EXON ANALYSIS WITH CABASE, A NOVEL CARBONIC ANHYDRASES DATABASE TOOL

UNIVERSITY OF TURKU
Department of Future Technologies / Faculty of Science and Engineering
Master of Science in Technology Thesis
Master's Degree Programme in Digital Health and Life Sciences
December 2020
Sebastián Ignacio Zúñiga Norman

UNIVERSITY OF TURKU

Department of Future Technologies/Faculty of Science and Engineering

ZÚÑIGA NORMAN, SEBASTIÁN IGNACIO: Exon analysis with CAbase, a novel Carbonic Anhydrases database tool.

Master of Science in Technology Thesis, 64 p., 14 app. p.

Master's Degree program in Digital Health and Life Sciences

December 2020

The Carbonic Anhydrases are a superfamily of proteins, a type of metalloenzyme that are found ubiquitously throughout the most diverse clades in the tree of life. Since first identified as a catalyst for the reversible hydration and dehydration of carbon dioxide almost 90 years ago, the interest in research for its pharmacological and industrial applications is constantly growing and counting with tools like the one here discussed becomes prevailing.

When assembly errors or gaps are in exonic regions, misses or misplacing exons occur and it is because of these errors that approaches like *homology-based gene models* fail due to the lack of correct data. Instead, a similar length of poorly matching sequence may be predicted. Performing exon analysis for the carbonic anhydrases in the context of CAbase aims to identify the mis predicted exons and refers to a set of actions that are possible to perform over them, be they computationally or humanly conducted. Analysis like the comparison of any relevant exon to a desired prototype, the curated addition of labels to them, indicating different important features, or simply, a visual guide to understand their distribution in transcripts or genes of interest.

CAbase is a web-based Third Party Annotation System (TPA). A toolkit that allows to browse, annotate and perform analysis over sets of data related to the Carbonic Anhydrases protein family. CAbase is an effort to centralise, normalise and analyse information and annotations over the CA family.

To date, 47 CAs (14 Human) genes have been incorporated and currently, more than 50 thousand exons are possible to be characterized based on their splicing site as well as analyse their predicted translations and alignment to the precursor homolog.

CAbase continues to be developed, more improvements are included and in conjunction with the usage by the experts in the field of CAs, this *inferal* type of TPA, could become a central tool in the CA research community, creating better consensus data on this ever-growing protein family.

Keywords: *Carbonic Anhydrases, Exon Analysis, Bioinformatics online tool, Third party Annotation system, Genome annotation, Protein family database*.

TABLE OF CONTENTS

Abbreviations and Acronyms

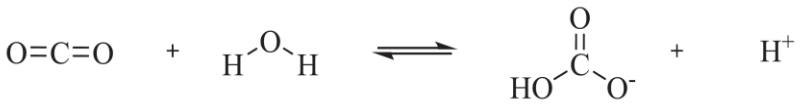
AS	ALTERNATIVE SPLICING
API	APPLICATION PROGRAMMING INTERFACE
CA	CARBONIC ANHYDRASE
CAHZ	CARBONIC ANHYDRASE FOR ZEBRAFISH
CARP	CARBONIC ANHYDRASE-RELATED PROTEIN
CDS	CODING DNA SEQUENCE
GFF	GENERIC FEATURE FORMAT
GPI	GLYCOSYLPHOSPHATIDYLINOSITOL
HCA	HUMAN CARBONIC ANHYDRASE
MRNA	MESSENGER-RNA
MSA	MULTI SEQUENCE ALIGNMENT
MVC	MODEL VIEW CONTROLLER
ORF	OPENING READING FRAME
RDBMS	RELATIONAL DATA BASE MANAGEMENT SYSTEM
UI/UX	USER INTERFACE / USER EXPERIENCE
TPA	THIRD PARTY ANNOTATION [DATABASE]

To Gabriel and Vanessa,
for their infinite support and love.
To Anna,
whose support made it possible, jag älskar dig.
For Nikolas,
with whom I hope one day share the passion for science.
Thanks to Dr. Martti Tolvanen,
mentor and friend, for his impossible care, dedication and passion.

Chapter 1

INTRODUCTION, THE CARBONIC ANHYDRASES

Carbonic Anhydrases are a superfamily of proteins, a type of metalloenzymes that catalyses the reversible reaction of carbon dioxide hydration to bicarbonate and proton. To this date, 8 genetic families have been identified showing the convergent evolution present throughout all life forms. All life kingdoms use this catalysation for managements of high levels if CO_2, pH regulation and other important processes. First discovered 87 years ago, in 1933 from erythrocytes in human blood (Brinkman R et al., 1932), Carbonic Anhydrases have been at the front of research, given its capacities to catalyse a reaction in one of the most abundant carbon forms, a core molecule to all life on Earth.



Equation 1–1: The reversible reaction of carbon dioxide into bicarbonate and proton.

While the carbon dioxide hydration reaction is very effective at pH values above 12, it is particularly slow at a pH value of 7.5 and lower, which usually marks many tissues and organisms.

1.1. The α-CA family

Whilst all the CAs from the eight currently identified families (α-, β-, γ-, δ-, ζ-, η-, θ- and ι-CAs)(Alterio et al., 2012), have in common the CO_2 hydration catalysation, with the catalytically active isoforms conformed by a hydroxyl molecule bound to a metal ion, these families present low sequence similarities and different three-dimensional structures among each other. While from delta- to iota-CAs are linked to more specific and simpler forms of life, gamma-CAs are related to bacteria and plants and are probably an earlier form (Smith et al., 1999), and beta-CAs appear in all organisms groups except from vertebrates (Smith & Ferry, 2000). However, it is the alpha-CA family of our interest, the one present in vertebrates, also present in protozoa, algae corals bacteria and the cytoplasm of green plants.

This family is present among many living organisms, rendering it as the most populous family of CAs. Within the α-CA family there are differences in their cellular location, oligomeric arrangement as well as their catalytic activity, nonetheless all of them have Zn (Zinc) as their metal ion found in the active site together with three histidine residues and a hydroxide ion as part of their fundamental structure.

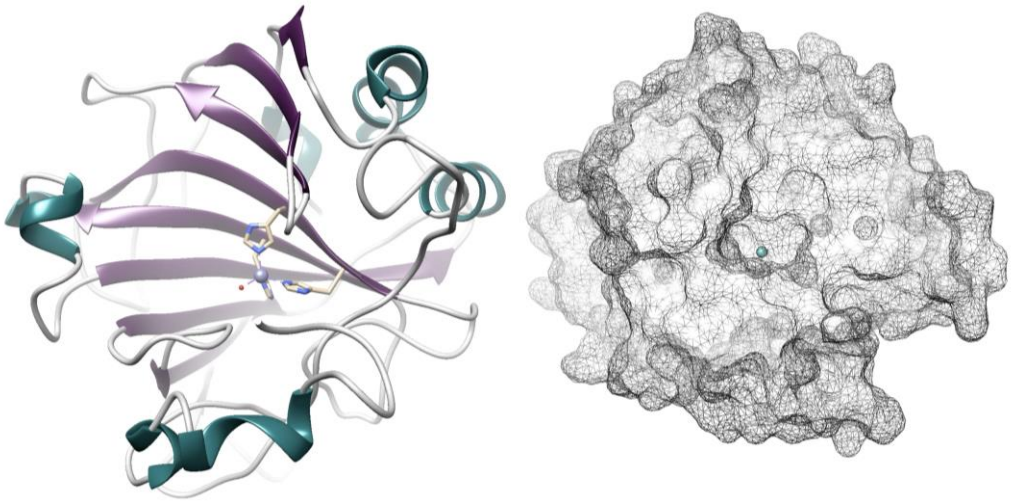


Figure 1-1 Human CA II. Secondary structure (left) showing the coordinating histidine residues as *blue and beige sticks* the Zn *light purple sphere*. *(Right)* the surface representation of hCA II, at the centre, the Zn^{2+} ion (*cyan sphere)* in the catalytic pocket. PDB accession ID: *1FW2*. Rendered using UCSF *Chimera 1.14* (Pettersen et al., 2004)

Different $\boldsymbol{\alpha}$-CAs (CA, here onwards) isozymes have been identified in mammals and in general three distinct groups of CA can be outlined. One group can be addressed as the *cytoplasmic* CAs, another the *membrane-bound* and a third group of rather interesting members, called *Carbonic Anhydrase-related proteins* (CARP). The first group include the mammalian CA I, II, III, V, VII and XIII found in the cytoplasm, with the exception of the mitochondrial confined CAs V namely, CA VA and CA VB. The second, consist of mammalian CAs IV, IX, XII, XIV and XV. Lastly, the CARP group is composed by CAs VIII, X and XI, this CAs have lost classical CA activity as they have lost one or more of the histidine residues that coordinate the zinc ion in the catalytic pocket and their roles have not been completely understood to date. Yet, while CARP VIII has been found to be associated with motor coordination in human and mouse, CARP XI has been associated with several cancers and CARP X the precursor for CARP XI, suggests a regulation in the circadian cycles due to their night/day expression likelihood (Aspatwar et al., 2013).

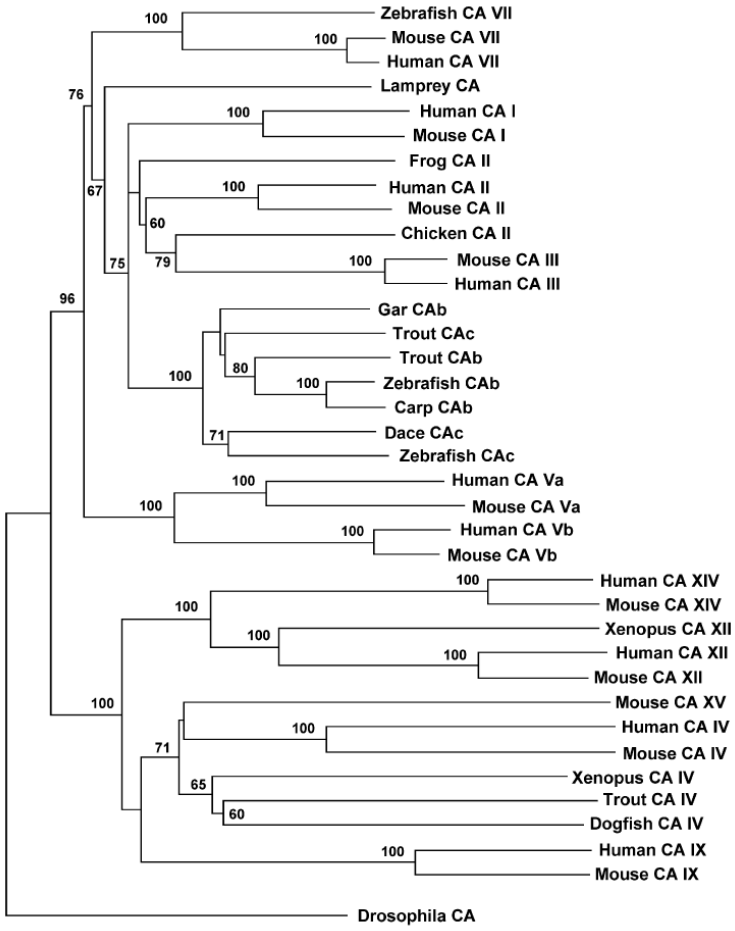


Figure 1-2: Phylogenetic analysis for both cytoplasmic and membrane bound CAs. This phylogenetic tree was constructed using neighbour joining by using parsimony analysis with support for nodes assessed using bootstrap analysis, and ordered using drosophila CA as an outgroup. Fig. 1A in (Aspatwar et al., 2013).Open access.

1.2. Human CAs

hCA isoforms play an important role in many physiological functions, their normal or abnormal catalytic activity has been linked to diseases, epilepsies, obesity, altitude sickness and glaucoma (Supuran & De Simone, 2015) as well as the participation in some cancer progression due to alternative splicing. Malentacchi et. al, described how CA IX isoform, when in lack of exons 8 and 9 (AS) influences cell proliferation and tumour progression by regulation of pH under hypoxic conditions, becoming a urinary marker for bladder cancer(Malentacchi et al., 2012). Indicating the clear relevance of understanding different splicing events for the CA family.

Special attention can be put on the case of CA XV, although characterized for other mammalians it was not found expressed in humans nor chimpanzees (Hilvo et al., 2005), this study pointed out that several similarities with CA IV, therefore being a GPI-anchored membrane protein, an important structural feature. Until, its expression in humans was found in the thick ascending limb of Henle (Saari et al., 2010). These studies have opened new possibilities for gene ancestry analyses, grouping the most recently added hCA isozyme with the known CA IV, revealing yet another GPI-linked CA, CA XVII, while lost in mammals it is found in fishes as multiple isozymes as well as a single isozyme in non-fish vertebrates (Tolvanen et al., 2013).

Figure 1-3 MSA for catalytically active hCA isoforms, depicting the Histidines coordinating the Zn ion in cyan. Glu106, and Thr199 in grey, while in red, the Histidine proton shuttle. MSA produced with Mview (Brown et al., 1998).

1.3. Non-Vertebrate CAs

CAs in fishes are studied with great emphasis, due to the abundant presence of CAs in their gills, they follow a non-trivial homologue relationship to their tetrapod counterparts, particularly, in the case of the cytoplasmic CAs, in a 2-to-4 orthologue relationship. For CAs I, II, III and XIII of tetrapod, *cahz* and *ca2* can be found. These two genes were part of a duplication process that took place only in teleost fishes, the vast majority of species under bony fishes, 96% of all fishes (Datovo & Vari, 2013). They have been proposed to be grouped under the clade called ca17a and ca17b (Ferreira-Martins et al., 2016). This, with the caveat that a CA17 name had already been proposed by Tolvanen et al., 2013. Although, to this date, this naming problem was brought to a discussion by the authors, it shed a light in naming conventions and a need to structurally classify the newly sequenced and discovered CAs.

1.4. Latest aditions to the CA superfamily

The next two classes of CAs with proven activity have been recently characterized, $\boldsymbol{\theta}$- and ι-CA in 2016 and 2019 respectively:

1.4.1 The $\boldsymbol{\theta}$-CA

First described after identifying a novel CA in the lumen of pyrenoid-penetrating thylakoid of the diatom *Phaeodactylum tricornutum,* a photosynthetic aquatic microeukaryote, showed no significant sequence homology with previously characterized CAs, placing it in its own family group (Datovo & Vari, 2013). This protein named Pt43233, have been recognized as an important player in the CO2-concentrating mechanism (CCM). Thus, understanding the role of these enzymes in the CCM could offer directions for improving photosynthetic yields with genetic engineering.

The crystallography structure of Pt43234, a homologous protein from *P. tricornutum* was solved (Jin et al., 2016), showing high sequence identity in a specific region with Pt43233, these two monomers showed $\boldsymbol{\beta}$-fold presenting the typical features of $\boldsymbol{\beta}$-CAs. Structurally, certain similarities can be drawn against the $\boldsymbol{\beta}$-CA class, the general fold and the amino acids that make the metal-binding site are similar, yet, the

active-site is built differently, since the residues come from different parts of the sequence.

```
          1[        .         .         .         .         :         .         .         .80
Pt43233     FETTMAKVQQAFTGAATNDQLVAKTKSALSRFGFGS-NSLVATSFCSDEVNRPLETDFAKEFKDTFSLGGLAGFPFSGV
Pt43234     MNVTLTAVKKAFPDALTNAELVAMVSKRLSQFGYHKYNTLLATSLCSDEVTRPLEQDFGEVYGKHFTMGGLAGFPFGGL

         81         .         1         .         .         .         .         :         .160
Pt43233     TGFGAMAKHIPDGGSCLVVYGPHVGVDLDGNVGTVNRRGREKGGTCCGSAVAAAGYISKVFNGEADPAPAVPESS----
Pt43234     TGFGAMAGHIPDGGSCLLIYGSHVGVSWEGKWGTVARRGREKGGACCGSAVAAAQAVTQAYQATLDESSSSSSSSATST

        161         .         .         .         2         .         .         .         .240
Pt43233     ---------MDAQQLYVGNMLLPYAERIGNAQDAMVELPYATYEPLDDLMQKIVAKGCGKVGGDGKIALLGGLQINTPA
Pt43234     PVYRYSNDPLDAQQGYVRDMLRPYAATlseaedvMVTLPVSVYDAQQKLVTRILDEGSNHIDGDGQIAVVGGIQINTPK

        241         :         .     ]262
Pt43233     GCPDYFLPLRFEVRDNQNNVLDNLL
Pt43234     EMSDFFVVRRFCIRDSSGNMVENFM
```

Figure 1-4 Sequence alignment Pt43233 (263-521) and Pt43234. Zn coordinating residues in *red,* catalytic Asp-Arg in *gray*, whereas the stabilizing histidine in *blue*.

1.4.2 The ι-CA

With the emergence of bacterial antibiotic resistance worldwide (Davies, 1996), understanding of functional bacterial CAs has increased. Recently discovered in the marine diatom *T. pseudonana.* A newly sequenced CA that has low identity with any known CA has given birth to a new family the ι-CA (*iota*-CA) (Jensen et al., 2019). This bacterial CA usually prefers Mn^{2+} over Zn^{2+} as a cofactor. Furthermore, the ι-CA identified in the genome of *Burkholderia territorii* (BteCAι) proved to be a good catalyst in the hydration of CO2 and sensitive to inhibition (Del Prete et al., 2020).

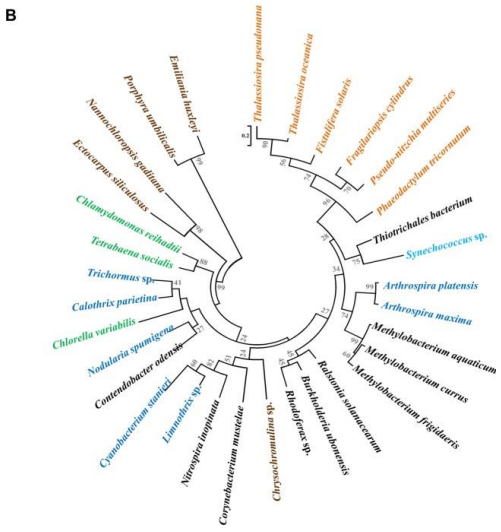


Figure 1-5 Phylogenetic tree of proteins containing LCIP63-like domains. Bootstrap values are shown between nodes. The scale represents the number of substitutions per site. Diatoms (orange), bacteria (black), cyanobacteria (blue), green algae (green), and other Chromista algae (brown) are shown. Fig. S4B from (Jensen et al., 2019). Open access.

Using protonography (De Luca et al., 2015) has been proven that BteCAι can be presented as a dimer, the resulting molecular model showed a *butterfly-shaped* di-dimer.

1.5. Applications of CAs

Their involvement in crucial processes to sustain life, such as pH homeostasis, transportation of CO_2 , respiration, as well as biosynthetic transformations like gluconeogenesis, ureagenesis, etc. makes carbonic anhydrases an ideal target for pharmacological applications. Among an extensive drug target repertoire, especially on CAIs, a recent work proposes CA IX as a novel candidate in liquid biopsy (De Luca et al., 2015), showing the potential of a transmembrane-CA as a cancer marker.

Also, CAs are among the fastest catalytic enzymes known to date, the hydration reaction accelerated by CAs lights a path in what could be a CA-based solution for CO_2 capture, putting CAs as an outstanding research opportunity for environmental purposes(De Luca et al., 2015)(González & Fisher, 2014).

1.6. Scope of this work

Considering how spread and varied the carbonic anhydrases are inserted in life forms as previously described, there is need for having a centralized database that can hold information about them for researchers alike. Even more, to process information and annotate this gene and protein family is a necessary development. On top of this, having an application with concurrent access that allows to perform annotations over the data about CAs is at this point, a necessary tool in the field.

