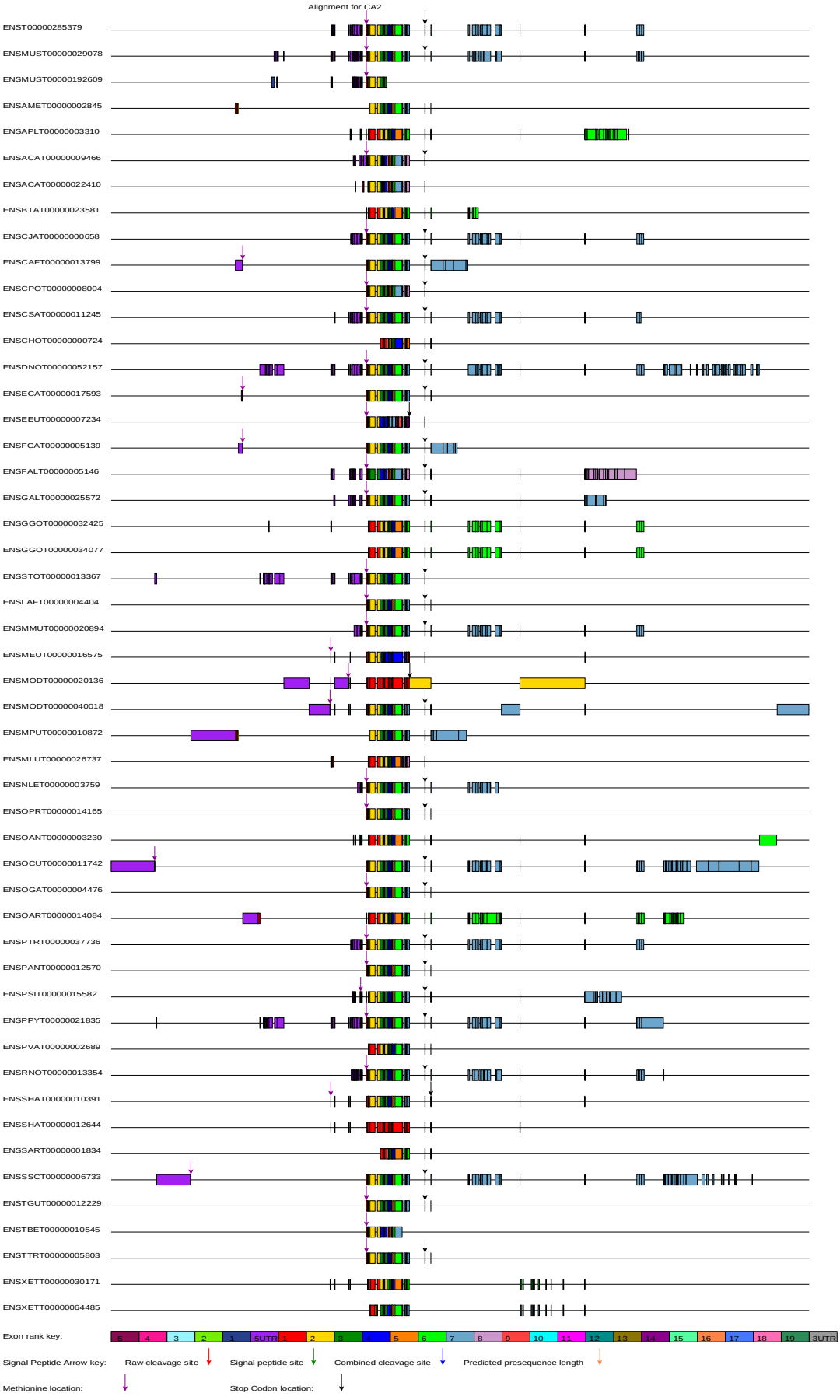

Figure 65 : The exon MSA schematic of all of the protein coding cDNA transcripts of CA2 from the PRANK MSA position 0 to the end.

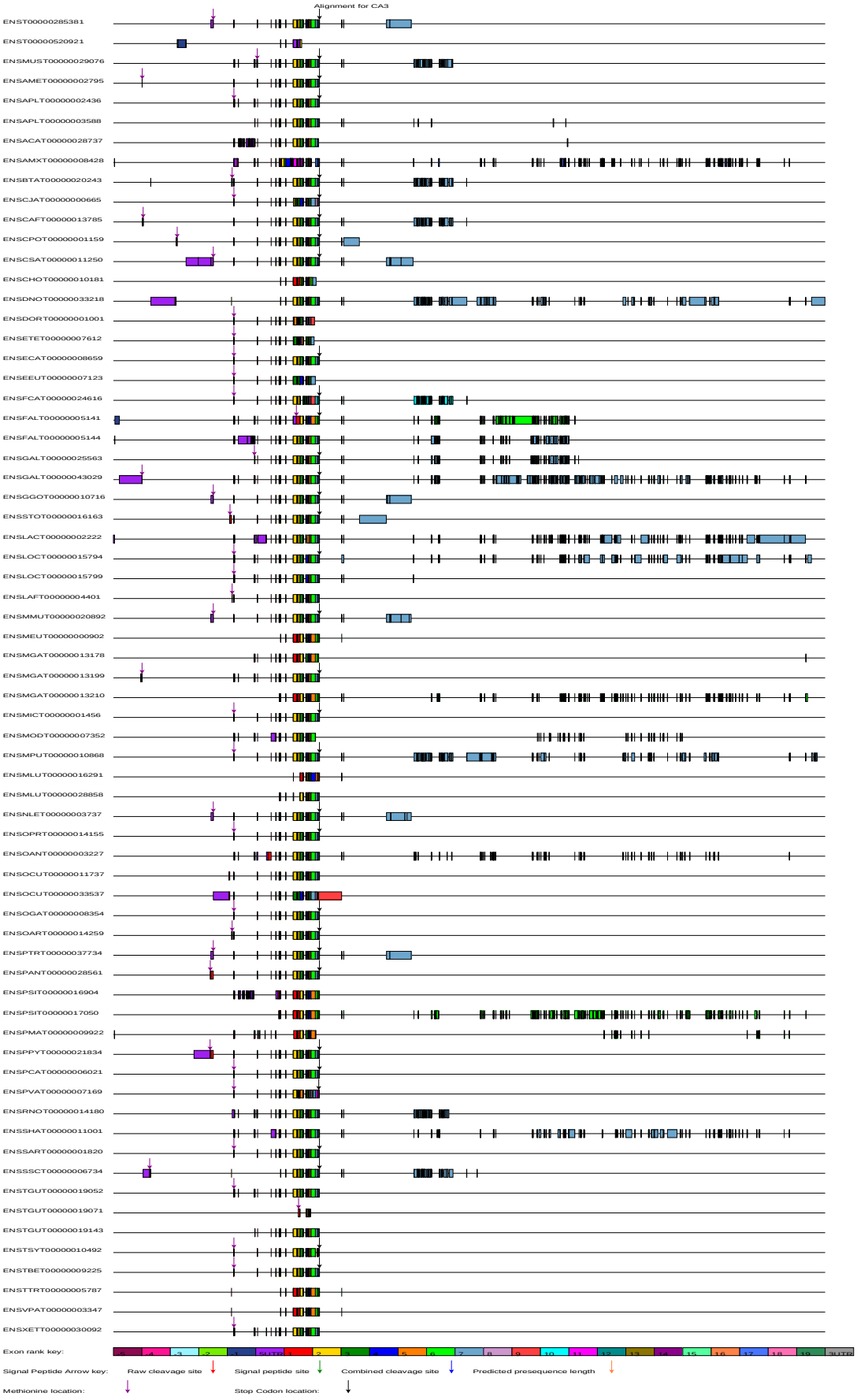

Figure 66 : The exon schematic of all of the protein coding cDNA transcripts of CA3 from the PRANK MSA position 0 to the end.