Chapter 2

LITERATURE REVIEW

Before proceeding to describe the proposed solution to the problem previously described, multiple platforms can be found online dedicated to study, analyse, review and inform about genomic information, some including exonic. To a greater extent, some web platforms are dedicated to particular protein families. In this chapter, different tools are reviewed, some of which have serve as a model when the proposed tool that is discussed and analysed in the following chapters.

2.1. Protein family databases

In an effort to analyse the state of the art of tools to analyse protein families (super- and sub-families as well) the following databases have been analysed and when possible, extracted their origin, main aim, technologies used to construct it, their current status and maintenance. For this a resource list created and curated, has been used to find protein family dedicated databases (Health Sciences Library Systems, 2014). Several of them are no longer accessible due to several factors, fortunately one of them being the availability of upgraded versions residing in different servers. From this list, their current development and technologies, it is a safe assumption that research groups dedicated to study protein families develop their own pipelines, scripts and in general utilize well-

established bioinformatics and genomic browsing tools. There is nevertheless, an application that stand out for its fast growth, continuous and active development due to the constant interest in the protein family that it tackles

2.1.1 BioCatNet

"The BioCatNet database system is a repository of sequence, structure and biocatalytic data on protein families to facilitate protein engineering." This online database system, or collection of databases, for protein families is maintained by the bioinformatics group at the Institute of biochemistry and Technical Biochemistry at the University of Stuttgart, Germany. Already in its version 4, it separates different protein families utilising the BioCatNet concept:

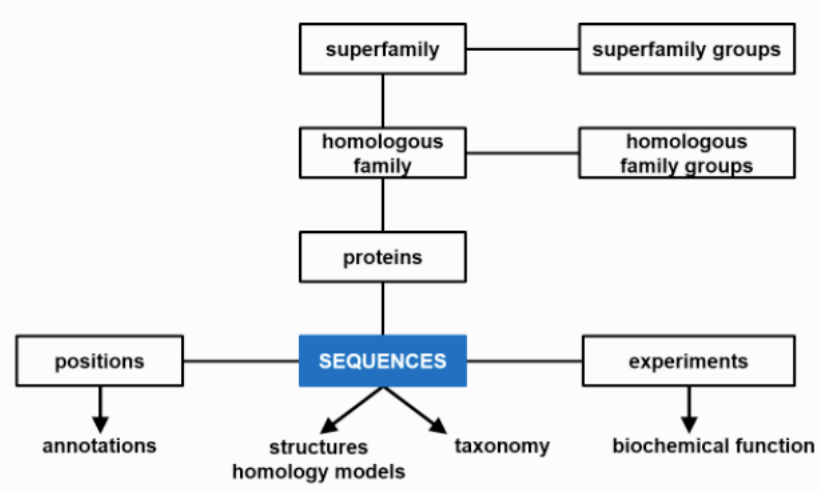


Figure 2-1 BioCatNet, Hierarchically grouping of protein families

Where proteins are assigned to homologous families by similarity of their sequence. Then a similar system to DWARF (Fischer et al., 2006).

Abbreviation	Database	No. of sequences	No. of structures
CYPED	CYtochrome P450 Engineering Database	52674	595
ExED	Expansin Engineering Database	15089	21
GH19ED	Glycoside Hydrolase 19 Engineering Database	22461	27
HYED	Hydratase Engineering Database	2046	3
IRED	Imine Reductase Engineering Database	1409	8
LCCED	LaCCase and multicopper oxidase Engineering Database	51058	229
LED	Lipase Engineering Database	280638	1557
oTAED	ω-Transaminase Engineering Database	114655	234
SDRED	Short-chain Dehydrogenase / Reductase Engineering Database	168212	688
TEED	Thiamine diphosphate-dependent Enzymes Engineering Database	119567	308
TEMLACED	TEM LACtamase Engineering Database	483	65
TTCED	TriTerpene Cyclase Engineering Database	2794	18

Table 2-1 Database collection available under BioCatNet

This collection of databases allows to browse using the native DataTables package for displaying the tables of information, providing a good UX while browsing data. Some earlier versions allowed the direct querying of the database, this was a big security breach.

2.1.2 CyBase

This database dedicated to the cyclic proteins(Fischer et al., 2006), it hosts and analyses their sequences and structures together with applications in protein discovery. Although currently maintained and updated it lacks on the use of current technologies, making difficult to browse the data. It possess appropriate tools to analyse structures, such as the termini distance distributions, yet the tools are hard to use in modern browsers.

This example of a database shows that regardless of the validity of the data, if wrongly presented can deter the usage by having a poor UI/UX implementation

2.1.3 GPCRdb

Originally introduced in 2015(Fischer et al., 2006), GPCRdb is a continuous development by the David E. Gloriam group. This Database is focused on G-protein coupled receptors. This portal on continuous development (Kooistra et al., 2020) focuses on GPCR protein family, an abundant protein that regulates many pathways in human physiology, making this protein family account for 34% of targets of marketed drugs (Hauser et al., 2017). Its due to this that many research groups focuses on them, making GPCRs the most heavily studied drug targets. GPCRdb contains reference data, interactive visualisation and experiment design tools for this protein family, as stated by their website their current statistics vary from holding 423 Human proteins with 43,808 orthologues; 500 GPCR structures with 418 of them refined. In terms of usability, their platform has had more than 38 thousand users in the last year, showing the broad interest in this protein family.

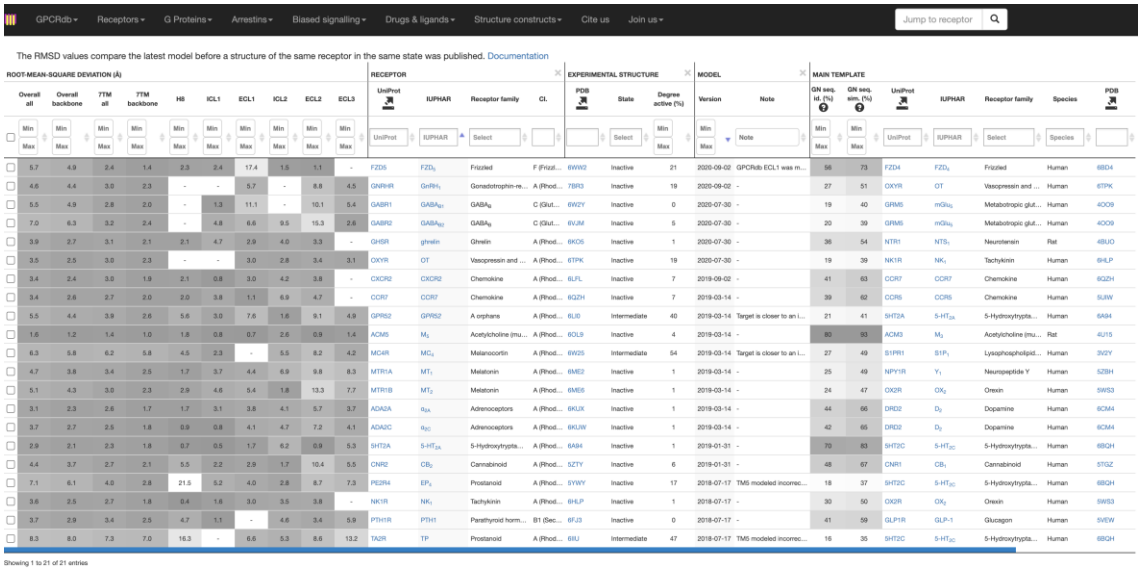


Figure 2-2 GPCRdb Model Statistics view: In the latest version for GPCRdb (Kooistra et al., 2020) compelling views as the Model statistics analyses the accuracy of its models upon release of experimental structures. Views like this make use of interactive tables to visualize and condense the information as the user needs, in-table searches, exportable information as well as column modifiers and filtering.

Since GPCRdb curates sequence alignments, structures and receptor mutation and builds interactive diagrams to visualise receptor residues and relationships (phylogenetic trees). It is a great comparison tool for what a protein family portal can aim for.

2.1.4 CAbase as command-line

In 2015 a first approach to tackle a centralised CA database took part as a thesis at the University of Tampere, Isokangas, 2016, a python-script oriented approach, with programmatic access to ensembl, NCBI and other databases. This project with similar initial motivations became a specific pipeline for CA and their Exons analysis tool, it also provided the possibilities to perform protein analysis by using third-party software such as SignalP (Nielsen et al., 1997) and TargetP (Armenteros et al., 2019). The approach taken, produced two pipelines, one that handled the data and a second one that analysed the conservation of exons per CA, visually displaying the location for cleavage sites.

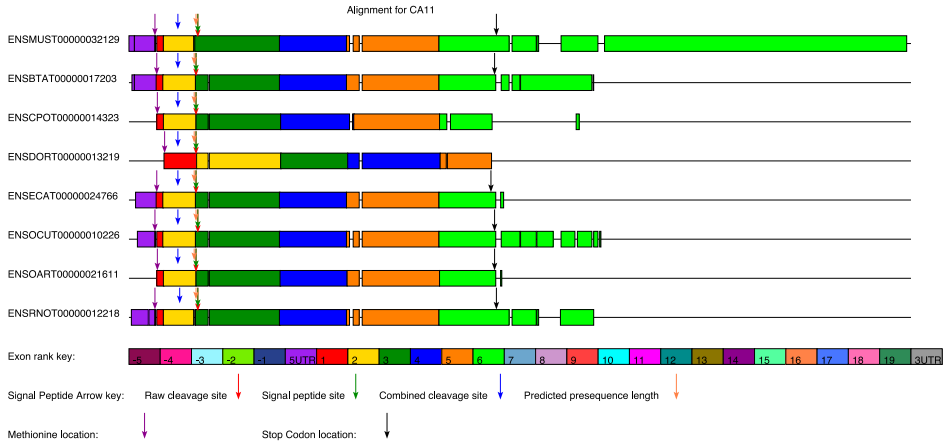


Figure 2-3 Phylogenetically aware codon alignment of cDNA sequences using PRANK(Löytynoja, 2014) in CAbase

It is with certainty that the results obtained in this work are the direction that this present work intends to follow, an in-depth analysis for CA-related exons.

2.2. Ensembl

As part of the European Bioinformatics Institute, the Ensembl genome database project is a major actor in bioinformatics and a recurrent tool to find, browse, compare and analyse genomic data.

Ensembl provides services and tools for comparative genomics procedures, allowing to understand gene evolution, differences and similarities between species at the gene levels, inferring gene functions based on homology and highly conserved regions, important when identifying functional genes.

Comparative genomics is carried out in different taxa clouds, Ensembl Compara, for vertebrates, Ensembl Metazoa Compara, for metazoa, etc. At the same time, there is a Pan-taxonomic, a service which takes a subset of each of the divisions of taxa available in ensembl, in addition to the previous, plants, fungi, protist are some of the examples.

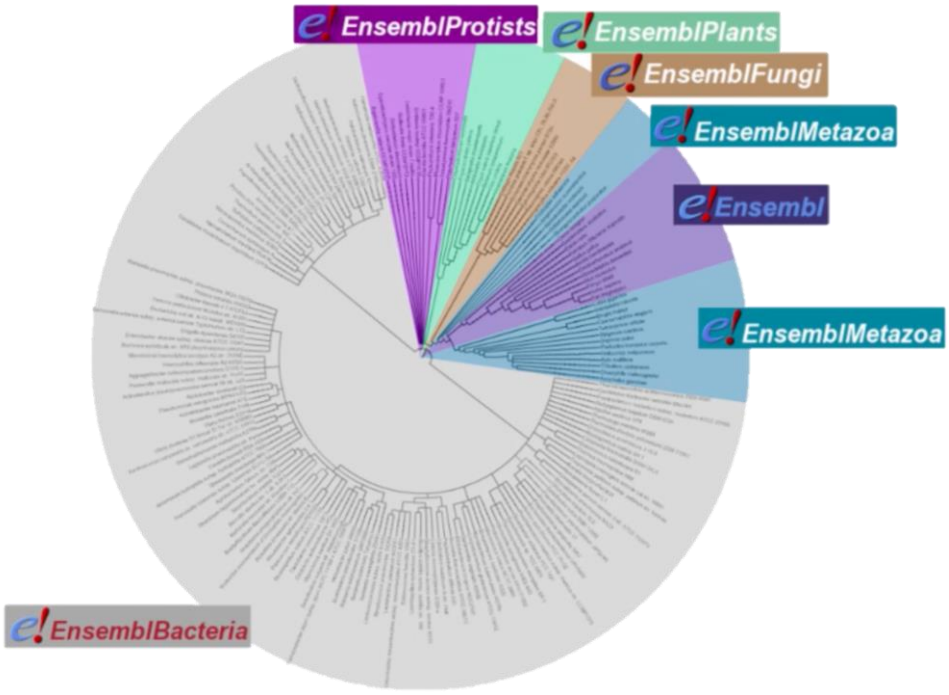


Figure 2-4 Ensembl Pan-taxonomic Compara. Comparative genomics can be carried out in greater detail in each of the taxa clouds of ensebl, or in a more general manner in the pan-taxonomic view. Figure adapted from ensembl comparative genomics website.

Gene trees is one of the ways Ensembl uses to carry out comparative genomics. Gene trees are computed based on protein alignment, where each protein coding gene has a representative protein, they are clustered using BLAST (Altschul et al., 1990) to create multiple alignments which are then reconciled against the species tree allowing to infer the homologues, indicating orthologues and paralogues, genes that come from speciation and duplication events respectively.

The gene tree view graph has green areas that indicate alignment regions, and their different shades indicated the level of clustering, depending whether the node is collapsed or expanded. This view can be also exported in several formats. A more detailed

alignment view is possible, by selectign sub-regions in the gene tree and calling the WASABIAPP (Veidenberg & Löytynoja, 2021)

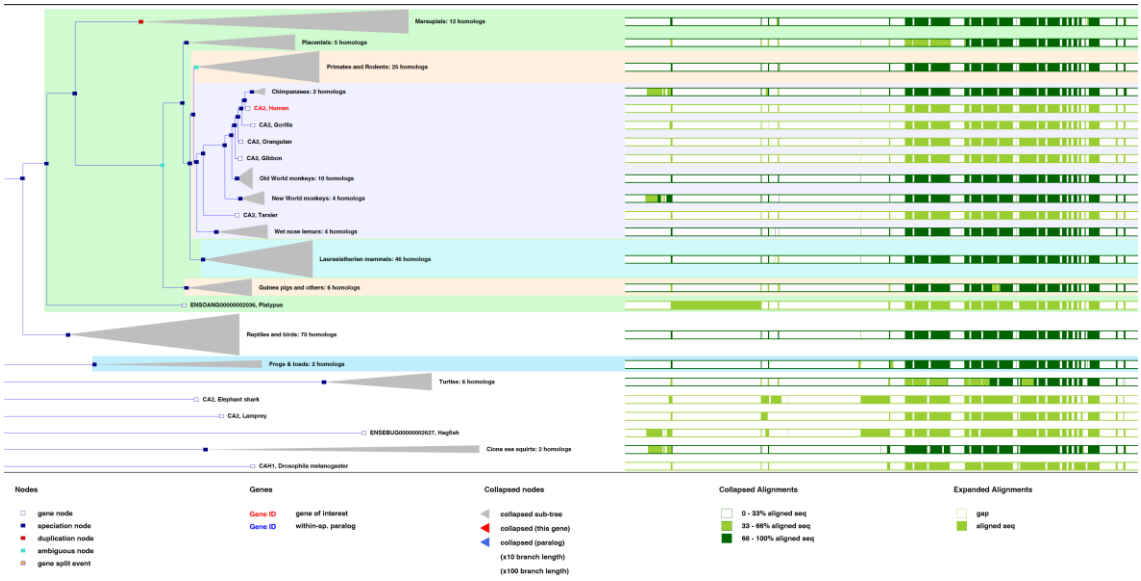
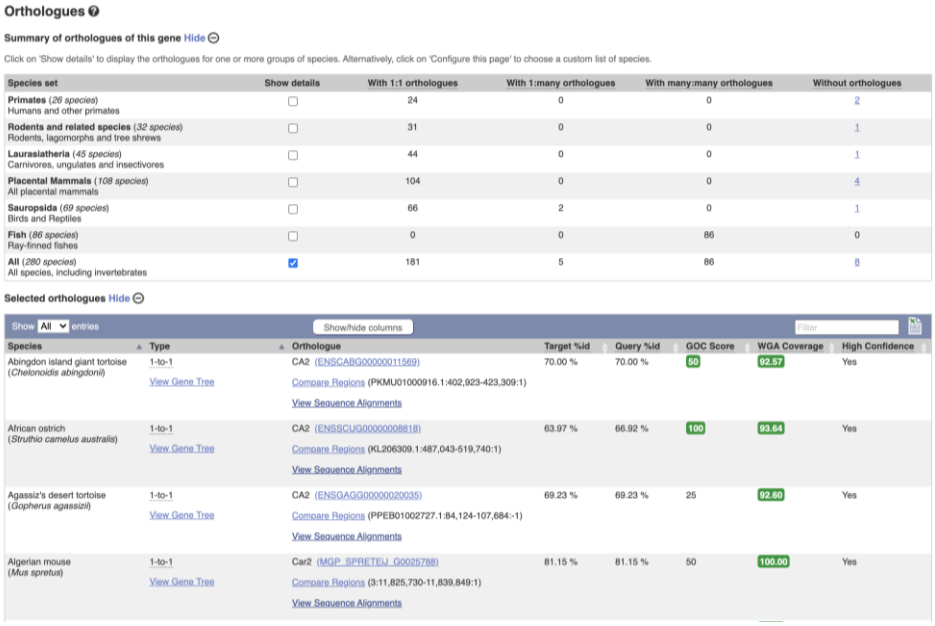


Figure 2-5 Gene tree view for Human CA2. Light purple shaded region shows the expanded gene tree view, where in red, the selected gene and species.

The orthologue section (of special interest for the current work), shows first, a summary the table for which the selected gene possess different types of orthology. This general table can be turned into a more detailed view, this table is built with raw HTML but allows minimum table interactions to browse the orthologues.



Orthologues

Summary of orthologues of this gene Hide

Click on 'Show details' to display the orthologues for one or more groups of species. Alternatively, click on 'Configure this page' to choose a custom list of species.

Species set	Show details	With 1:1 orthologues	With 1:many orthologues	With many:many orthologues	Without orthologues
Primates (*26 species*) Humans and other primates	☐	24	0	0	2
Rodents and related species (*32 species*) Rodents, lagomorphs and tree shrews	☐	31	0	0	1
Laurasiatheria (*45 species*) Carnivores, ungulates and insectivores	☐	44	0	0	1
Placental Mammals (*108 species*) All placental mammals	☐	104	0	0	4
Sauropsida (*69 species*) Birds and Reptiles	☐	66	2	0	1
Fish (*86 species*) Ray-finned fishes	☐	0	0	86	0
All (*280 species*) All species, including invertebrates	☑	181	5	86	8

Selected orthologues Hide

Show All entries | Show/hide columns | Filter

Species	Type	Orthologue	Target %id	Query %id	GOC Score	WGA Coverage	High Confidence
Abingdon island giant tortoise (*Chelonoidis abingdonii*)	1-to-1 View Gene Tree	CA2 (ENSCABG00000011569) Compare Regions (PKMU01000916.1:402,923-423,309:1) View Sequence Alignments	70.00 %	70.00 %	50	92.57	Yes
African ostrich (*Struthio camelus australis*)	1-to-1 View Gene Tree	CA2 (ENSSCUG00000008818) Compare Regions (KL206309.1:487,043-519,740:1) View Sequence Alignments	63.97 %	66.92 %	100	93.64	Yes
Agassiz's desert tortoise (*Gopherus agassizii*)	1-to-1 View Gene Tree	CA2 (ENSGAGG00000020035) Compare Regions (PPEB01002727.1:84,124-107,684:-1) View Sequence Alignments	69.23 %	69.23 %	25	92.60	Yes
Algerian mouse (*Mus spretus*)	1-to-1 View Gene Tree	Car2 (MGP_SPRETEJ_G0025788) Compare Regions (3:11,825,730-11,839,849:1) View Sequence Alignments	81.15 %	81.15 %	50	100.00	Yes

Figure 2-6 Ensembl Ortholog table view. As exemplified by the human CA2 gene.

The selection of specific orthologues allows to view the pairwise alignment between the selected gene and the orthologue of interest, computed by Clustal W(Thompson et al., 1994).

Chapter 3

TOOLS AND METHODOLOGY, BUILDING CABASE

As described in previous sections, CAs are a ubiquitous, well-studied and, in terms of dedicated tools, a de-centralized family of proteins. With this big scope in mind, a system in which the analysis of such proteins and its genetic units is possible has been conceived, enter, CAbase, a TPA:inferal database that aims to agglomerate information related to CAs. With an initial motivation to solve some of the problems encountered when newly sequenced CA members could be named, following naming conventions that will structure and give sense to their names, it became clear that the potential of a tool of this nature could extend beyond merely an informative archive.

Is for this reason that more dedicated bioinformatics tools were decided to be included as part of its development. In this chapter the proposed solution is described, from its architectural design and implementation, to its use and views. First an

introduction to the technologies used to build the TPA. Second, an overview of the system and its main capabilities are described. Lastly, the building of features for Exon analysis.

Figure 3-1CAbase logo: “From gene to protein”. Depicting a double helix and a GPI anchor

The status of novel approach that CAbase claims, is in part due to the type of development, it uses non- conventional technologies for Bioinformatic tools. CAbase is a PHP-based web-application, that uses the Laravel Framework, allowing a robust and secure multi-purpose application. Given its aim, it implements user access, with different roles accessing different sections and tools. From being merely informative for guest users, to allow the edit of raw sequences by higher administrative profiles.

3.1. MVC

Laravel framework is a web application framework with expressive and elegant syntax. A robust multi-purpose framework, backed-up by the most varied cross-industry projects. The Model View Controller pattern of programming was proposed as an approach for CAbase, allowing to modularize the different sections of the software.

3.1.1 Models

Models are the framework entities that define how to manipulate the data. It is a programming layer between the application and the data. Following, an example of a Model definition for the *Gene* model used by CAbase:

```
class EnsemblGene extends Model
{
    use SoftDeletes;

    public $table = 'ensembl_genes';

    protected $dates = ['deleted_at'];

    protected $primaryKey = 'ensembl_genomic_id';

    public $fillable = [
        'ensembl_genomic_id',
        'taxon_id',
        'assembly_name',
        'source',
        'db_type',
        'display_name',
        'logic_name',
        'description',
        'seq_region_name',
        'strand',
        'start',
        'end',
        'sequence',
        'version'
    ];

     public function taxonomy(){
        return $this->belongsTo(\App\Models\Taxonomy::class,'taxon_id','id');
```

```
29      }
30
31      public function homologyAsTarget(){
32          return $this->belongsTo(\App\Models\Homology::class,'ensembl_genomic_id',
    'target_genomic_id');
33      }
34
35      public function homologyAsSource(){
36          return $this->belongsTo(\App\Models\Homology::class,'ensembl_genomic_id',
    'source_genomic_id');
37      }
38
39      public function ensemblTranscripts(){
40          Return $this->hasMany(\App\Models\EnsemblTranscript::class,
    'ensembl_genomic_id', 'ensembl_genomic_id');
41      }
```

This model is related to other models via ordinal relationships, for example, a *Gene*, has many *Transcripts* Indicating that any given gene in the system can have, at least one transcript. This defines the flexibility of each Model and what data can be manipulated with them.

3.1.2 Controllers

In an MVC application the Controllers are the part that is in charge of handling the requests an passes data from the Model to the views. This layer is the link between the Model and the View. For a software that dissects its requests in an API, it is the Controller the one handling the HTTP requests (properly routed by the routing engine of the framework). In the following example the storage of a new Carbonic Anhydrase in the system is shown:

```
1   public function store(CreateCarbonicAnhydraseRequest $request)
2   {
3       $input = $request->all();
4
5       $carbonicAnhydrase = $this->carbonicAnhydraseRepository->create($input);
6
7       Flash::success('Carbonic Anhydrase saved successfully.');
8
9       return redirect(route('carbonicAnhydrases.index'));
10  }
```

The following table summarises how the Requests are handled and implemented in CAbase

Table 3-1**CAbase Request structure** Five basic manipulation are noted, followed by examples on how CAbase implements them. Note: if *soft delete* is implemented, the *destroy* Request does not permanently delete the resource, instead a PATCH request updates the resource indicating the date of alleged deletion

Request	Description	Method	URI *example*
index	*List all resources*	GET	**/resource** /ensembl/genes/
show	*Show specific resource*	GET	**/resource/{resourceId}** /ensembl/transcripts/ENSABRT00000021194
store	*Store a new resource*	POST	**/resource** /ensembl/proteins/
update	*Update the uniquely identifiable resource*	PATCH	**/resource/{resourceId}** /ensembl/exons/ENSABRE00000115621
edit	*Show possible edits for a resource*	GET	**/resource/{resourceId}/edit** /taxonomies/9606/edit
destroy	*Delete specific id.*	DELETE	**/resource/{resourceId}** /carbonic_anhydrase_isozymes/2

Another example in which the Controller plays a crucial role in CAbase, is the handling of the tables, the DataTable entity, these entities will render views containing interactive tables.

```php
public function dataTable($query){
  $dataTable = new EloquentDataTable($query);
  return $dataTable
     ->addIndexColumn()
     ->addColumn('action', 'ensembl/genes.datatables_actions')
     ->addColumn('fastaSeq',function(EnsemblGene $ensemblGene){
         return view('ensembl/genes.datatables_details',compact('ensemblGene'));
     })
     ->rawColumns(['fastaSeq','action']);
}

public function query(EnsemblGene $model){
   return $model->newQuery()->with(['taxonomy']);
}

public function html(){
   return $this->builder()
     ->columns($this->getColumns())
     ->setTableId('ExonsDataTable')
     ->minifiedAjax()
     ->dom('Bfrtilp')
     ->pageLength(50)
     ->lengthMenu([10, 50, 100, 200, 500])
     ->autoWidth(false)
     ->addAction([
         'width' => '120px',
         'printable' => false,
         'title' => '',
         'class' => 'noColVis'
        ])
 }
```

In order to generate a DataTable, the controller needs 3 parts. First, the data from the Model and its relationships, EnsemblGene with *taxonomy* in the example above (in order to render the species attributes), this in the *query* function,. Secondly, the *html* function that will build the table for the view. Lastly, the *dataTable* function that will return the dataTable object to the view.

3.1.3 Views

Views are the visual representation of the application; this layer of the software is responsible for displaying the data that the Controller gathered from the Model. In Laravel

an engine to render dynamic data is used, called *Blade templating,* this engine extends HTML by injecting data retrieved from the Models. The following code block shows how the Transcript sequences are included in the view, pre-processed before building the final HTML passed to the client:

```
<div class="row">
  <div class="col-md-4">
   Genomic sequence:<pre>{!! toFASTA($ensemblTranscript->genomic_sequence,60)!!}</pre>
  </div>
  <div class="col-md-4">
   cDNA sequence:<pre>{!! toFASTA($ensemblTranscript->cdna_sequence,60)!!}</pre>
  </div>
  <div class="col-md-4">
   CDS sequence:<pre>{!! toFASTA($ensemblTranscript->cds_sequence,60)!!}</pre>
  </div>
</div>
```

Or how the Exon edit view form is dynamically created

```
<div class="card-body">
    {!! Form::model($ensemblExon, ['route' => ['exons.update',
    $ensemblExon->ensembl_exon_id], 'method' => 'patch']) !!}

    @include('ensembl/exons.edit_fields')

    {!! Form::close() !!}
</div>

```

For the previous example, the @include directive will search for the edit fields in the following route:

```
app/resources/views/ensembl/exons/edit_fields.blade.php
```

This file, another blade template will include the rest of the form. Then, the final view can be observed:

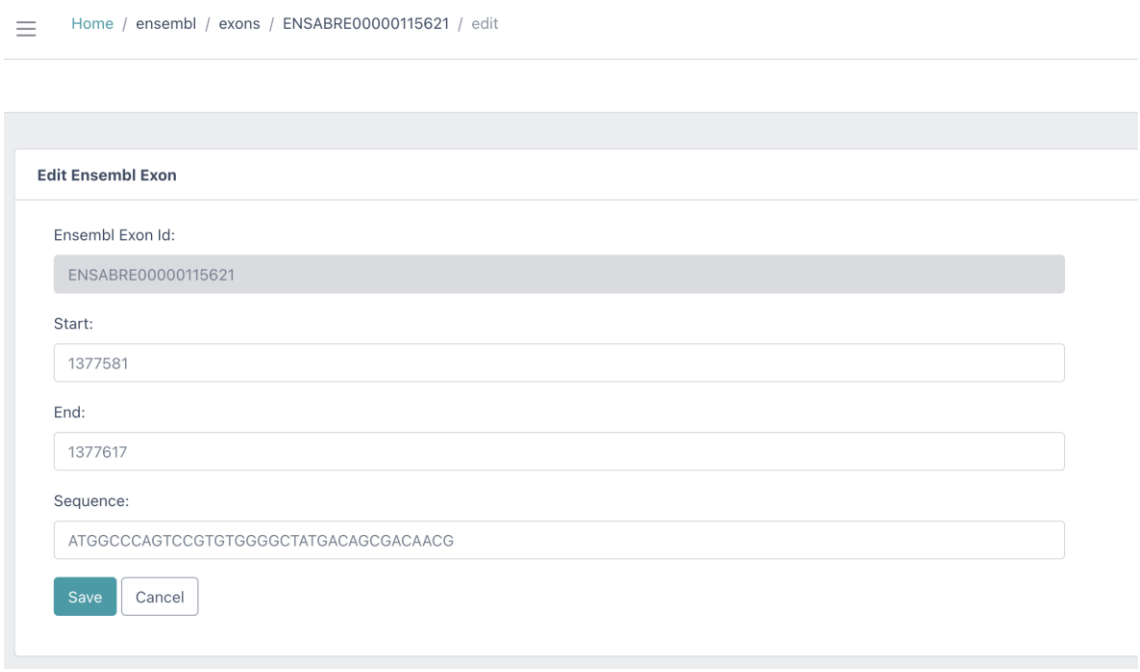


Figure 3-2 Exon edit view: The three editable attributes of an exon are shown in the form.

3.2. The Database

As any TPA, there is a heavy database load on the application. For this, CAbase relies MySQL as its RDBMS. An overview of the main accessed tables can be seen in the following figure, the full diagram is listed in Appendix 2:

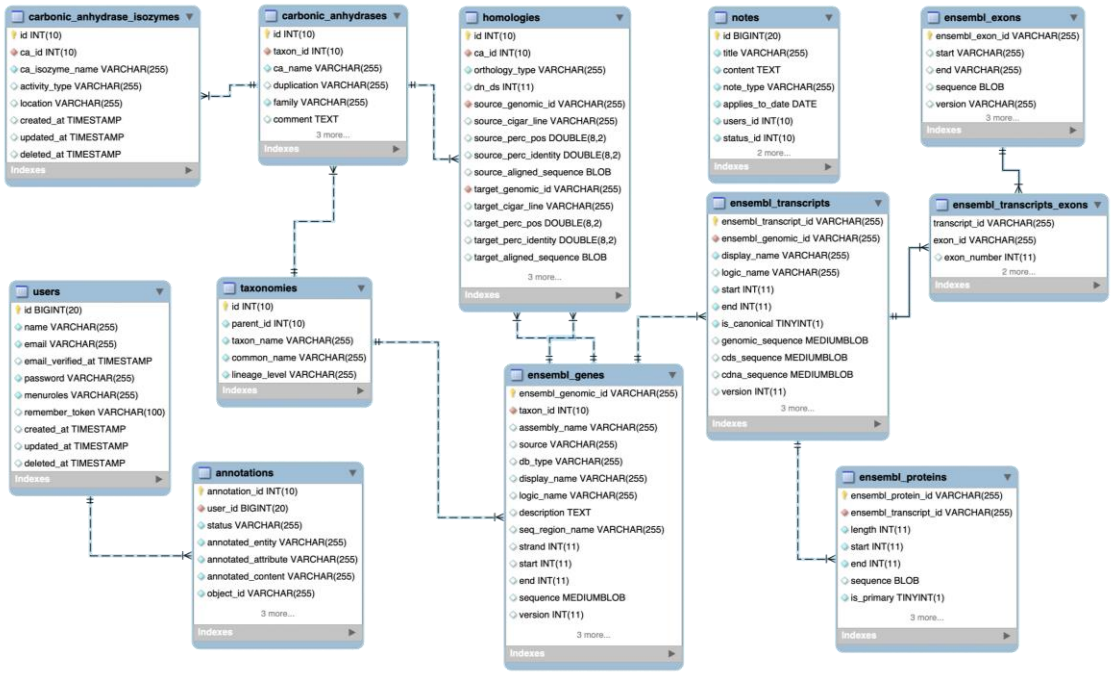


Figure 3-3 Database Model diagram for CAbase

3.3. Third-party tools

3.3.1 TaxaDB

As one of the first steps in development, CAbase has included the full NCBI taxonomies entries from the Taxonomy Database (Federhen, 2012), using a third-party application, TaxaDB (Gourlé, 2019), this application is python based and focuses on building a local taxonomy database for faster accession and local manipulation of the data. Nevertheless, for the usage of CAbase it was requested for modification, allowing the local implementation of the database to include not only scientific names, but also common species names.

3.3.2 Yajra DataTables

In order to build interactive tables, as the one implemented by 2.1.3, a plugin like jQuery's DataTables was needed. Yajra DataTables is a package that allows the usage of DataTables, powerful tables with the Laravel framework. Making use of the full potential by accessing the data via the Models and configuring the tables via Controller as shown in 3.1.2.

3.3.3 MView

Depicting the alignments and the colouring for their conservation has been done incorporating a Perl-based application, MView (Brown et al., 1998). Here an example of a call to the program from the CAbase environment

```
exec("resource_path('views/mview/mview')"
                ." ".$fileInput." > ".$fileHtml
                ." -in plain"
                ." -width 60"
                ." -coloring identity"
                ." -html data"
                , $output);
            $this->comment( implode( PHP_EOL, $output ) );
```

This block uses the app and uses the output to be later manipulated. The output is an HTML version of the alignment with a fixed width. This information is persisted as a file.

In MView, the *percent of coverage* reported in each alignment row is calculated with respect to the reference sequence, the first sequence in $fileinput:

$$\%_{cov} = \frac{R_A}{U_l} \times 100$$

Where R_A is the number of residues in row aligned with reference row; U_l = length of ungapped reference row.

As for the *identity percentage* reported in each alignment row is calculated with respect to the aligned portion of the reference sequence.

$$\%_{id} = \frac{\mathrm{I}}{U_l(A)} \times 100$$

Where I is the number of identical residues; $U_l(A)$ = length of ungapped reference row over aligned region.

3.4. CA Classification

In order to solve possible future naming conflicts, CAbase proposes a naming system for consensus CAs. In order to achieve this, a simple three-way entity relationship, ensures each species gets both its gene and isoforms named.

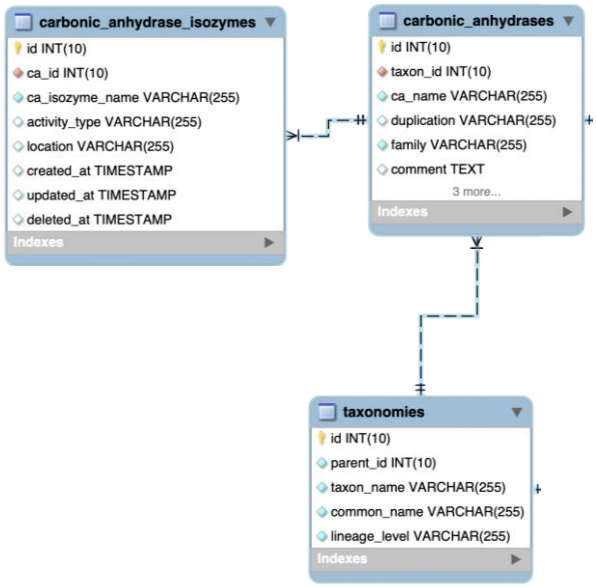
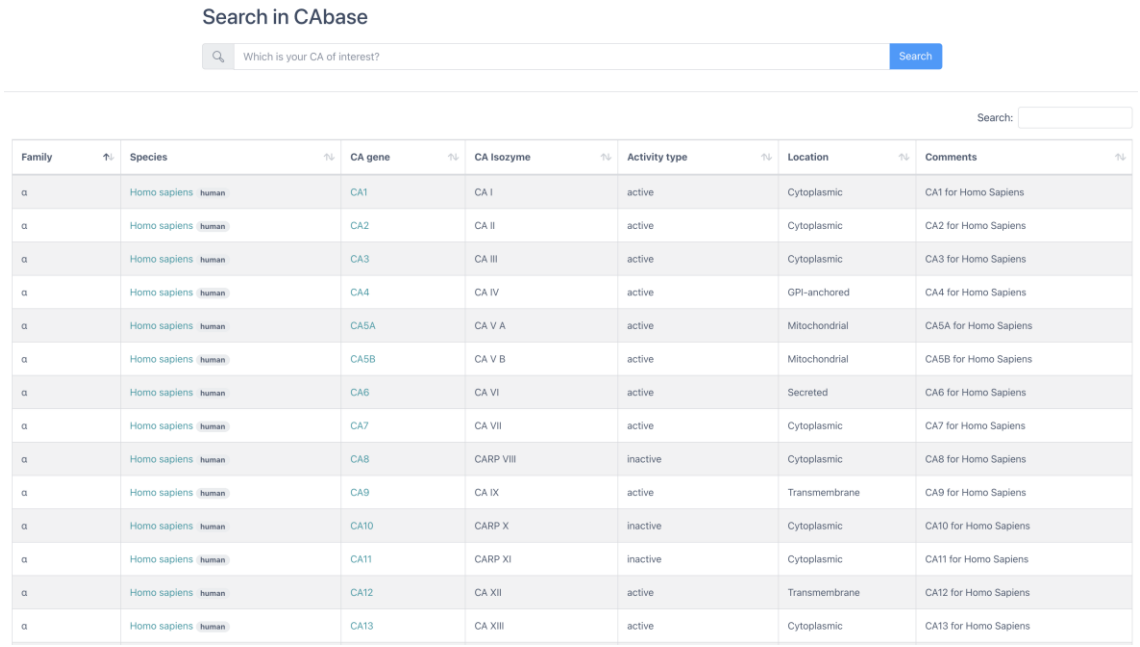


Figure 3-4 CAbase Gene-Isoform-Taxonomy database relationship model

With this structure, the following table (and landing page) of CAbase can be generated:



Family	Species	CA gene	CA Isozyme	Activity type	Location	Comments
α	Homo sapiens human	CA1	CA I	active	Cytoplasmic	CA1 for Homo Sapiens
α	Homo sapiens human	CA2	CA II	active	Cytoplasmic	CA2 for Homo Sapiens
α	Homo sapiens human	CA3	CA III	active	Cytoplasmic	CA3 for Homo Sapiens
α	Homo sapiens human	CA4	CA IV	active	GPI-anchored	CA4 for Homo Sapiens
α	Homo sapiens human	CA5A	CA V A	active	Mitochondrial	CA5A for Homo Sapiens
α	Homo sapiens human	CA5B	CA V B	active	Mitochondrial	CA5B for Homo Sapiens
α	Homo sapiens human	CA6	CA VI	active	Secreted	CA6 for Homo Sapiens
α	Homo sapiens human	CA7	CA VII	active	Cytoplasmic	CA7 for Homo Sapiens
α	Homo sapiens human	CA8	CARP VIII	inactive	Cytoplasmic	CA8 for Homo Sapiens
α	Homo sapiens human	CA9	CA IX	active	Transmembrane	CA9 for Homo Sapiens
α	Homo sapiens human	CA10	CARP X	inactive	Cytoplasmic	CA10 for Homo Sapiens
α	Homo sapiens human	CA11	CARP XI	inactive	Cytoplasmic	CA11 for Homo Sapiens
α	Homo sapiens human	CA12	CA XII	active	Transmembrane	CA12 for Homo Sapiens
α	Homo sapiens human	CA13	CA XIII	active	Cytoplasmic	CA13 for Homo Sapiens

Figure 3-5 CAbase Isoforms summary. This visible part of a more complete table, is the summary from the CAs in human, showing the distinctive groups.