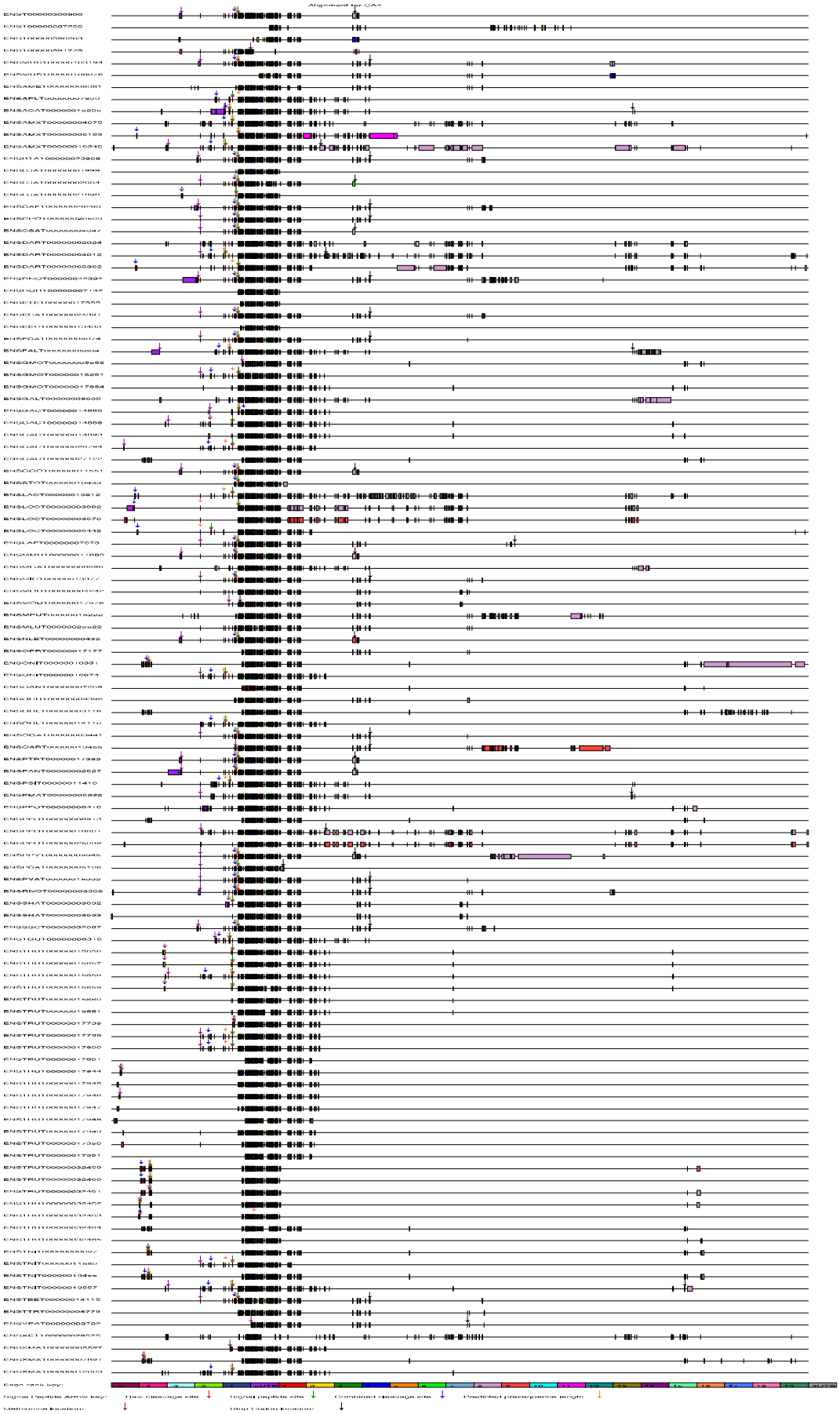

Figure 67 : The exon schematic of all of the protein coding cDNA transcripts of CA4 from the PRANK MSA position 0 to the end.

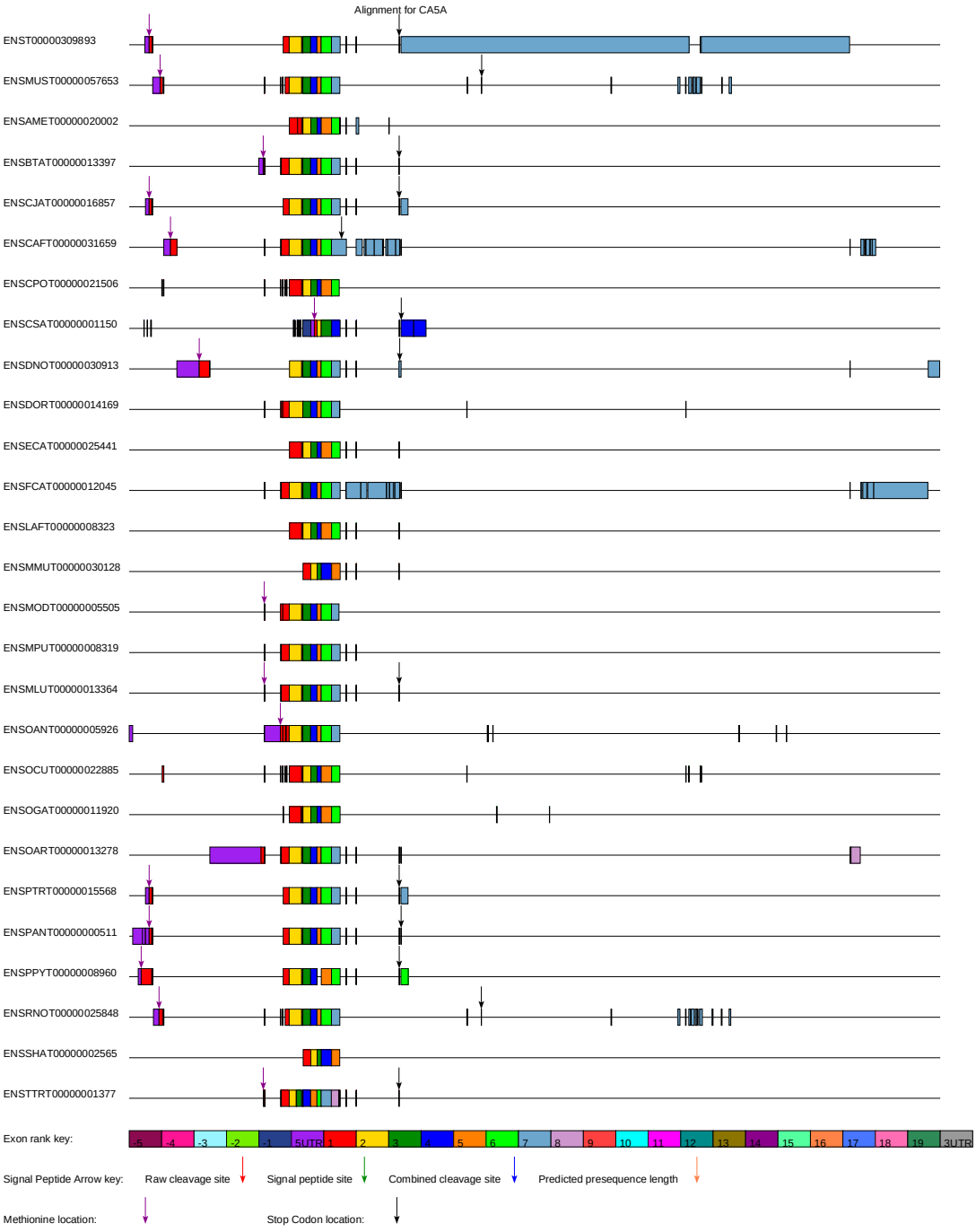

Figure 68 : The exon schematic of all of the protein coding cDNA transcripts of CA5A from the PRANK MSA position 0 to the end.

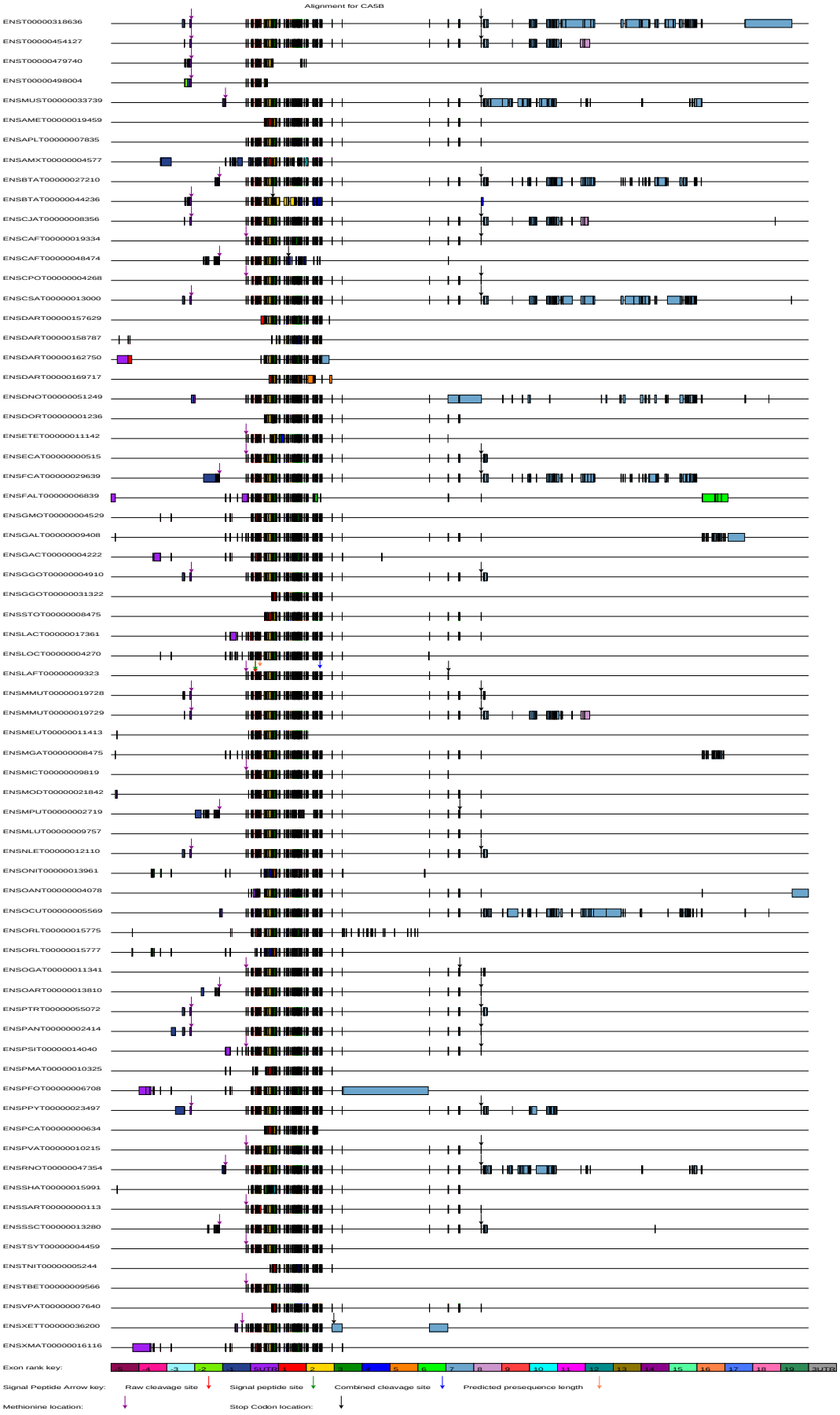

Figure 69 : The exon schematic of all of the protein coding cDNA transcripts of CA5B from the PRANK MSA position 0 to the end.