3.5. Ensembl as data provider

Originally developed in the Perl language, since 2012 has released a language-agnostic API (besides the opening of a more compelling Perl API). The source of raw data for CAbase is incorporated by the software by making use of this API. Prior to populate the database, the API endpoints(Kumari & Kline, 2020) are here discussed. All the following endpoints have the parameter content-type set to application/json in order to obtain an easy to manipulate JSON object (Bray, 2017).

3.5.1 Homology Endpoint

The homology-based approach that CAbase possess and moreover the ortholog approach is tackled by this endpoint. Particularly CAbase uses

```
2  https://rest.ensembl.org/homology/symbol/9606/ca4/
```

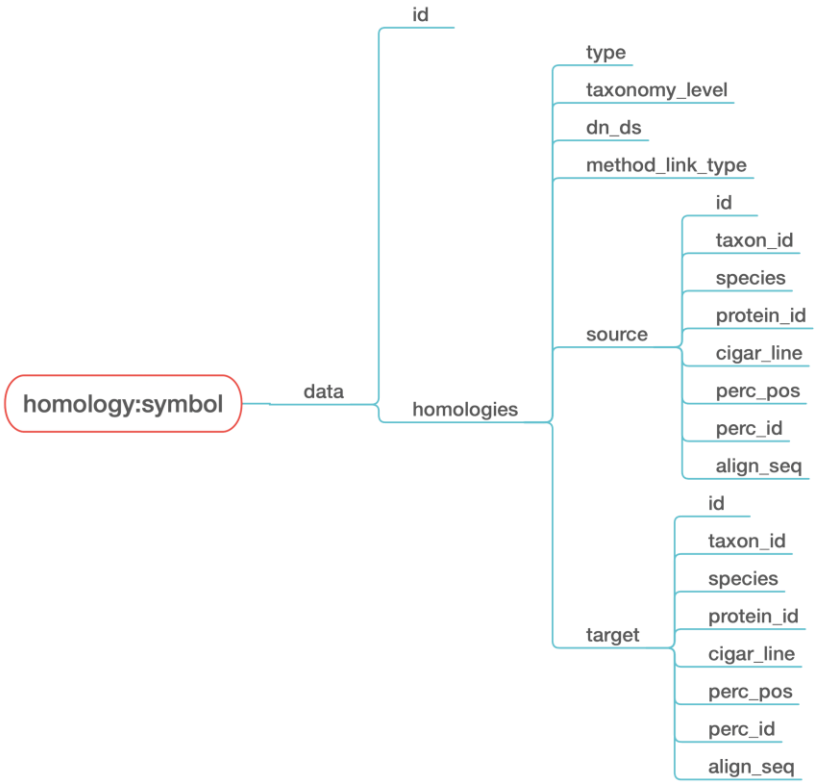


Figure 3-6 Ensembl Homology symbol endpoint hierarchical representation of the response body

3.5.2 Lookup Endpoint

CAbase utilises this endpoint to obtain the main set of information for each entity, Genes, Transcripts, Translations and Exon. The endpoint:

```
3  https://rest.ensembl.org/lookup/id/ENSG00000167434?expand=1;
```

shows the two required parameters, the StableID and expand 1, Each entity is identified with its StableID and the expand parameter will show, for a gene ID, its Transcripts, Exons and Translation. CAbase only stores protein coding transcripts.

As Ensembl (and the major Biological databases too) stores the data, it might happen that the same exon is under different transcripts

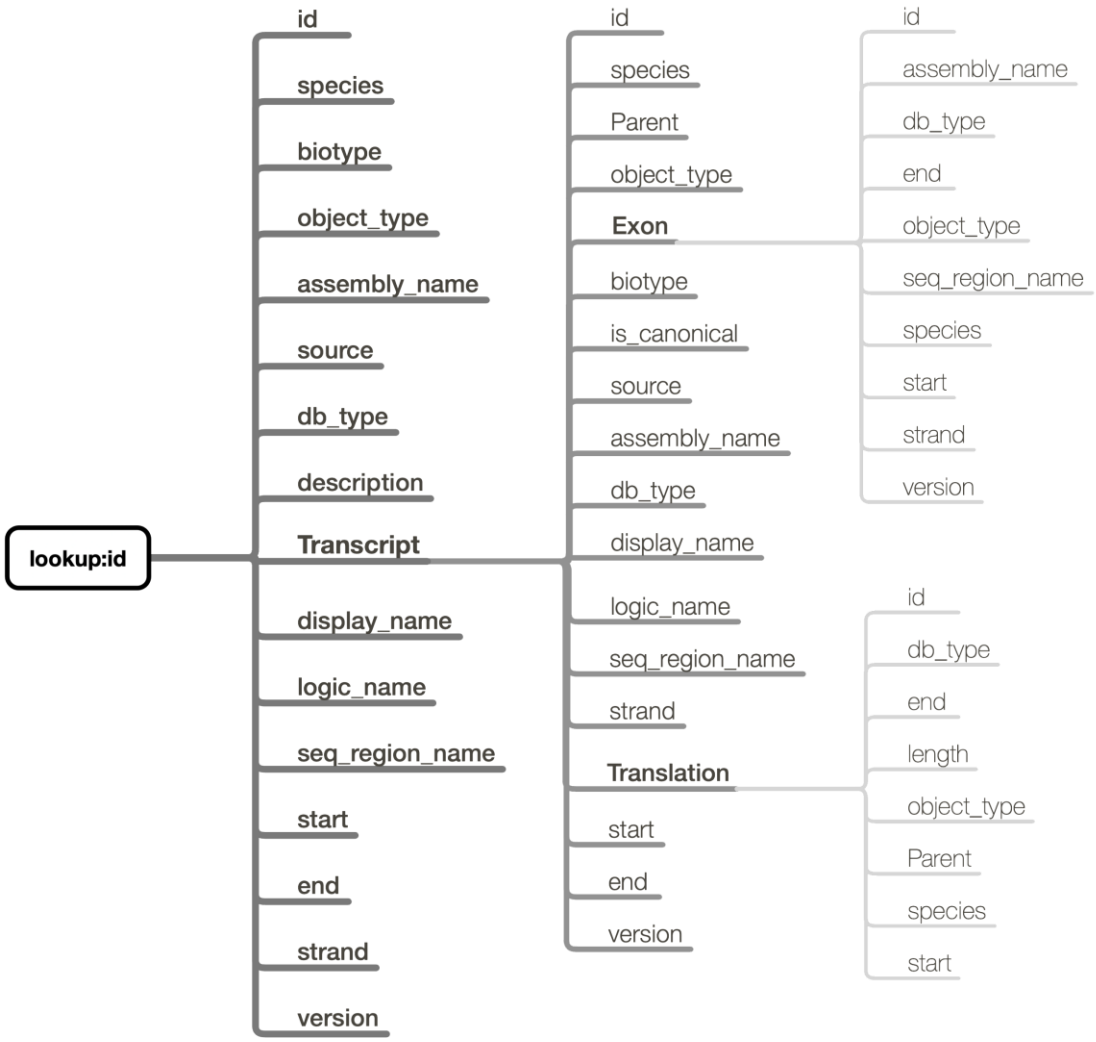


Figure 3-7 Ensembl Lookup endpoint response body

CAbase does not hold duplicated information that comes from higher hierarchical levels, i.e., the species for the exon or transcript, as well as the as db_type.

3.5.3 Data retrieval

In order to obtain the raw data from which CAbase and its users will perform analysis and annotations, it is necessary to persist the data for it to be manipulated. The way in which CAbase obtains this data is utilizing a common interface in web design, a step-by-step wizard, or assistant. It has been designed this way due to the Homology-

based approach to obtain the different genes and their information. The assistant view can be summarized by the following steps with their use of different endpoints of the REST API:

1. Select the CAbase of interest
   1.1. Select the species source of the homology
   1.2. Select one of their associated CAs
2. Collect their homologue
   2.1. The main Gene Stable ID for the selected CA is shown
   2.2. Using the homology/symbol endpoint a summary of the number of homologies and orthologues are displayed
3. Collecting the entities data
   3.1. With the list of ortholog genes, their information is obtained by using the lookup/id endpoint
   3.2. Using the sequence/id endpoint, the sequence of each entity is obtained
      3.2.1. Genomic for Genes
      3.2.2. CDS, cDNA and Genomic for Transcripts
      3.2.3. Peptide Sequence for Translations
4. All the previously collected data is persisted in CAbase

In the following figures, the process of bringing novel data into the system is exemplified by downloading information for *ca12* gene for zebrafish (ENSDARG00000045644).

1. Selecting the Reference species and the Reference Carbonic Anhydrase:

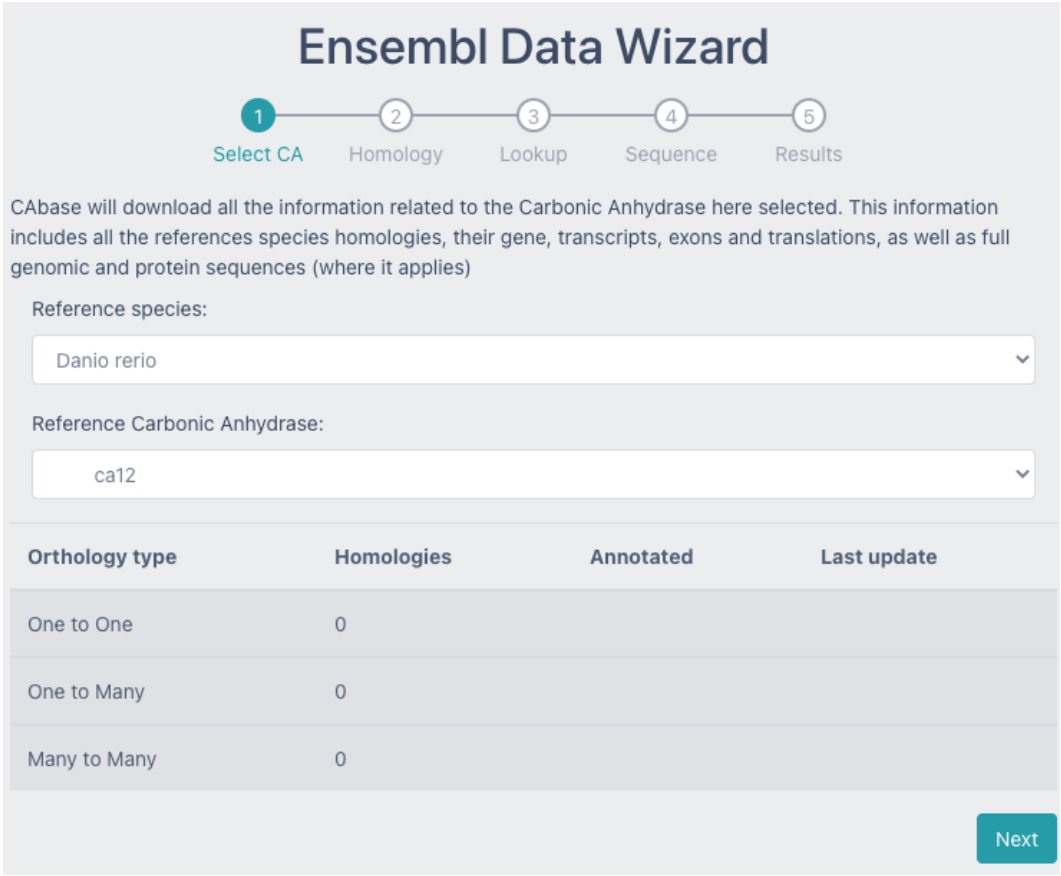


Figure 3-8 Ensembl wizard, step 1 CA selection

When the aforementioned CA is selected, a table displays that there are no orthologues found in CAbase, indicating it is new data being incorporated to the system.

2. Accepting the symbol found.

The second step will show the symbol found using the homology/symbol endpoint:

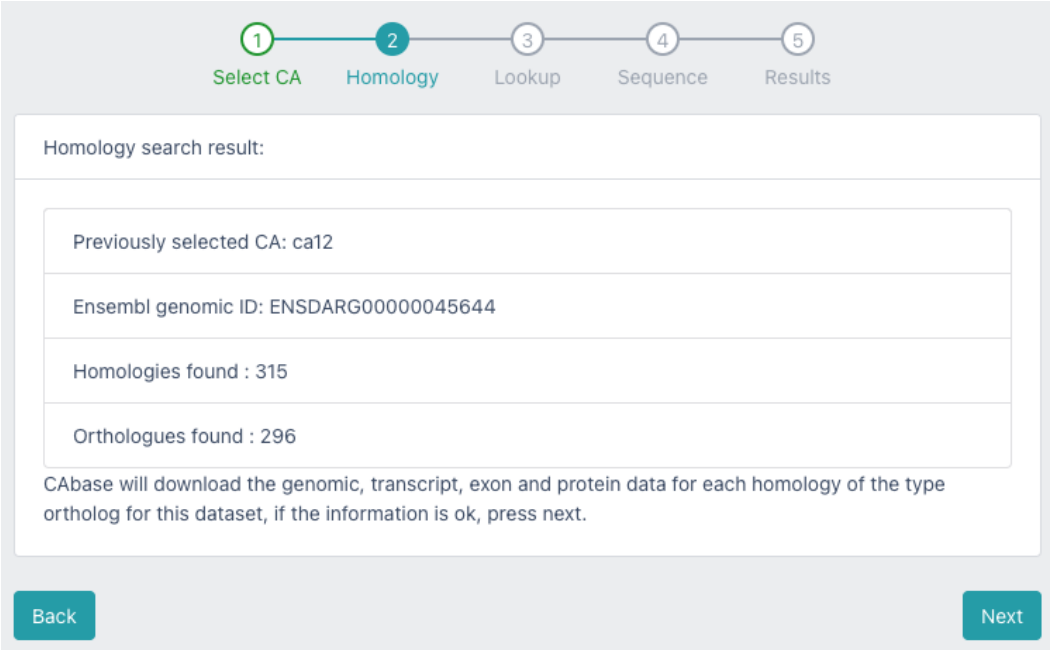


Figure 3-9 Ensembl wizard, step 2 Homology search

CAbase will discard the paralogs homology, accepting only orthologs in ‘one to one’, ‘one-to-many’ and ‘many to many’ relationships.

3. Downloading the gene tree information

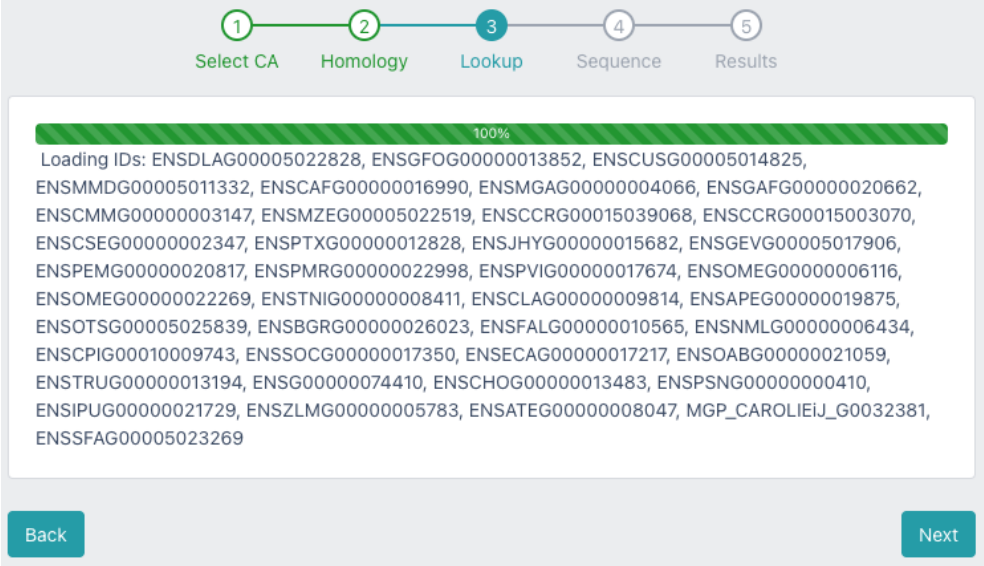


Figure 3-10 Ensembl wizard, step 3 Detailed gene information retrieval

The figure shows the culmination of the download, this process utilises a POST request, meaning a group of maximum 500 identifiers are sent as a request to the *lookup* endpoint to the ensembl servers. Given that this request is used utilising the parameter 'expand', that retrieves the whole gene tree (transcripts-translations-exons) and in order to reduce the server's workload, batches of 50 symbols are sent per request. It is worth mentioning this endpoints response body does not contain sequences, for this reason a separate step for this is needed.

4. Downloading the gene tree information

In a similar manner to the previous step, the sequences are retrieved using the same homology dictionary (gene ids), expanded by the ids for transcripts, translations and exon ids from the previous step.

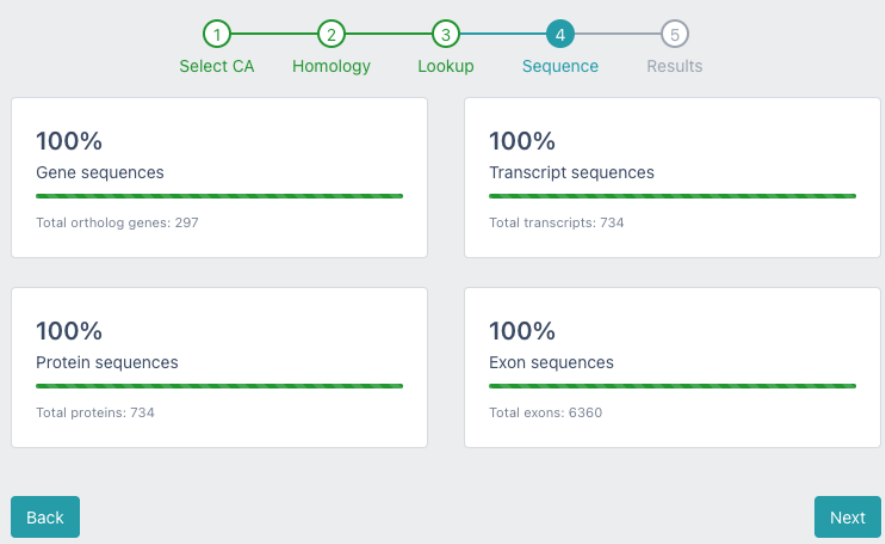


Figure 3-11 Ensembl wizard, step 4 Sequences. Indicating the number of sequences in download process.