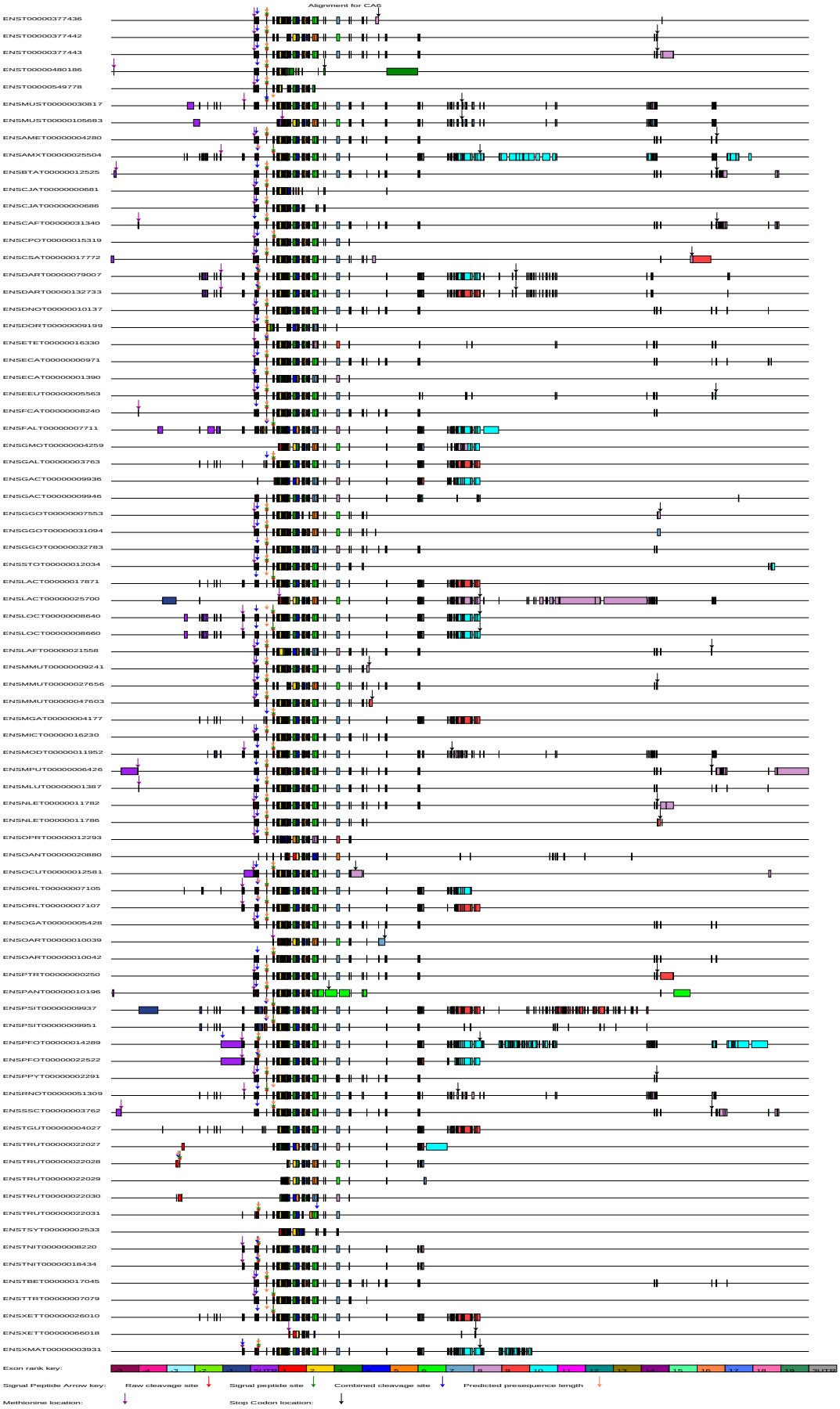

Figure 70 : The exon schematic of all of the protein coding cDNA transcripts of CA6 from the PRANK MSA position 0 to the end.

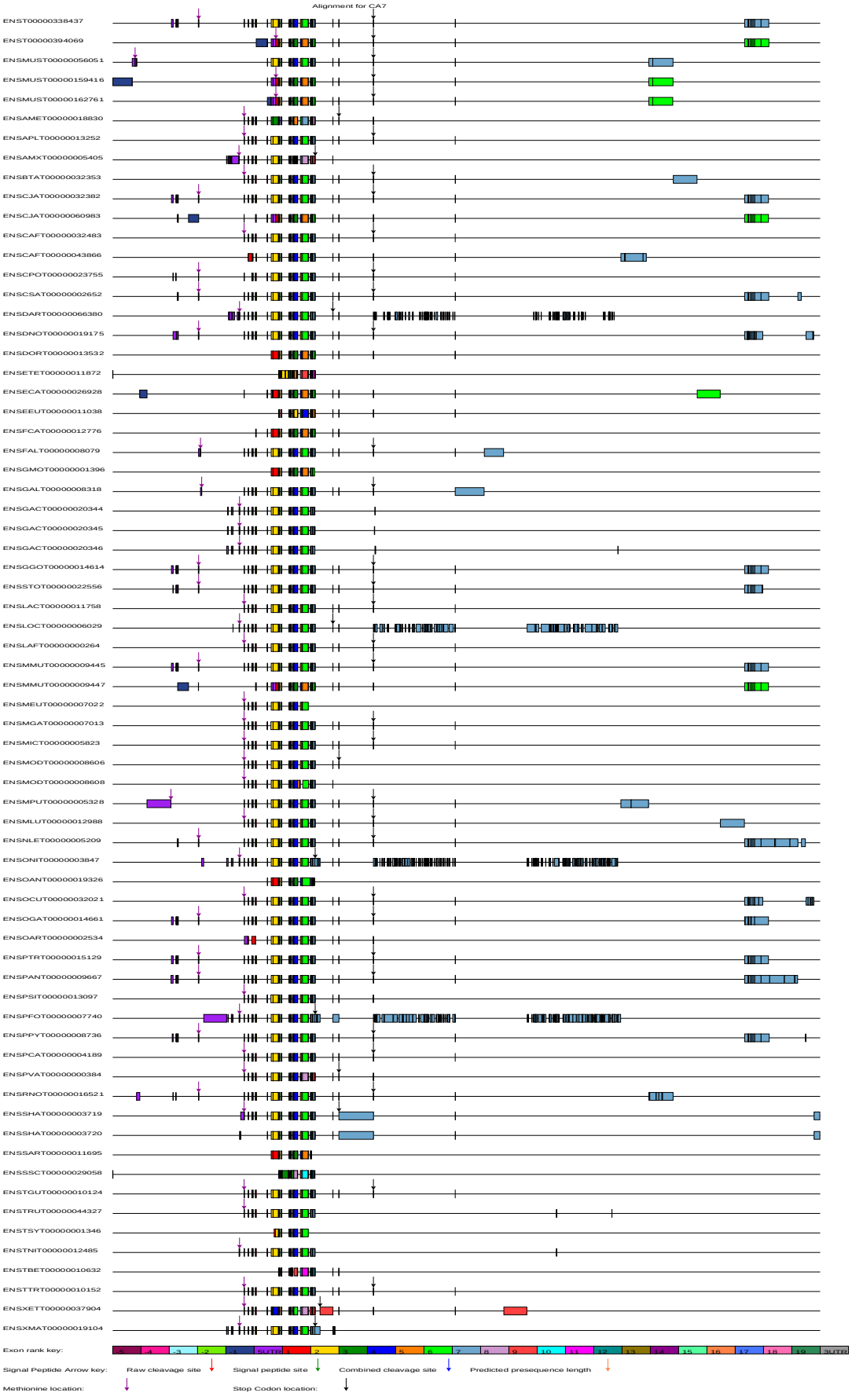

Figure 71 : The exon schematic of all of the protein coding cDNA transcripts of CA7 from the PRANK MSA position 0 to the end.