5. Results table

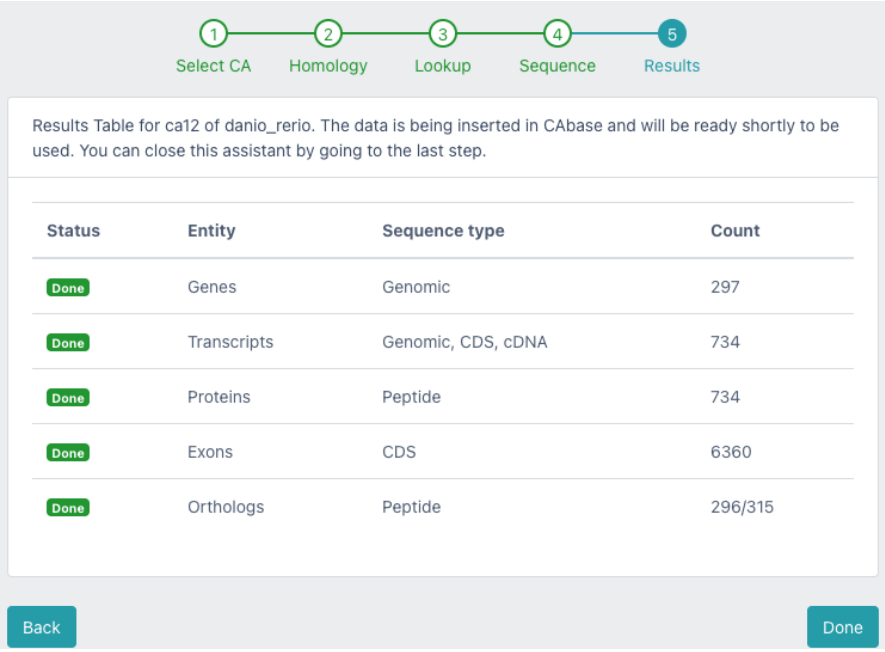


Figure 3-12 Ensembl wizard, step 5 Results.

3.6. The platform

CAbase uses a profile system that differentiates 3 types of users, *guest*, *researcher* and *administrator*. Guest user can only access the homepage and browse the different Carbonic Anhydrases in the system, whereas the researcher can have full access to the information and its distinction with the administrator profile is that in order to perform any change in the data, this has to pass a process of validation. With the exception of the Manual annotation's description, the views here onwards described are proposed from the administrator's point of view, in which any changes proposed are immediately reflected in the views.

3.7. The entities tables

As seen in 3.2 each of the basic information is hold in separated entities, where each of them has a separate view in the platform. These views are CAs with CAs Isozymes as a separate view), then, the Homology table, finally the ensembl entities, Genes, Proteins, Transcripts and Exons, with this information the following analysis views have been developed. Since the Exon analysis is the focus of the present work, the Exon view is now described.

3.7.1 Exon view

The information that the Exon holds has been already discussed in , and to this there are some attributes that needs to be added. As ensembl (and the major Biological databases too) stores the data, it might happen that the same exon is under different transcripts for this reason, each exon might have a different exon number, hence, the link to the transcript(s) that the exon is present indicates in parenthesis the exon number for that transcript. Each exon, is a row in the table that has 3 actions view, edit and delete.[1] The editable information is the raw sequence (CDS), its start and the ending positions.

[1] View is accessible by all users, Edit and Delete is seen by administrator user whereas the researcher will see Annotate.

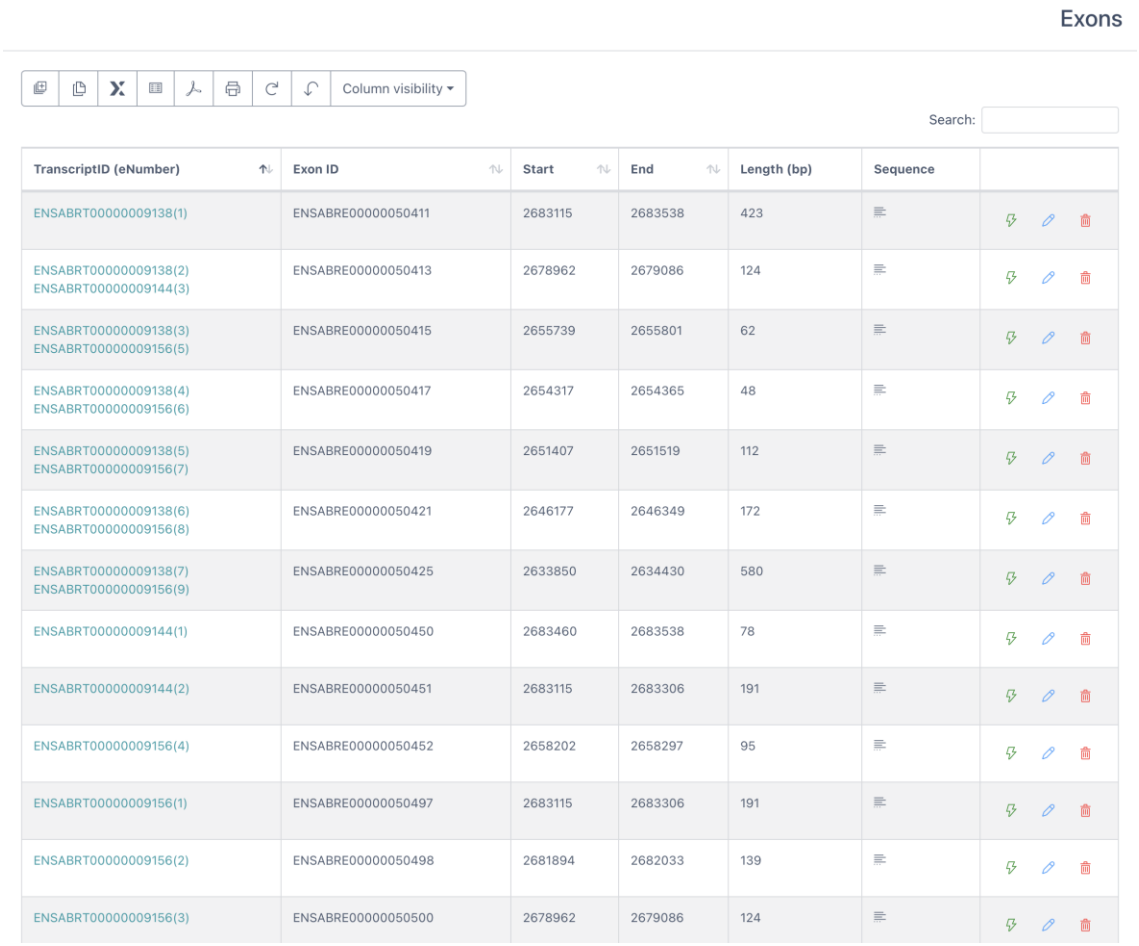


Figure 3-13 Full Exon Table.

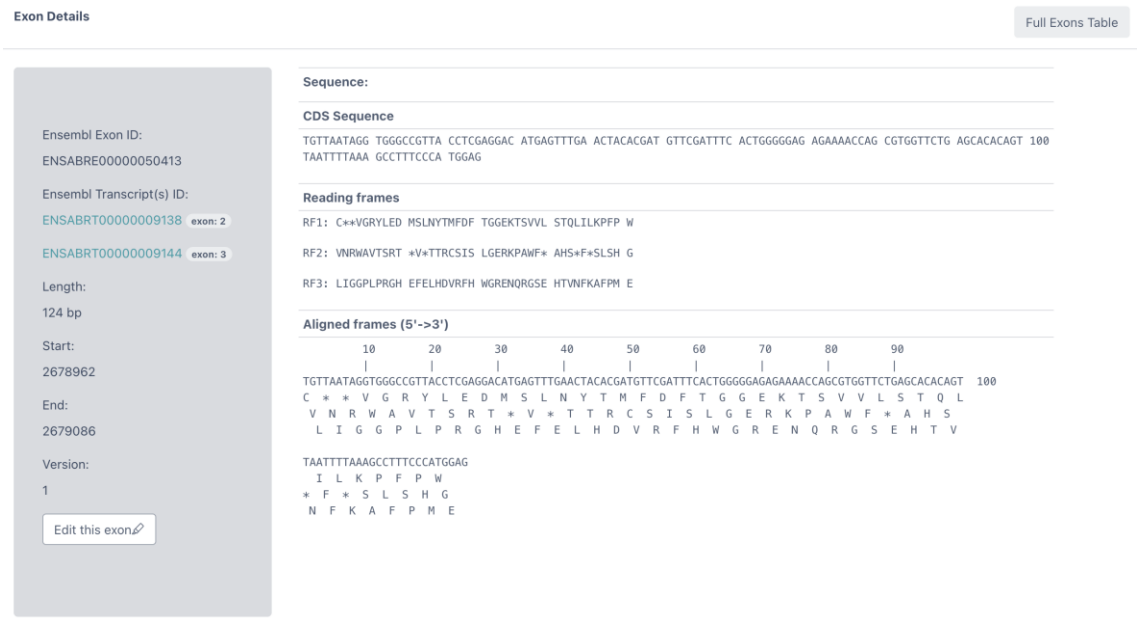


Figure 3-14 **Detailed Exon view** will display the CDS code in a maximum of 100 bp per line. Reading frames are possible to visualize too.

Exemplified by the exon view, this general view (Figure 3-13), and detailed view (Figure 3-14) can be applied to all the previously discussed entities of the system. In the views themselves, upper buttons allow to export the table to different formats (plain text, csv, xls or PDF) or access the printable view. Reset and reload buttons are available in case the view wants to be taken back to its original state. Colum visibility allows to have a clear view on the attributes of interest. The views implement a *saved state* where the column viewed, their reordering and sorting is saved for future logged sessions[2].

3.8. Predicted translation

This function indicates, if any of the possible peptide sequences from the exon's CDS code is matched within the translation of the transcript, with the following convention for the labels:

RF: x aa	The Reading frame in which there is a *complete match* in the full translation, next to the label, the amino acid chain length of the exon
RF: x aa	Indicates a partial match in the full translation, indicating the ORF with a greater length matched. Next to it, the length of the match.
No RF	There is no match found within the full peptide sequence. Matches of single aa are not counted.

3.9. Exon Tagging

On both sides of the exons, reside the intronic regions, parts of non-coding DNA. these sequences are proportionally much more extense than the exonic or coding regions. The importance of identifying these sites for each exon cannot be understated, and they can allow to better understand the role they play in the expression pathway of the gene.

[2] This applies when the user logs in with the same browser

This process is not only for the canonical spliced transcriptome. Analysing all exon for the Carbonic anhydrase family can incorporate new knowledge in areas such as Alternative Splicing, mechanism that plays an utterly fundamental role in gene function and expression.

The feature called 'Exon tagging' is a process that analyses all exon presents in CAbase and annotates them by looking for the predecessor and successor dinucleotides, taking advantage of the fact that introns have consensus dinucleotides in each end. 'GT' is in the beginning of most introns (after each exon except the last one) and AG is in the end of each intron, before each exon except the first one. They serve as marker for exon cleavage to transform pre-mRNA into mRNA. Knowing the exon start and endpoints, the transcript length and its start and endpoints also makes it easy to retrieve those dinucleotides and with them, label each exon with the following annotation structure:

1. 'Splicing OK': The exon is tagged with this label when the exons neighbouring introns end and start with the consensus dinucleotides. If the transcript starts or end at different positions than the first or last exons, the end or the start of the non-coding regions, respectively, are not considered.

2. 'Non-standard 5'': When only the 5' ending of the neighbouring intron does not contain the expected dinucleotides.

3. 'Non-standard 3'': When only the 3' ending of the neighbouring intron does not contain the expected dinucleotides.

4. 'Non-standard 3' & 5' ': The exon doesn't have any of the 2 expected endings and start of the flanking introns..

A 5th and 6th tags have been proposed that indicate a higher degree of manual curation:

5. 'OnlyExon': When the exon does not possess intronic regions because its length is marked the same as the transcript that it belongs.

6. 'Error-': The calculations might fail due to erroneous or missing data, inconsistent lengths, in these cases an extra start of the tag indicates it.

The following code block shows the approach taken to tag the exon. This is the script that runs on the server side and persist the data in a simple if-else manner.

```
$genes = EnsemblGene::with('ensemblTranscripts')->get();
foreach($genes as $gene){
    foreach ($gene->ensemblTranscripts as $transcript){
        $tSeq = str_split($transcript->genomic_sequence);
        $nExons = count($transcript->ensemblExon);
        if($nExons>1)
            foreach ($transcript->ensemblExon as $exon){
                $eLength = $exon->end - $exon->start;
                if($gene->strand == '1'){
                    $posStart = $exon->start - $transcript->start;
                    $posEnd = $posStart + $eLength;}
                else{
                    $posStart = $transcript->end - $exon->end;
                    $posEnd = $transcript->end - $exon->start;}
                if($exon->pivot->exon_number == 1){$diNuc5='--';}
                if($exon->pivot->exon_number == $nExons){$diNuc3='--';}
                if ($diNuc5 == $intronEnd || $diNuc5 == '--') $splicingStart = TRUE;
                else $splicingStart = FALSE;
                if ($diNuc3 == $intronStart || $diNuc3 == '--') $splicingEnd = TRUE;
                else $splicingEnd = FALSE;

                if ($splicingStart && $splicingEnd) $tag = $tags[0];
                elseif(!$splicingStart && $splicingEnd) $tag = $tags[1];
                elseif($splicingStart && !$splicingEnd) $tag = $tags[2];
                else $tag = $tags[3];

                if($transcript->ensemblExon()->updateExistingPivot($exon->ensembl_exon_id,[
                    'tag' => $tag,
                    'nucleotides5'=>$diNuc5,
                    'nucleotides3'=>$diNuc3
                    ]))};
```

3.9.1 Manual annotations

CAbase user management system allows three different profiles, administrator, guest and researcher. Researchers are the profiles that can perform manual annotations over different entities in the system. These annotations undergo a process from when they are created, until they are public information to guest and general public of the system.

At first, the researcher browses the entity that wishes to annotate, this can be achieved from any of the previously discussed views of the system, in particular, the exon analysis related views

As described before, this can only be performed by *researcher* users. In any entity or view that a change in the information wants to be perform, it will undergo a validation process of the annotation.

Where the user initiates the process and the data is persisted with a status of *new* annotation. This marks the start on which the administrator has to take action and the new data proposed can have the following status:

a) Accepted: Data in this status is reflected for all users
b) Rejected: For which no further interaction takes place.
c) More info: For this status, the researcher provides more information.

The manual annotation process is considered complete when the modified data is set to either a) or b).

Chapter 4

RESULTS AND DISCUSSION

In the present work, the results obtained using CAbase for different exon analyses are described in this chapter, utilizing graphical and programming codes as resources. The resulting Exon analysis tool, uses the features previously described. Firstly, a review of the present CAs in the system and how the information in CAbase relates to current literature. Second results obtained in selected CAs for their Exon analysis view.

4.1. Carbonic Anhydrases

Of the total of 47 CAs as genesis for the dataset can be summarized in the following table, in which the amount of orthologs is shown, this amount is calculated with the relationships ‘One to one’ and ‘One to many’, since from these homologues is possible to extract more meaningful information for this analysis, still, CAbase holds the 3 types of ortholog.

Table 4-1 Source CAs and their orthologs

	CA gene	StableID	Orthologs
Human	CA1	ENSG00000133742	156
	CA2	ENSG00000104267	196
	CA3	ENSG00000164879	249
	CA4	ENSG00000167434	453
	CA5A	ENSG00000174990	109
	CA5B	ENSG00000169239	302
	CA6	ENSG00000131686	286
	CA7	ENSG00000168748	294
	CA8	ENSG00000178538	293
	CA9	ENSG00000107159	176
	CA10	ENSG00000154975	206
	CA11	ENSG00000169605	160
	CA12	ENSG00000074410	280
	CA13	ENSG00000185015	187
	CA14	ENSG00000118298	262
Mouse	Car1	ENSMUSG00000027556	168
	Car2	ENSMUSG00000027562	211
	Car3	ENSMUSG00000027559	264
	Car4	ENSMUSG00000000805	468
	Car5a	ENSMUSG00000025317	124
	Car5b	ENSMUSG00000031373	317
	Car6	ENSMUSG00000028972	301
	Car7	ENSMUSG00000031883	294
	Car8	ENSMUSG00000041261	308
	Car9	ENSMUSG00000028463	191
	Car10	ENSMUSG00000056158	221
	Car11	ENSMUSG00000003273	137
	Car12	ENSMUSG00000032373	295
	Car13	ENSMUSG00000027555	201
	Car14	ENSMUSG00000038526	277
	Car15	ENSMUSG00000090236	235
Zebrafish	cahz	ENSDARG00000011166	104
	ca2	ENSDARG00000014488	79
	ca4b	ENSDARG00000042293	275
	ca4a	ENSDARG00000043589	255
	ca4c	ENSDARG00000044512	264
	ca5a	ENSDARG00000101778	306
	ca6	ENSDARG00000056499	287
	ca7	ENSDARG00000045139	294
	ca8	ENSDARG00000039098	293
	ca10a	ENSDARG00000052644	286
	ca10b	ENSDARG00000009568	286
	ca12	ENSDARG00000045644	279
	ca14	ENSDARG00000061697	257
	ca15a	ENSDARG00000015654	47
	ca15b	ENSDARG00000040510	68
	ca15c	ENSDARG00000002259	47

Also, due to the known differences among classes of CA, the results have been grouped by location of the CAs isozymes.

Table 4-2 CA distribution by attribute

	Peptide chain (aa)			Gene Length (nt)		
Activity	avg()	max()	min()	avg()	max()	min()
Active	291,6	890	67	19 333,04	266907	219
Inactive	260,2	1100	79	145 937,42	1345934	240
Location						
Cytoplasmic	253,6	627	71	18 083,23	266907	219
Mitochondrial	284,5	890	67	27 958,32	227238	240
Transmembrane	330,0	678	79	20 971,99	266907	240
GPI-anchored	294,6	876	96	10 826,80	206163	532
Secreted	367,6	704	78	17 075,16	80351	576

Interestingly, the gene length for inactive CAs (CARPs) is considerably greater than those active, the biggest gene being ENSOARG00020021387, from domestic sheep.

Table 4-3 CA isoforms with their average and maximum length of peptides (aa) and nucleotides (nt)

Isoform	Location	AVG aa	MAX aa	AVG nt	MAX nt
CA I	Cytoplasmic	240,78	627	20 335,36	266 907
CA II	Cytoplasmic	253,68	393	15 167,71	266 907
CA III	Cytoplasmic	262,16	527	23 504,53	189 013
CA IV	GPI-anchored	294,57	876	10 826,80	206 163
CA V A	Mitochondrial	277,66	726	23 114,54	200 314
CA V B	Mitochondrial	289,34	890	31 341,56	227 238
CA VI	Secreted	367,58	704	17 075,16	80 351
CA VII	Cytoplasmic	255,06	527	10 926,88	72 227
CARP VIII	Cytoplasmic	269,82	527	51 357,97	210 888
CA IX	Transmembrane	321,25	678	10 854,76	266 907
CARP X	Cytoplasmic	285,24	383	353 009,50	1 345 934
CARP XI	Cytoplasmic	192,21	1 100	6 094,63	27 506
CA XII	Transmembrane	339,87	578	34 055,19	223 441
CA XIII	Cytoplasmic	260,25	527	21 555,73	266 907
CA XIV	Transmembrane	325,93	527	13 930,12	117 423

A more detailed view in a per-isozyme table allows to see how transmembrane CAs are longer in comparison with the rest of CA types, with the exception of the only secreted CA, due to the pentraxin domain (Patrikainen et al., 2017).

4.2. Exon Analysis

This view is the development result of combining the different entities and their attributes. In this table view, each row of the table is a transcript, a *detailed-row* approach has been implemented. The following schematic shows the different sections that this view holds. Next, each section is described.

Figure 4-1 **Schematic representation of Exon Analysis table,** when expanded, each transcript will show 3 distinctive sections, (1) Exon table, (2) Pairwise Alignment and (3) Exon distribution prediction

The Exon table section shows the following information, the exon number, the linkable exon ID, the base pair length the splicing tag, the 5' and 3' dinucleotides of the flanking intronic zones where it applies and finally, the predicted translation per exon.

In the pairwise alignment section, the aligned peptide sequence visualization utilizing MView (MView) between the selected transcript aligned against the primary protein (the canonical transcript) of the selected source CA for the view. Alignment has been calculated using CLUSTAL W. The alignment is shown using the default colouring of MView based on identity.

The exon distribution prediction has been designed following patterns that are not common in other bioinformatic tools, trying to adjust the visualization for a better understanding of their distribution against a prototype transcript. The prototype transcript is the canonical transcript for the selected CA gene.

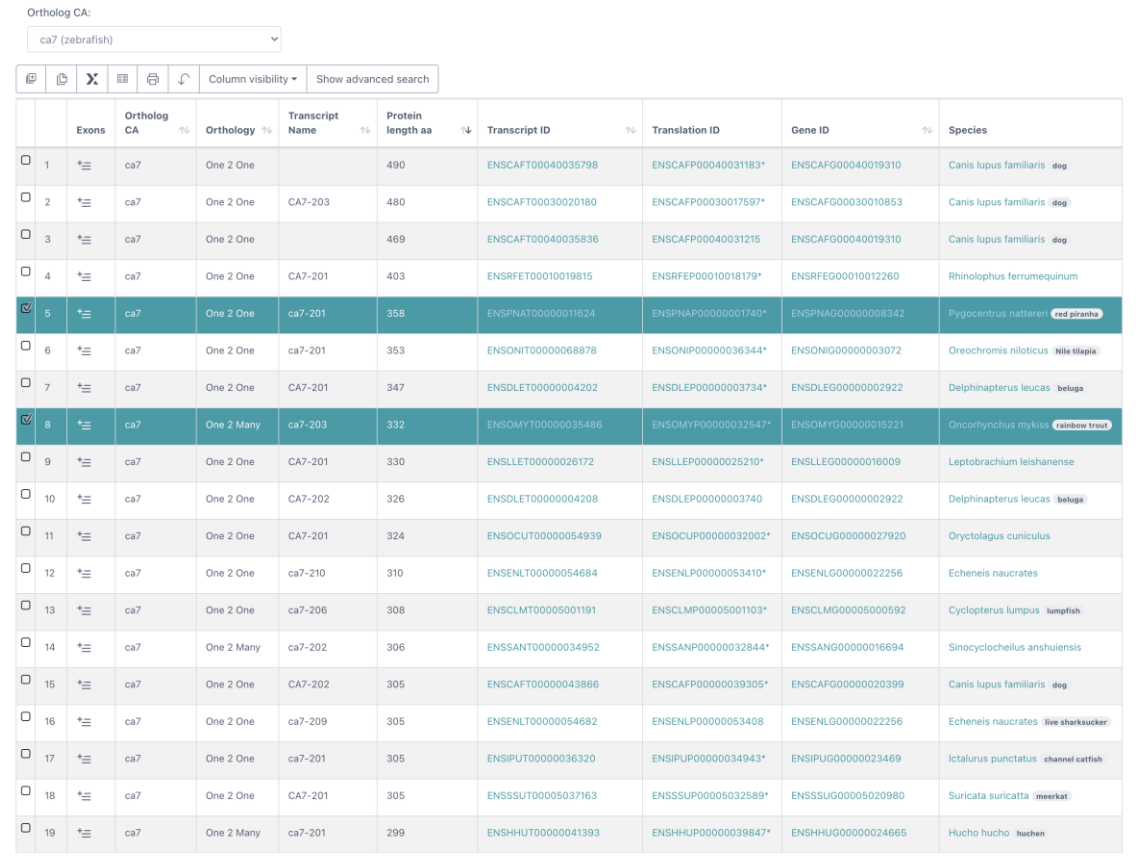
This Exon table allows to perform annotations over the whole transcript as well as per exon, taking the user to the previously described editing view.

Lastly, this view calculates the order of the exons based in their assorted starting points, depending on the strand. The response body from ensembl's *lookup endpoint* does not include the exon number for the transcript as an attribute, this information can be only derived from the index ordering of the object. Given that the response object ()The ordering usually matches with the index of the response. Nevertheless, cases where found in which following this pattern, the exon numbering was reversed

A complete view that took the Homology-based approach and agglomerated the information in a table as shown in this chapter.

The main table view possesses different sorting, search and filtering in order to best find the matches needed by the researcher. In the advanced search functionality different filtering options allow to only show and search within certain fields, possible to focus in desired attributes, i.e., the orthology type or only canonical translations per species.

This table is the first step in order to proceed onto an Exon analysis. This analysis is performed on a Transcript based set. From the previous table described in Figure 4-1, column 3 is a *button* that opens up the Exon analysis view per row.



		Exons	Ortholog CA	Orthology	Transcript Name	Protein length aa	Transcript ID	Translation ID	Gene ID	Species
☐	1		ca7	One 2 One		490	ENSCAFT00040035798	ENSCAFP00040031183*	ENSCAFG00040019310	Canis lupus familiaris dog
☐	2		ca7	One 2 One	CA7-203	480	ENSCAFT00030020180	ENSCAFP00030017597*	ENSCAFG00030010853	Canis lupus familiaris dog
☐	3		ca7	One 2 One		469	ENSCAFT00040035836	ENSCAFP00040031215	ENSCAFG00040019310	Canis lupus familiaris dog
☐	4		ca7	One 2 One	CA7-201	403	ENSRFET00010019815	ENSRFEP00010018179*	ENSRFEG00010012260	Rhinolophus ferrumequinum
☑	5		ca7	One 2 One	ca7-201	358	ENSPNAT00000011624	ENSPNAP00000001740*	ENSPNAG00000008342	Pygocentrus nattereri red piranha
☐	6		ca7	One 2 One	ca7-201	353	ENSONIT00000068878	ENSONIP00000036344*	ENSONIG00000003072	Oreochromis niloticus Nile tilapia
☐	7		ca7	One 2 One	CA7-201	347	ENSDLET00000004202	ENSDLEP00000003734*	ENSDLEG00000002922	Delphinapterus leucas beluga
☑	8		ca7	One 2 Many	ca7-203	332	ENSOMYT00000035486	ENSOMYP00000032547*	ENSOMYG00000015221	Oncorhynchus mykiss rainbow trout
☐	9		ca7	One 2 One	CA7-201	330	ENSLLET00000026172	ENSLLEP00000025210*	ENSLLEG00000016009	Leptobrachium leishanense
☐	10		ca7	One 2 One	CA7-202	326	ENSDLET00000004208	ENSDLEP00000003740	ENSDLEG00000002922	Delphinapterus leucas beluga
☐	11		ca7	One 2 One	CA7-201	324	ENSOCUT00000054939	ENSOCUP00000032002*	ENSOCUG00000027920	Oryctolagus cuniculus
☐	12		ca7	One 2 One	ca7-210	310	ENSENLT00000054684	ENSENLP00000053410*	ENSENLG00000022256	Echeneis naucrates
☐	13		ca7	One 2 One	ca7-206	308	ENSCLMT00005001191	ENSCLMP00005001103*	ENSCLMG00005000592	Cyclopterus lumpus lumpfish
☐	14		ca7	One 2 Many	ca7-202	306	ENSSANT00000034952	ENSSANP00000032844*	ENSSANG00000016694	Sinocyclocheilus anshuiensis
☐	15		ca7	One 2 One	CA7-202	305	ENSCAFT00000043866	ENSCAFP00000039305*	ENSCAFG00000020399	Canis lupus familiaris dog
☐	16		ca7	One 2 One	ca7-209	305	ENSENLT00000054682	ENSENLP00000053408	ENSENLG00000022256	Echeneis naucrates live sharksucker
☐	17		ca7	One 2 One	ca7-201	305	ENSIPUT00000036320	ENSIPUP00000034943*	ENSIPUG00000023469	Ictalurus punctatus channel catfish
☐	18		ca7	One 2 One	CA7-201	305	ENSSSUT00005037163	ENSSSUP00005032589*	ENSSSUG00005020980	Suricata suricatta meerkat
☐	19		ca7	One 2 Many	ca7-201	299	ENSHHUT00000041393	ENSHHUP00000039847*	ENSHHUG00000024665	Hucho hucho huchen

Figure 4-2 Main Exon Analysis view, this example shows each row an ortholog for *ca7 (zebrafish)* [stabID: *ENSDARG00000045139].* The table shows 19 out of 311 found orthologues, sorted by protein length. Multiple selection shows *red piranha* and *rainbow trout selected. Note: due to their length, some common name species have been removed for representational purposes.*

4.2.1 Exon Analysis Case

The following results were obtained by defining a CA of interest, the GPI-anchored CA4, Using *Car4* as a source for the homology, choosing the orthology type as 'One to one'

Nr.	Exon id	Length	Splicing	5'	3'	Predicted translation
1	ENSLOCE00000060269	54 (bp)	Splicing OK	--	GT	RF:1 18 MRAVLSLLTL LCSAPGAA
2	ENSLOCE00000060272	26 (bp)	Non-Standard 3' & 5'	GA	GG	RF:3 8 KKDRSYEF
3	ENSLOCE00000060275	32 (bp)	Non-Standard 3'	AG	GG	RF:3 10 LESHEYPISS
4	ENSLOCE00000060277	164 (bp)	Splicing OK	AG	GT	RF:3 54 PAHWADQHPY CGGRRQSPIN VVTRKAQYDS SLEPFTFEGY 40 DQTHSVTAEN LGHS
5	ENSLOCE00000060280	145 (bp)	Splicing OK	AG	GT	RF:3 48 HFALSSSVRT QGGHMPDWYK AVNFHLHWGN EAGPGSEHTI 40 DGEQFPME
6	ENSLOCE00000060288	53 (bp)	Splicing OK	AG	GT	RF:1 18 LNTKQSVNTT KHIMGGNV
7	ENSLOCE00000060306	27 (bp)	Non-Standard 3'	AG	CT	RF:1 9 LEQCMNALA
8	ENSLOCE00000060310	80 (bp)	Non-Standard 5'	AC	GT	RF:3 26 LSLQESSQDN PYYKVLIDAL DEVKHA
9	ENSLOCE00000060316	163 (bp)	Splicing OK	AG	GT	RF:3 54 NRTIITALQL VSILPKPQDL RKYYYYSGSV TVPDCDEAVV 40 WAVFETPIRI SREQ
10	ENSLOCE00000060328	107 (bp)	Splicing OK	AG	--	RF:1 36 LTAFSQKLLF RTEKPMKMTF RPTQPLNGRV VLRSSA

Figure 4-3 Exon table details for transcript ENSLOCT00000006442

The exon annotation based on their cleavage sites, shows that 4 exons have non-standard intron termination, start or both. This result, aligns with the predicted exon distribution as shown in the following figure:

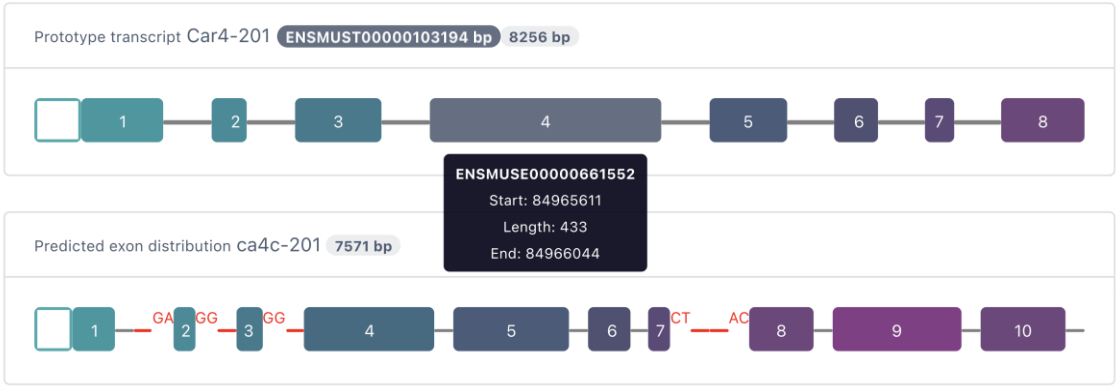


Figure 4-4 Exon distribution prediction for *ENSLOCT00000006442* against *ENSMUST00000103194*. Note: Mouse is hovering over the prototype's exon Nr. 4, denoting the difference in length for the same numbered exon.

As for the pairwise alignment, it was possible to observe that such prediction for this particular transcript the expected results were based in known base pairs that usually act in the splicing machinery. With a $\%_{id}$ of only 27%.

Figure 4-5 Pairwise Alignment for gene *ENSMUSG00000000805* with *ENSLOCG00000005339*.

As indicated by (Dou et al., 2006)., identifying exon cleavage zones can serve as an indicator for Alternative splicing Transcripts. This can be summarized by the following table where all the present exons in CAbase have been annotated with the splicing tag described in 3.9:

Table 4-4 Exon tagging results

Tag	Exons
Splicing OK	48 891
Non-Standard 5'	460
Non-Standard 3' & 5'	294
Non-Standard 3'	2 535
OnlyExon	17

From this, the distribution of splicing cites corresponds to the one by. yet, it begs the attention the tag on the 3-prime end, that stands with a 5,42% (considering the exons tagged on both ends). These exons are grouped by species and by doing so, this tagging could be attributed to the quality of the genomes.

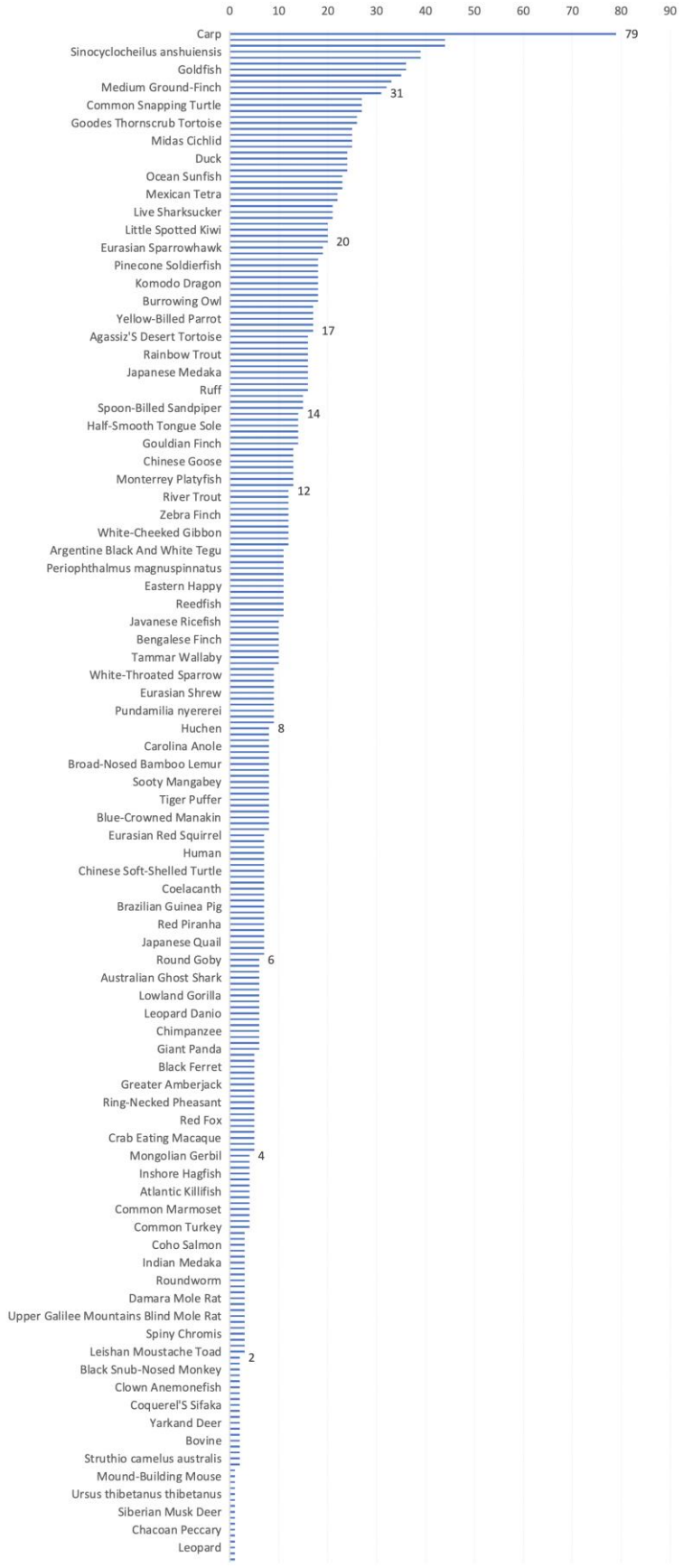


Figure 4-6 Frequencies of Non-Standard 3' Exon cleavage sites by nucleotide comparison

Chapter 5

CONCLUSIONS

In the present work, a new Exon analysis tool for the Carbonic Anhydrase protein family has been proposed. This novel approach utilises a web platform a novel tool for this type of analysis. Yet, it allows different user profiles to explore and centralise information regarding the carbonic anhydrases gene/protein family.

The usage of the solution described through this work, has been limited to a test environment, having produce no significant changes to the current Carbonic Anhydrases standardized information. In order to produce significant results, CAbase is set to be deployed and open to the scientific community, in particular to researchers in CAs.

The Exon analysis produced, both views and tools are arguably a good start in terms of the functionalities as a novel approach. The upcoming implementation of GFF3 will possibly be a major change in terms of the visualization of the exon alignment, producing results that are in line with current online major alignment tools, or the ones discussed in 2.1.4

Considering that the central work of this thesis project has been the Exon analysis part, the discussion should consider the designs of the platform, views, database

architecture as well as data sources. Is with this platform that the set of tools of CAbase can be expanded.

Although the platform has been conceived using web technologies that are not recurrent in daily use for bioinformatics analysis, the technologies chosen allow for both replicate this analysis and to incorporate external calls for such tools.

Where the present work fails is to discuss is the validity of the results obtained. Given the users constraints and the nature of the platform, a community of experts in this protein family coming together to create consensus data is required to validate the results obtained. This use case has not yet been implemented at the moment of writing this work, hence this has not produced meaningful data.

In second place, the results that are discussed in this chapter refer to the ones inherited to the design of the features and tools that CAbase has to offer to perform Exon Analysis as well as other projected usage.

5.1. Exon Analysis

The exon analysis made utilising CAbase can be divided into automated and manual work. In the first group the tools for exon tagging, the automated annotations that allow to deeper understand the splicing machinery, and once fully implemented, will allow to properly annotate the predicted (alternative) splicing of certain transcripts.

5.1.1 Exon tagging

As seen in the results section, different attempts were made on the exon tagging, to further explain the difference in the proportions where in the 3-prime end, 5% of all the present exons were tagged unexpectedly with a non-standard splicing site. Here these results are discussed.

First, for this discussion, the dinucleotides counted are the ones encountered in exons labelled as both ‘`Non-Standard 3''` and ‘`Non-Standard 3' & 5''` which are 2,743 and 298 out of 59,081 total exons annotated.

Second, the tools used for this analysis are direct SQL queries to the database, CAbase does not count with an interface for this type of analysis in a *case-by-case* basis.

Also, the exons that are discussed are the ones with valid genomic nucleotides. From the source of data, a near neglectable number of cases in which the sequence presented amino acids in their transcript's genomic sequence (13 exons in 4 transcripts). This will be further investigated. However, a bigger number of exons (or expected lengths of nucleotides positions) were encountered in non-determined regions, or regions with gaps, N-filled regions, 139 for this case.

Altogether, the CA-related exons of interest are 2,701. At the same time, the exons labelled for the 5' are 300.