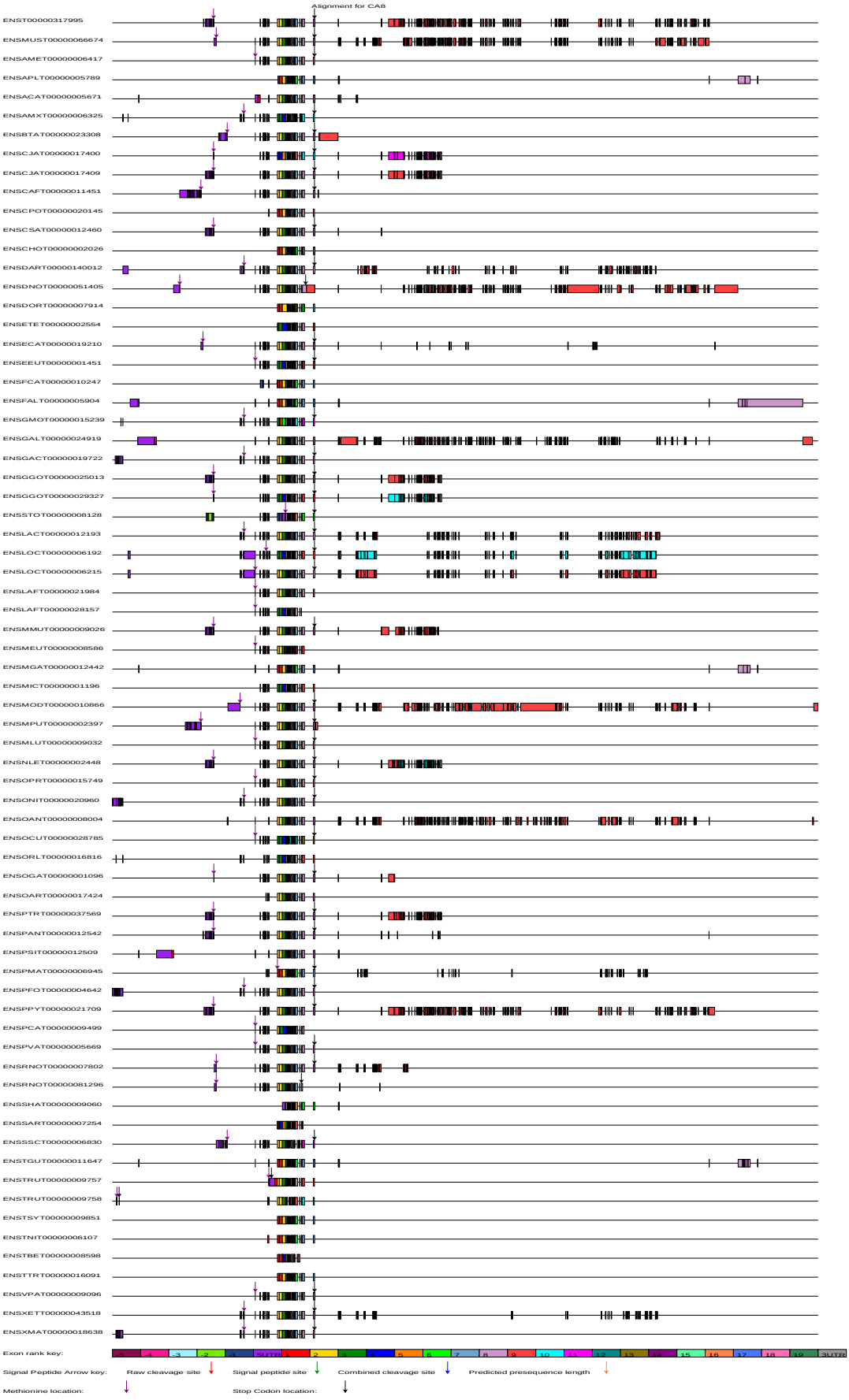

Figure 72 : The exon schematic of all of the protein coding cDNA transcripts of CA8 from the PRANK MSA position 0 to the end.

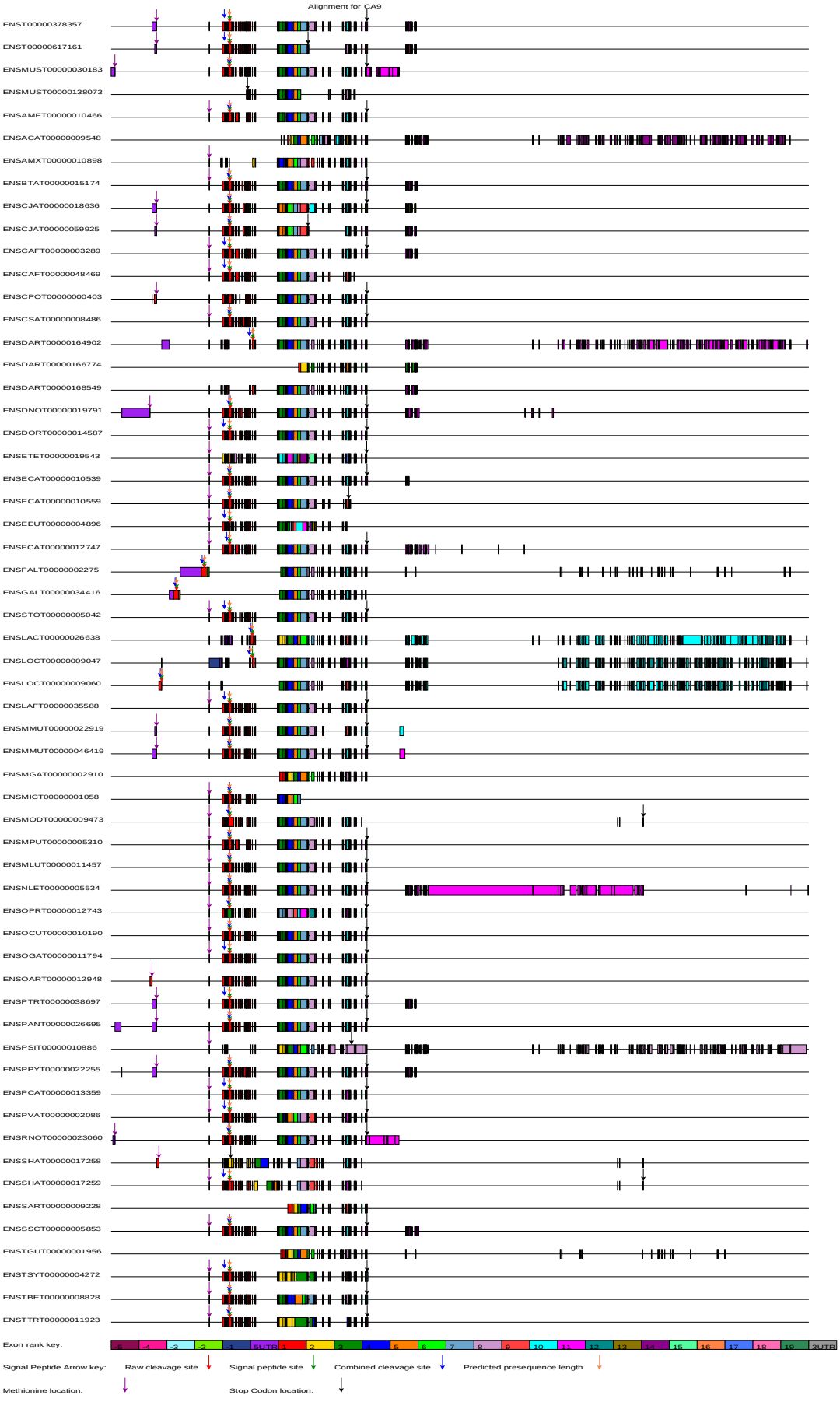

Figure 73 : The exon schematic of all of the protein coding cDNA transcripts of CA9 from the PRANK MSA position 0 to the end.

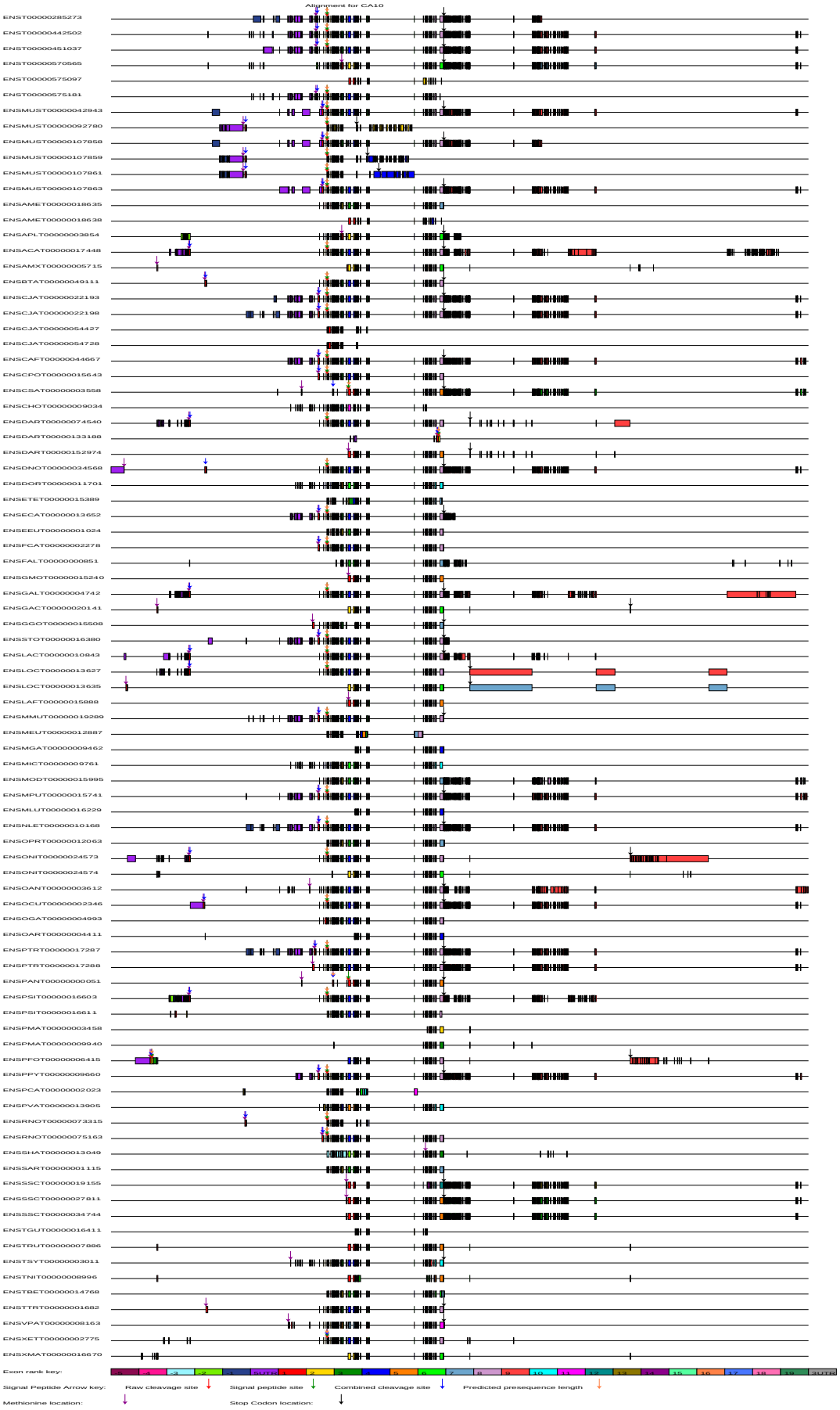

Figure 74 : The exon schematic of all of the protein coding cDNA transcripts of CA10 from the PRANK MSA position 0 to the end.

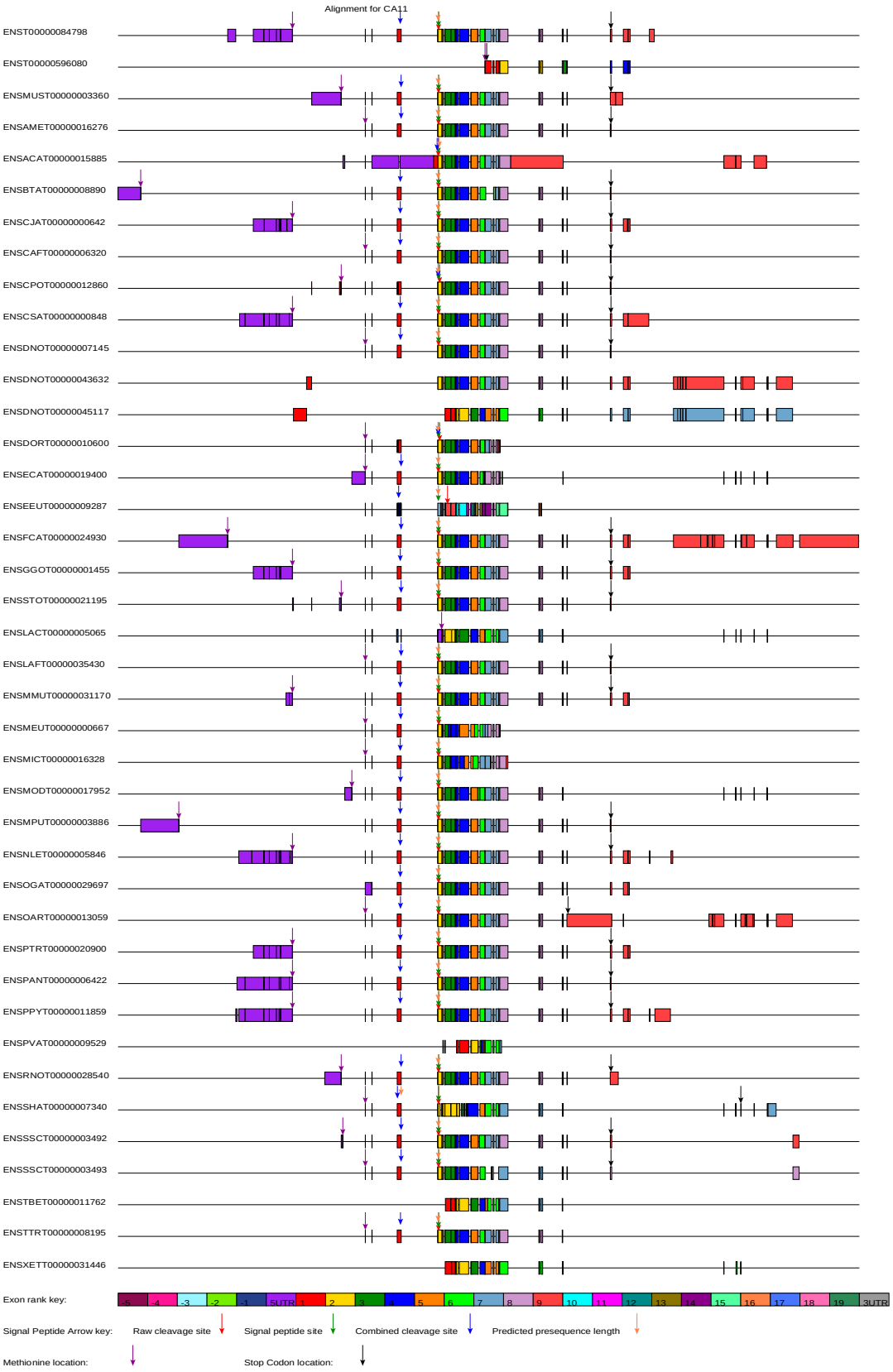

Figure 75 : The exon schematic of all of the protein coding cDNA transcripts of CA11 from the PRANK MSA position 0 to the end.