The dinucleotides on the 3-prime had to be dissected by classes at their lineage level, resulting in the following table:

Table 5-1 Non-Standard 3' splicing per organism classes

Class	Species	%	3' labels	%2
Mammals	106	41 %	795	28 %
Fishes	80	31 %	995	35 %
Amphibians	2	1 %	8	0 %
Birds	49	19 %	740	26 %
Reptiles	19	7 %	344	12 %

Where, interestingly, Birds seem to contribute more to the number of exons labelled as non-standard splicing on the 3'. However, for the mammalian class:

Table 5-2 Dinucleotides on the 3' end for mammalian

Dinucleotides	Exons	%
AA	26	3 %
AC	21	3 %
AG	24	3 %
AT	22	3 %
CA	27	3 %
CC	26	3 %
CG	6	1 %
CT	39	5 %
GA	49	6 %
GC	**350**	**44 %**
GG	55	7 %
TA	52	7 %
TC	17	2 %
TG	51	6 %
TT	30	4 %

It was not possible to prove the already discussed expected ratio concluded by (Burset et al., 2000), since the dinucleotide pair GC presents in a great percentage. Nevertheless, this analysis brought the overall case of correct, or expected splicing label of a 94% to a 95,5%. Furthermore, be this analysis expanded to a greater degree of nucleotides, i.e. 6 as Dou et al., in 2006 proposed and the conclusions could be more significant.

5.1.2 Exon distribution prediction

The graphical interpretation of the exon distribution lacks of features inherited to the common exon representation. Intronic regions need to be better and more proportionately displayed. It is firmly believed that with the integration of the GFF3 file format to read exon positioning, the visualization will improve considerably. This in conjunction with the already implement front-end will allow a refreshed way of visualizing and navigating over the exon unit. As pointed out by (Movassat et al., 2019), the exon size (and not necessarily sequence) conservation, is a strong predictor of Alternative Splicing.

5.1.3 Pairwise Alignment

The Pairwise alignment available in CAbase, it can only be obtained against a prototype transcript, hence little flexibility for the analysis is provided by CAbase. A server-side implementation of pairwise alignment or even multiple sequence alignment is possible given the described modularized development.

5.2. The designed platform

Even in fields diametrically opposite, platforms of this type offer a sustainable and scalable solution, where online users can concurrently perform actions over centralised data. The creation of such tool in the present field, and narrowed it for a determined protein family is a solid first step in creating a complex and scalable platform that could provide the necessary means for a targeted group of scientific knowledge, to consolidate information and create consensus data.

Albeit CAbase performance has only been tested extensively in local environments and in a more seldom manner in a commercial hosting service, the amount of data does not propose any computational exigence thus far. It is the views themselves that can be evaluated in their performance, and it was found that although easier to program, Yajra's DataTables package comes short in comparison with the features of the original package (DataTables, s.f.). Even though Yajra's package continues to adapt the original jQuery version into Laravel, more complex sorting, searches, filtering and data display in general have a longer expected time to be implemented. A good improvement is to migrate to the original package which would allow more tables features, a good example of the usage of this package can be found in 2.1.3. This, together with the additional editor package (paid-package) or the CloudTables© (subscription) version would take the UI/UX of CAbase to an *eye-catching* usability and design.

First, a pipeline with a more detailed planning is required to produce a that can take CAbase to be an official TPA:inferal of database, and become a platform that can not only creates consensus data in a specific area such as Carbonic Anhydrases, but that can funnel information to well-established platforms that can broaden the knowledge here inferred, namely NCBI/GeneBank, EMBL-EBI/Ensembl itself.

The iterative process of creation of such platform proves to be implementable in other areas or protein families that deserve the centralisation of their ever-growing scientific research.

REFERENCES

Brinkman R, Margaria R, Meldrum NU, & Roughton FJW. (1932). The CO2 catalyst present in blood. *Journal of Physiology*, *75*, 3–4.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, *22*(22), 4673–4680. https://doi.org/10.1093/nar/22.22.4673

Davies, J. (1996). Origins and evolution of antibiotic resistance. In *Microbiología (Madrid, Spain)* (Vol. 12, Issue 1, pp. 9–16). Microbiol Mol Biol Rev. https://doi.org/10.1128/mmbr.00016-10

Nielsen, H., Engelbrecht, J., Brunak, S., & Von Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engineering*, *10*(1), 1–6. https://doi.org/10.1093/protein/10.1.1

Nielsen, H., Engelbrecht, J., Brunak, S., & von Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engineering Design and Selection*, *10*(1), 1–6. https://doi.org/10.1093/protein/10.1.1

Brown, N. P., Leroy, C., & Sander, C. (1998). MView: A web-compatible database search or multiple alignment viewer. *Bioinformatics*, *14*(4), 380–381. https://doi.org/10.1093/bioinformatics/14.4.380

Smith, K. S., Jakubzick, C., Whittam, T. S., & Ferry, J. G. (1999). Carbonic anhydrase is an ancient enzyme widespread in prokaryotes. *Proceedings of the National*

Academy of Sciences of the United States of America, *96*(26), 15184–15189. https://doi.org/10.1073/pnas.96.26.15184

Burset, M., Seledtsov, I. A., & Solovyev, V. V. (2000). Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Research*, *28*(21), 4364–4375. https://doi.org/10.1093/nar/28.21.4364

Burset, M., Seledtsov, I. A., & Solovyev, V. V. (2000). Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Research*, *28*(21), 4364–4375. https://doi.org/10.1093/nar/28.21.4364

Smith, K. S., & Ferry, J. G. (2000). Prokaryotic carbonic anhydrases. *FEMS Microbiology Reviews*, *24*(4), 335–366. https://doi.org/10.1111/j.1574-6976.2000.tb00546.x

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera - A visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, *25*(13), 1605–1612. https://doi.org/10.1002/jcc.20084

Hilvo, M., Tolvanen, M., Clark, A., Shen, B., Shah, G. N., Waheed, A., Halmi, P., Hänninen, M., Hämäläinen, J. M., Vihinen, M., Sly, W. S., & Parkkila, S. (2005). Characterization of CA XV, a new GPI-anchored form of carbonic anhydrase. *Biochemical Journal*, *392*(1), 83–92. https://doi.org/10.1042/BJ20051102

Fischer, M., Thai, Q. K., Grieb, M., & Pleiss, J. (2006). DWARF - A data warehouse system for analyzing protein families. *BMC Bioinformatics*, *7*. https://doi.org/10.1186/1471-2105-7-495

Dou, Y., Fox-Walsh, K. L., Baldi, P. F., & Hertel, K. J. (2006). Genomic splice-site analysis reveals frequent alternative splicing close to the dominant splice site. *RNA*, *12*(12), 2047–2056. https://doi.org/10.1261/rna.151106

Esbaugh, A. J., & Tufts, B. L. (2006). The structure and function of carbonic anhydrase isozymes in the respiratory system of vertebrates. *Respiratory Physiology and Neurobiology*, *154*(1–2), 185–198. https://doi.org/10.1016/j.resp.2006.03.007

Wang, C. K. L., Kaas, Q., Chiche, L., & Craik, D. J. (2008). CyBase: A database of cyclic protein sequences and structures, with applications in protein discovery and engineering. *Nucleic Acids Research*, *36*(SUPPL. 1), D206. https://doi.org/10.1093/nar/gkm953

Saari, S., Hilvo, M., Pan, P., Gros, G., Hanke, N., Waheed, A., Sly, W. S., & Parkkila, S. (2010). The most recently discovered carbonic anhydrase, CA XV, is expressed in the thick ascending limb of Henle and in the collecting ducts of mouse kidney. *PLoS ONE*, *5*(3). https://doi.org/10.1371/journal.pone.0009624

Löytynoja, A., & Goldman, N. (2010). webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. *BMC Bioinformatics*, *11*, 579. https://doi.org/10.1186/1471-2105-11-579

Alterio, V., Di Fiore, A., D'Ambrosio, K., Supuran, C. T., & De Simone, G. (2012). Multiple binding modes of inhibitors to carbonic anhydrases: How to design specific drugs targeting 15 different isoforms? In *Chemical Reviews* (Vol. 112, Issue 8, pp. 4421–4468). American Chemical Society. https://doi.org/10.1021/cr200176r

Malentacchi, F., Vinci, S., Melina, A. Della, Kuncova, J., Villari, D., Giannarini, G., Nesi, G., Selli, C., & Orlando, C. (2012). Splicing variants of carbonic anhydrase IX in bladder cancer and urine sediments. *Urologic Oncology: Seminars and Original Investigations*, *30*(3), 278–284. https://doi.org/10.1016/j.urolonc.2010.05.009

Federhen, S. (2012). The NCBI Taxonomy database. *Nucleic Acids Research*, *40*(D1), D136. https://doi.org/10.1093/nar/gkr1178

Aspatwar, A., Tolvanen, M. E. E., & Parkkila, S. (2013). An update on carbonic anhydrase-related proteins VIII, X and XI. In *Journal of Enzyme Inhibition and Medicinal Chemistry* (Vol. 28, Issue 6, pp. 1129–1142). J Enzyme Inhib Med Chem. https://doi.org/10.3109/14756366.2012.727813

Tolvanen, M. E. E., Ortutay, C., Barker, H. R., Aspatwar, A., Patrikainen, M., & Parkkila, S. (2013). Analysis of evolution of carbonic anhydrases IV and XV reveals a rich history of gene duplications and a new group of isozymes. *Bioorganic and Medicinal Chemistry*, *21*(6), 1503–1510. https://doi.org/10.1016/j.bmc.2012.08.060

Boone, C. D., Habibzadegan, A., Gill, S., & Mckenna, R. (2013). Carbonic anhydrases and their biotechnological applications. In *Biomolecules* (Vol. 3, Issue 3, pp. 553–562). MDPI AG. https://doi.org/10.3390/biom3030553

Datovo, A., & Vari, R. P. (2013). The Jaw Adductor Muscle Complex in Teleostean Fishes: Evolution, Homologies and Revised Nomenclature (Osteichthyes: Actinopterygii). *PLoS ONE*, *8*(4). https://doi.org/10.1371/journal.pone.0060846

González, J. M., & Fisher, S. Z. (2014). Carbonic Anhydrases in Industrial Applications. In *Sub-cellular biochemistry* (Vol. 75, pp. 405–426). Subcell Biochem. https://doi.org/10.1007/978-94-007-7359-2_20

Health Sciences Library Systems. (2014). Online Bioinformatics Resources Collection Individual protein families. (The Health Sciences Library System) Retrieved from https://www.hsls.pitt.edu/obrc/index.php?page=individual_protein_families

Löytynoja, A. (2014). Phylogeny-aware alignment with PRANK. *Methods in Molecular Biology*, *1079*, 155–170. https://doi.org/10.1007/978-1-62703-646-7_10

Supuran, C. T., & De Simone, G. (2015). Carbonic Anhydrases as Biocatalysts: From Theory to Medical and Industrial Applications. In *Carbonic Anhydrases as Biocatalysts: From Theory to Medical and Industrial Applications*. Elsevier Inc. https://doi.org/10.1016/C2012-0-13548-1

De Luca, V., Del Prete, S., Supuran, C. T., & Capasso, C. (2015). Protonography, a new technique for the analysis of carbonic anhydrase activity. *Journal of Enzyme Inhibition and Medicinal Chemistry*, *30*(2), 277–282. https://doi.org/10.3109/14756366.2014.917085

Ferreira-Martins, D., McCormick, S. D., Campos, A., Lopes-Marques, M., Osório, H., Coimbra, J., Castro, L. F. C., & Wilson, J. M. (2016). A cytosolic carbonic anhydrase molecular switch occurs in the gills of metamorphic sea lamprey. *Scientific Reports*, *6*. https://doi.org/10.1038/srep33954

Isokangas, L. (2016). *Development of CAbase and an Exon Analysis Pipeline for Visual assessment of Predicted Genes for the Carbonic Anhydrases*. https://trepo.tuni.fi/handle/10024/99277

Isberg, V., Mordalski, S., Munk, C., Rataj, K., Harpsøe, K., Hauser, A. S., Vroling, B., Bojarski, A. J., Vriend, G., & Gloriam, D. E. (2016). GPCRdb: An information system for G protein-coupled receptors. *Nucleic Acids Research*, *44*(D1), D356–D364. https://doi.org/10.1093/nar/gkv1178

Jin, S., Sun, J., Wunder, T., Tang, D., Cousins, A. B., Sze, S. K., Mueller-Cajar, O., & Gao, Y. G. (2016). Structural insights into the LCIB protein family reveals a new group of β-carbonic anhydrases. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(51), 14716–14721. https://doi.org/10.1073/pnas.1616294113

Kikutani, S., Nakajima, K., Nagasato, C., Tsuji, Y., Miyatake, A., & Matsuda, Y. (2016). Thylakoid luminal Θ-carbonic anhydrase critical for growth and photosynthesis in the marine diatom Phaeodactylum tricornutum. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(35), 9828–9833. https://doi.org/10.1073/pnas.1603112113

Patrikainen, M. S., Tolvanen, M. E. E., Aspatwar, A., Barker, H. R., Ortutay, C., Jänis, J., Laitaoja, M., Hytönen, V. P., Azizi, L., Manandhar, P., Jáger, E., Vullo, D., Kukkurainen, S., Hilvo, M., Supuran, C. T., & Parkkila, S. (2017). Identification and characterization of a novel zebrafish (Danio rerio) pentraxin-carbonic anhydrase. *PeerJ*, *2017*(12). https://doi.org/10.7717/peerj.4128

Hauser, A. S., Attwood, M. M., Rask-Andersen, M., Schiöth, H. B., & Gloriam, D. E. (2017). Trends in GPCR drug discovery: New agents, targets and indications.

Nature Reviews Drug Discovery, *16*(12), 829–842. https://doi.org/10.1038/nrd.2017.178

Bray, T. (2017). *The JavaScript Object Notation (JSON) Data Interchange Format*. https://doi.org/10.17487/RFC8259

Gourlé, H. (2019, 08 07). *Taxadb*. Retrieved from https://taxadb.readthedocs.io

DataTables. (n.d.). *DataTables, Add advanced interaction controls to your HTML tables*. (SpryMedia Ltd) Retrieved from www.datatables.net

Jensen, E. L., Clement, R., Kosta, A., Maberly, S. C., & Gontero, B. (2019). A new widespread subclass of carbonic anhydrase in marine phytoplankton. *ISME Journal*, *13*(8), 2094–2106. https://doi.org/10.1038/s41396-019-0426-8

Armenteros, J. J. A., Salvatore, M., Emanuelsson, O., Winther, O., Von Heijne, G., Elofsson, A., & Nielsen, H. (2019). Detecting sequence signals in targeting peptides using deep learning. *Life Science Alliance*, *2*(5). https://doi.org/10.26508/lsa.201900429

Movassat, M., Forouzmand, E., Reese, F., & Hertel, K. J. (2019). Exon size and sequence conservation improves identification of splice-altering nucleotides. *RNA*, *25*(12), 1793–1805. https://doi.org/10.1261/rna.070987.119

Del Prete, S., Nocentini, A., Supuran, C. T., & Capasso, C. (2020). Bacterial ι-carbonic anhydrase: a new active class of carbonic anhydrase identified in the genome of the Gram-negative bacterium Burkholderia territorii. *Journal of Enzyme Inhibition and Medicinal Chemistry*, *35*(1), 1060–1068. https://doi.org/10.1080/14756366.2020.1755852

Ozensoy Guler, O., Supuran, C. T., & Capasso, C. (2020). Carbonic anhydrase IX as a novel candidate in liquid biopsy. In *Journal of Enzyme Inhibition and Medicinal Chemistry* (Vol. 35, Issue 1, pp. 255–260). Taylor and Francis Ltd. https://doi.org/10.1080/14756366.2019.1697251

Kooistra, A. J., Mordalski, S., Gásp´gáspár, G., Andy-Szekeres, P. ´, Esguerra, M., Mamyrbekov, A., Munk, C., Gy¨, G., Keser, G. M., & Gloriam, D. E. (2020). GPCRdb in 2021: integrating GPCR sequence, structure and function. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkaa1080

Kooistra, A. J., Mordalski, S., Gásp´gáspár, G., Andy-Szekeres, P. ´, Esguerra, M., Mamyrbekov, A., Munk, C., Gy¨, G., Keser, G. M., & Gloriam, D. E. (2020). OUP accepted manuscript. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkaa1080

Kumari, E., & Kline, W. (2020). *Captive-Portal Identification in DHCP and Router Advertisements (RAs)*. https://doi.org/10.17487/RFC8910

Veidenberg, A., & Löytynoja, A. (2021). Evolutionary Sequence Analysis and Visualization with Wasabi. In *Methods in molecular biology (Clifton, N.J.)* (Vol. 2231, pp. 225–240). Methods Mol Biol. https://doi.org/10.1007/978-1-0716-1036-7_14

APPENDIX 1

CABASE INSTALLATION

In this appendix, the installation of CAbase is covered. From cloning the Bitbucket repository, to access the portal in a new deployment. Also, the package dependency manager for both front- and back-end, npm and composer respectively, are covered.

Steps to deploy CAbase

1. Clone the repository:

```
1  git@bitbucket.org:NachoSinColor/cabase.git
```

2. create .env file in server directory

```
2  cp .env.example .env
```

3. edit .env and add environment variables needed, i.e., mysql credentials

```
APP_NAME=CAbase
APP_ENV=local
APP_KEY=base64:/gZPFx67GXsm3h9xjEThfBUJGUrJDYrhtV/ss+M3LC4=
APP_DEBUG=true
APP_URL=http://cabase.app

LOG_CHANNEL=stack

DB_CONNECTION=mysql
DB_HOST=localhost
DB_PORT=3306
DB_DATABASE=CAbase
DB_USERNAME=root
DB_PASSWORD=secret
```

4. The following two options are for local development. For production server deployment skip this step.

A) With Laravel Homestead

Laravel Homestead is a ready for development virtual machine that includes all the necessary applications in an Ubuntu distribution package.

1. Install vagrant
2. Install virtual box (or other VM provider)
3. Install Homestead
4. Add domains to host file
5. Run Vagrant machine:

```
vagrant up (--provision for the first run or if the file has been modified)
```

B) Laravel valet

Similar to Homestead, Valet is a lightweight approach for MacOS systems

1. Install Valet
2. Terminal in the CAbase-app/public folder

```
valet park;

valet link;

valet open;
```

5. Next, the necessary Laravel configuration continues

```
composer install

php artisan key:generate

php artisan config:cache

composer dump-autload

npm install // install packages.json

npm run dev // to create assets and public content
```

6. Migrate the database (create the tables)

```
php artisan migrate
```

7. Seed the database (feed initial data)

Feed initial data such as Taxonomy information, Initial CAs (Human, Mouse, Zebrafish)

```
php artisan db:seed
```

8. Taxonomy table initialization

```
php artisan taxa:build
```

9. go to the configured URL. i.e., cabase.local/login
   user:admin
   pass:password
10. These steps allow the basic initialization of CAbase, allowing to explore routes views and getting acquainted with the system in general.