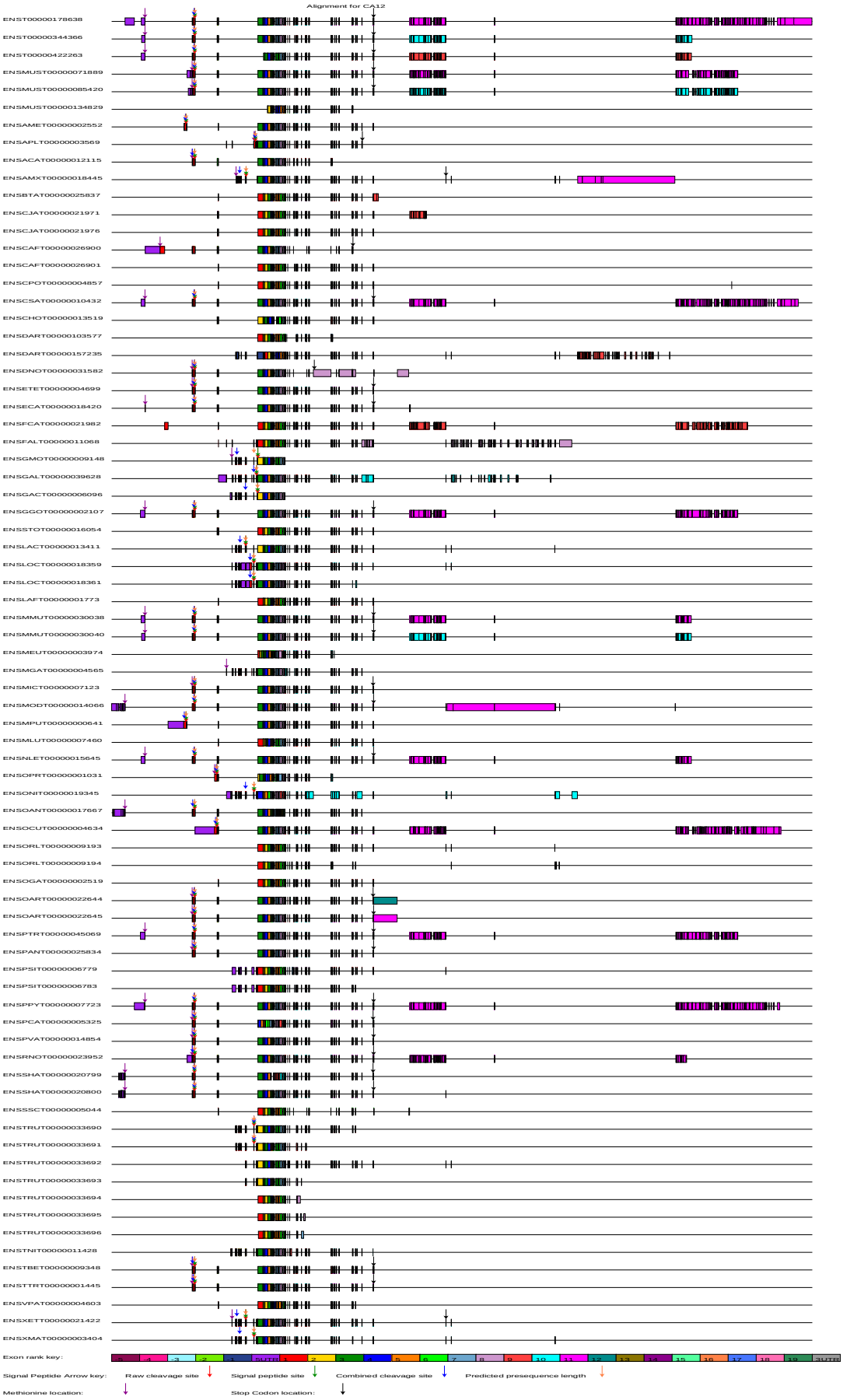

Figure 76 : The exon schematic of all of the protein coding cDNA transcripts of CA12 from the PRANK MSA position 0 to the end.

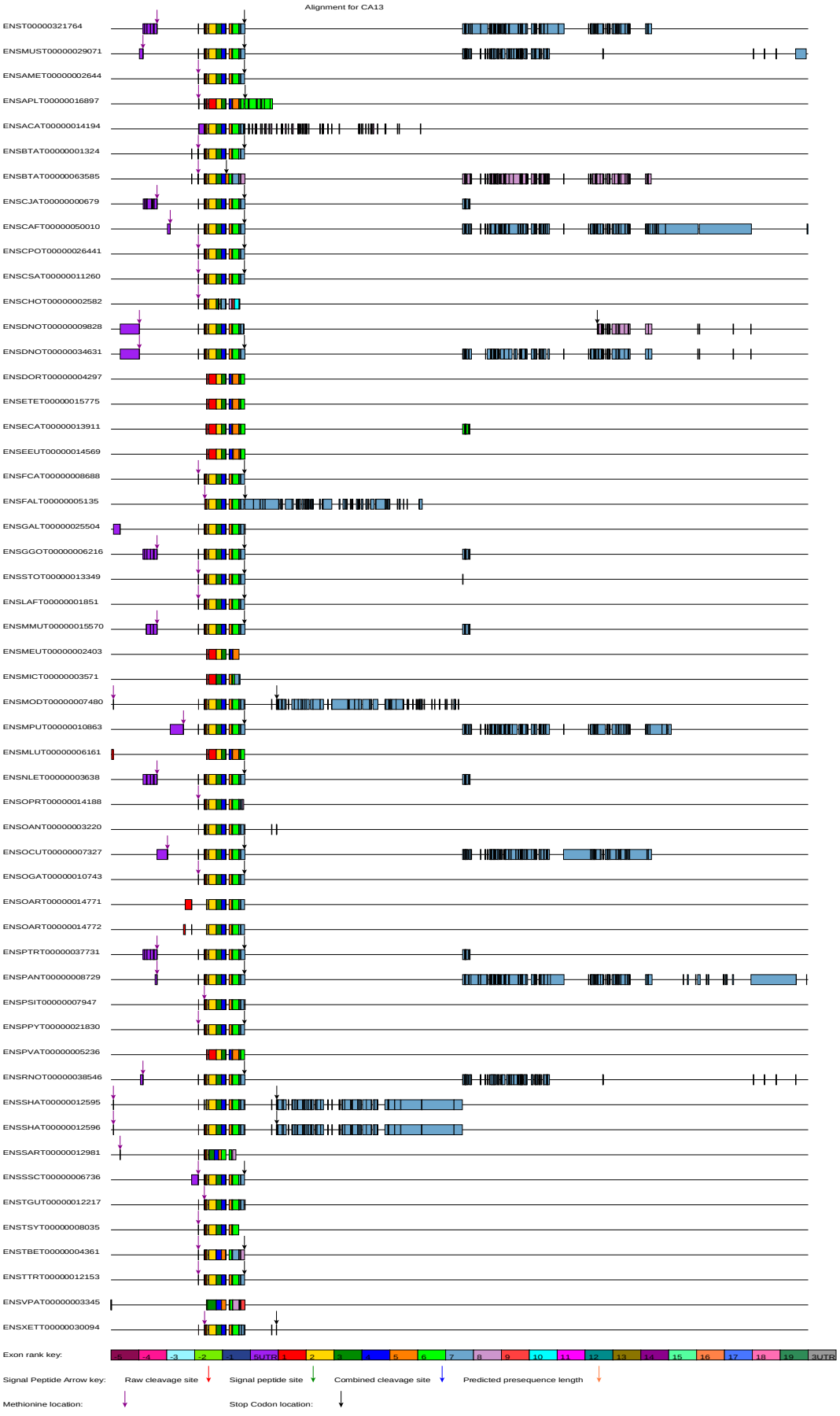

Figure 77 : The exon schematic of all of the protein coding cDNA transcripts of CA13 from the PRANK MSA position 0 to the end.

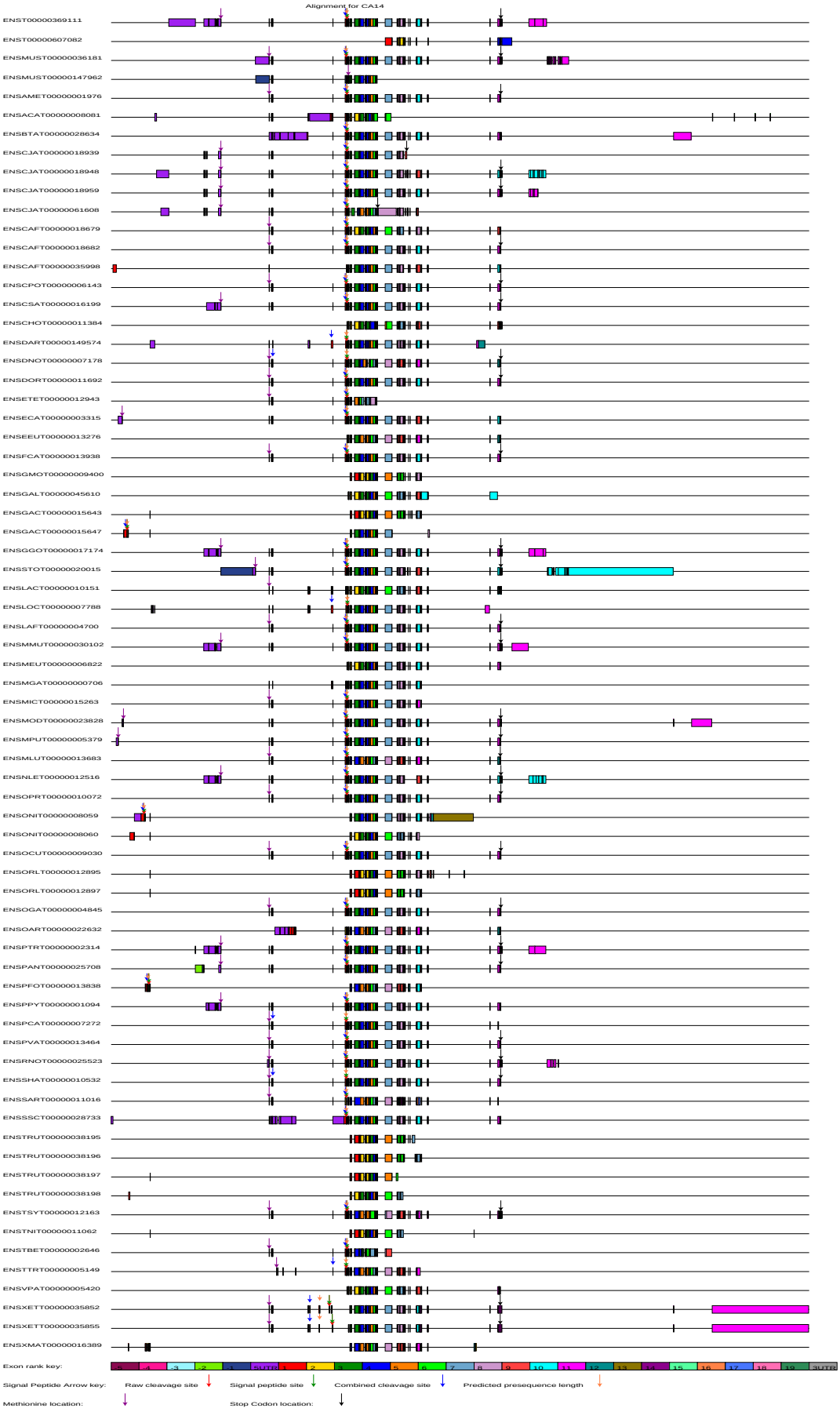

Figure 78 : The exon schematic of all of the protein coding cDNA transcripts of CA14 from the PRANK MSA position 0 to the end.

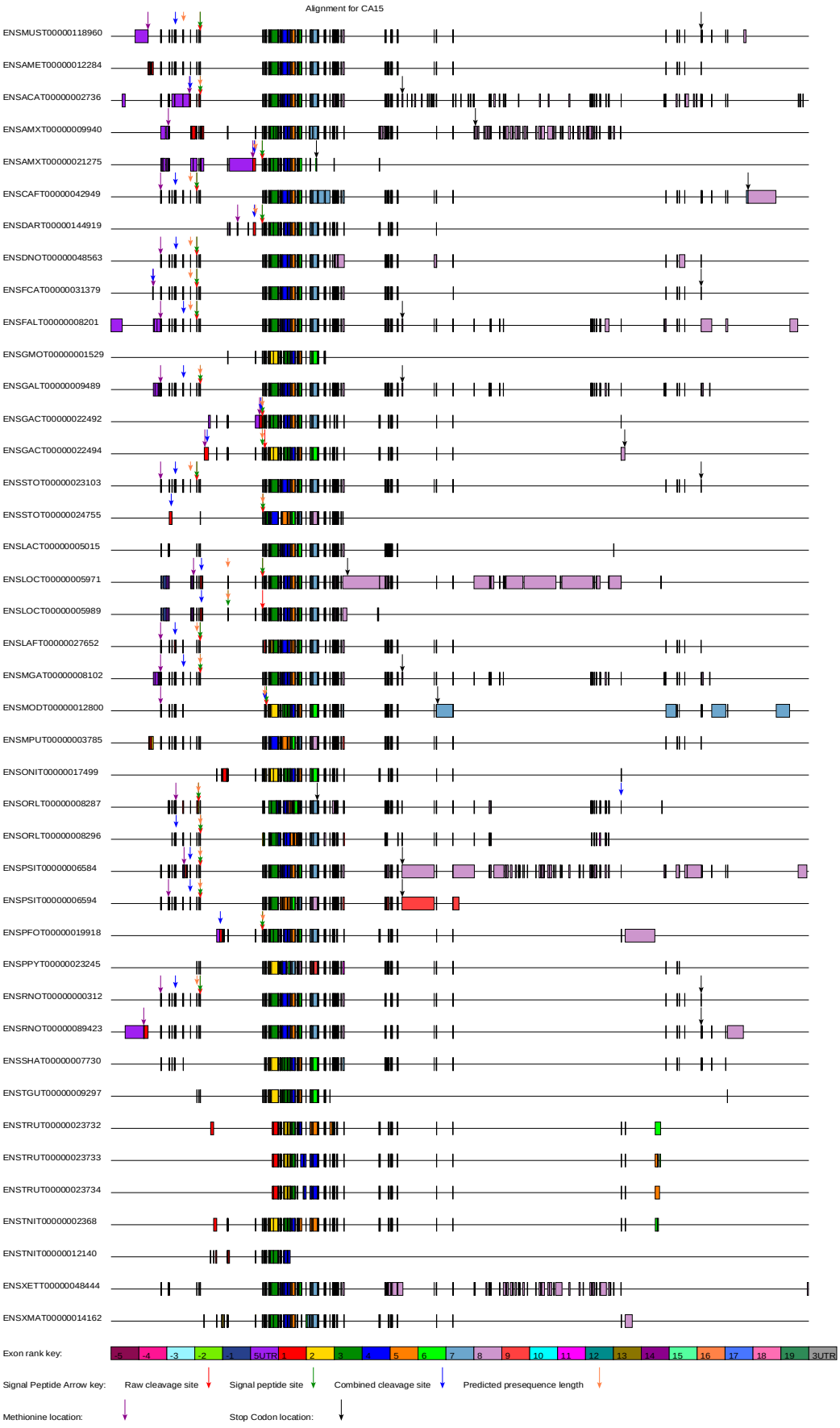

Figure 79 : The exon schematic of all of the protein coding cDNA transcripts of CA15 from the PRANK MSA position 0 to the end.