APPENDIX 2

CABASE DATABASE MODEL

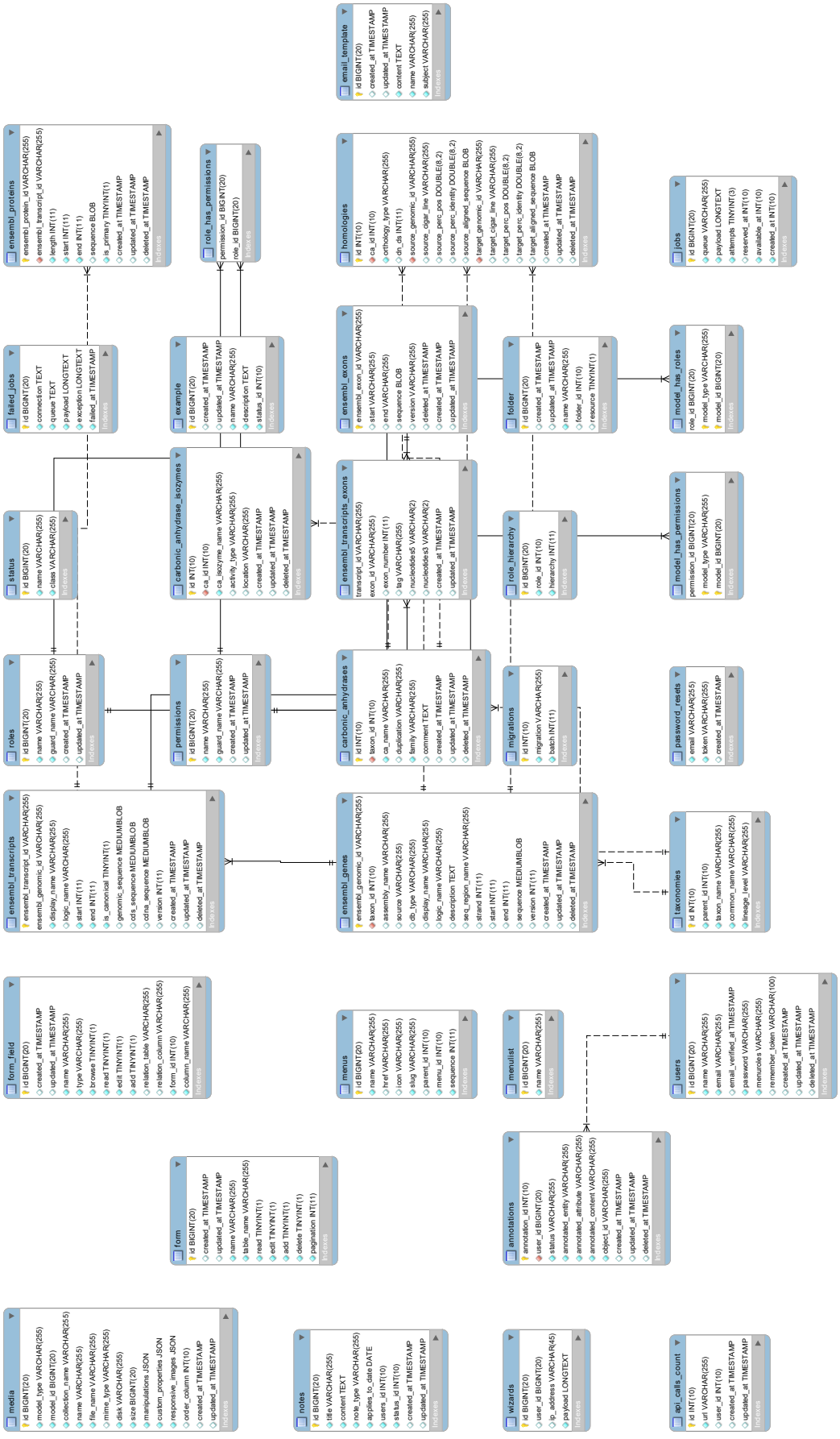


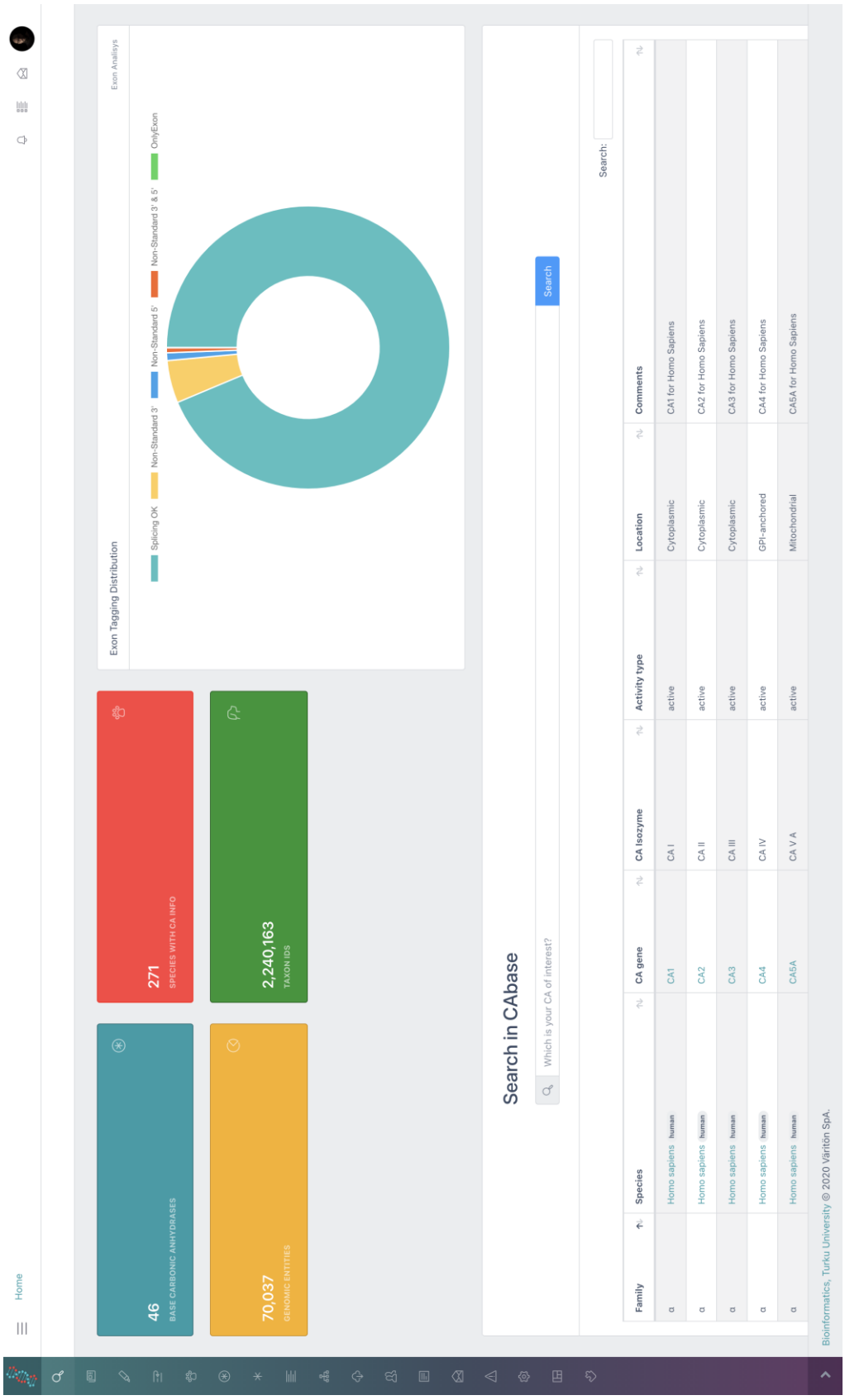
Figure 0-1 CAbase Entity Relationship diagram

APPENDIX 3

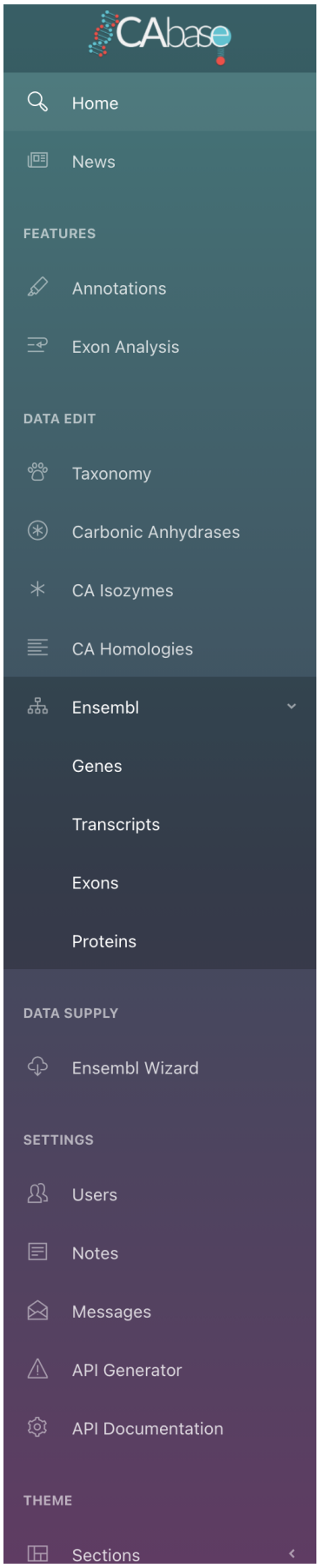
CABASE DATABASE UI

CAbase User interface is composed by the following list of views. This appendix is generated with administrator privileges, accessing to all the data available. The exon view is excluded from this appendix as it was cover extensively in Chapter 3 and 4.

5.3. Homepage


Home
Exon Analisys
Exon Tagging Distribution
Spliceing OK
Non-Standard 3'
Non-Standard 5'
Non-Standard 3' & 5'
OnlyExon
271
SPECIES WITH CA INFO
2,240,163
TAXON IDS
46
BASE CARBONIC ANHYDRASES
70,037
GENOMIC ENTITIES
Search in CAbase
Which is your CA of interest?
Search
Search:
Family
Species
CA gene
CA Isozyme
Activity type
Location
Comments
α
Homo sapiens human
CA1
CA I
active
Cytoplasmic
CA1 for Homo Sapiens
α
Homo sapiens human
CA2
CA II
active
Cytoplasmic
CA2 for Homo Sapiens
α
Homo sapiens human
CA3
CA III
active
Cytoplasmic
CA3 for Homo Sapiens
α
Homo sapiens human
CA4
CA IV
active
GPI-anchored
CA4 for Homo Sapiens
α
Homo sapiens human
CA5A
CA V A
active
Mitochondrial
CA5A for Homo Sapiens
Bioinformatics, Turku University @ 2020 Variton SpA.

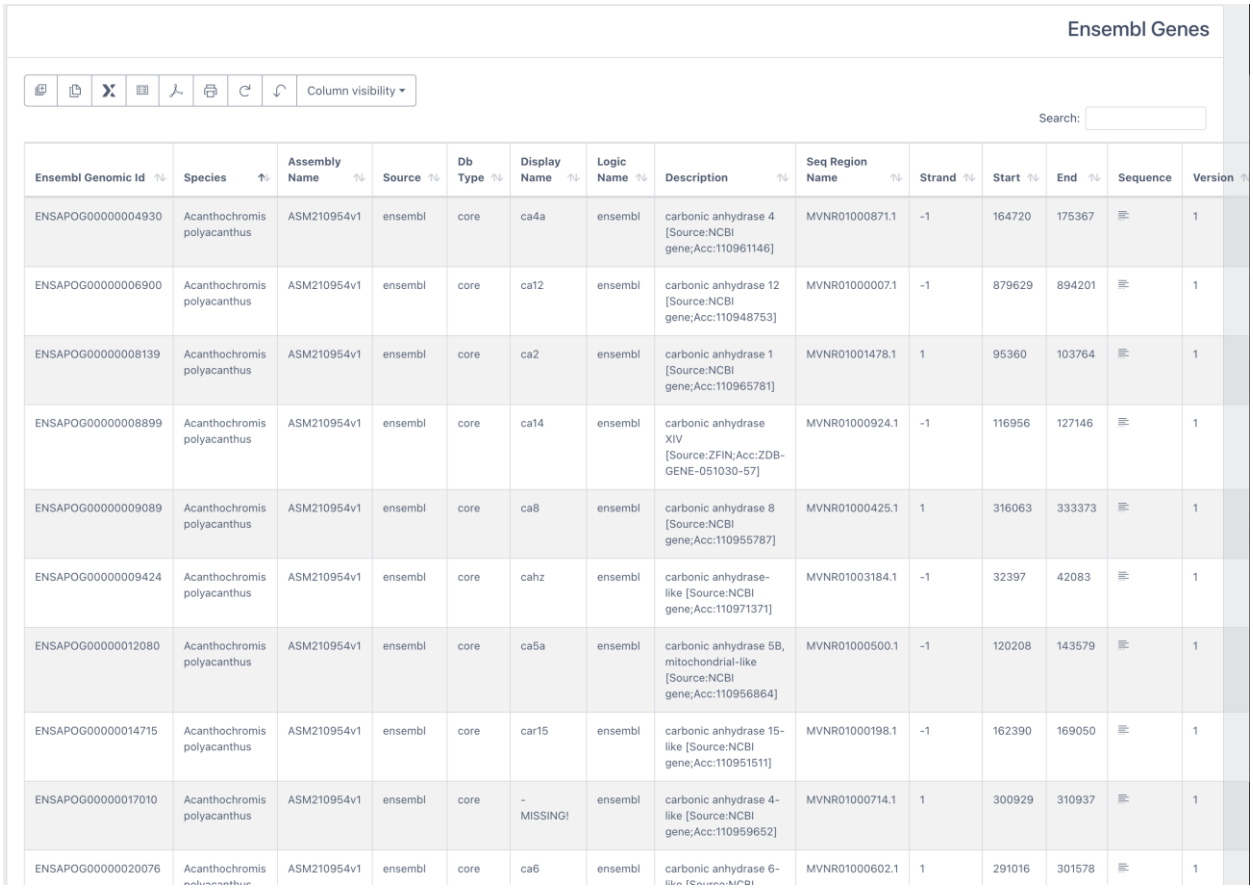
5.4. Side Menu


CAbase
Home
News
FEATURES
Annotations
Exon Analysis
DATA EDIT
Taxonomy
Carbonic Anhydrases
CA Isozymes
CA Homologies
Ensembl
Genes
Transcripts
Exons
Proteins
DATA SUPPLY
Ensembl Wizard
SETTINGS
Users
Notes
Messages
API Generator
API Documentation
THEME
Sections

5.5. User Management

5.6. Entity tables

5.6.1 Genes General View



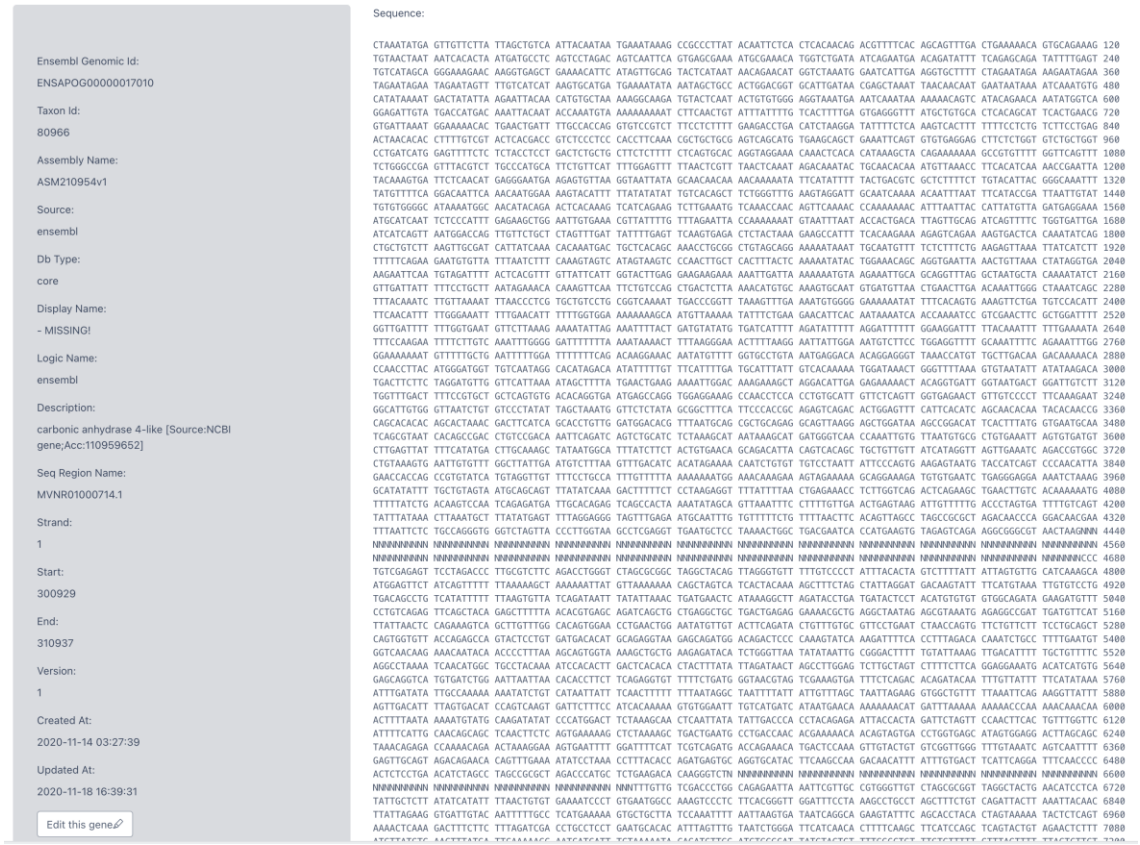
Ensembl Genes

Column visibility

Search:

Ensembl Genomic Id	Species	Assembly Name	Source	Db Type	Display Name	Logic Name	Description	Seq Region Name	Strand	Start	End	Sequence	Version
ENSAPOG00000004930	Acanthochromis polyacanthus	ASM210954v1	ensembl	core	ca4a	ensembl	carbonic anhydrase 4 [Source:NCBI gene;Acc:110961146]	MVNR01000871.1	-1	164720	175367		1
ENSAPOG00000006900	Acanthochromis polyacanthus	ASM210954v1	ensembl	core	ca12	ensembl	carbonic anhydrase 12 [Source:NCBI gene;Acc:110948753]	MVNR01000007.1	-1	879629	894201		1
ENSAPOG00000008139	Acanthochromis polyacanthus	ASM210954v1	ensembl	core	ca2	ensembl	carbonic anhydrase 1 [Source:NCBI gene;Acc:110965781]	MVNR01001478.1	1	95360	103764		1
ENSAPOG00000008899	Acanthochromis polyacanthus	ASM210954v1	ensembl	core	ca14	ensembl	carbonic anhydrase XIV [Source:ZFIN;Acc:ZDB-GENE-051030-57]	MVNR01000924.1	-1	116956	127146		1
ENSAPOG00000009089	Acanthochromis polyacanthus	ASM210954v1	ensembl	core	ca8	ensembl	carbonic anhydrase 8 [Source:NCBI gene;Acc:110955787]	MVNR01000425.1	1	316063	333373		1
ENSAPOG00000009424	Acanthochromis polyacanthus	ASM210954v1	ensembl	core	cahz	ensembl	carbonic anhydrase-like [Source:NCBI gene;Acc:110971371]	MVNR01003184.1	-1	32397	42083		1
ENSAPOG00000012080	Acanthochromis polyacanthus	ASM210954v1	ensembl	core	ca5a	ensembl	carbonic anhydrase 5B, mitochondrial-like [Source:NCBI gene;Acc:110956864]	MVNR01000500.1	-1	120208	143579		1
ENSAPOG00000014715	Acanthochromis polyacanthus	ASM210954v1	ensembl	core	car15	ensembl	carbonic anhydrase 15-like [Source:NCBI gene;Acc:110951511]	MVNR01000198.1	-1	162390	169050		1
ENSAPOG00000017010	Acanthochromis polyacanthus	ASM210954v1	ensembl	core	- MISSING!	ensembl	carbonic anhydrase 4-like [Source:NCBI gene;Acc:110959652]	MVNR01000714.1	1	300929	310937		1
ENSAPOG00000020076	Acanthochromis	ASM210954v1	ensembl	core	ca6	ensembl	carbonic anhydrase 6-	MVNR01000602.1	1	291016	301578		1

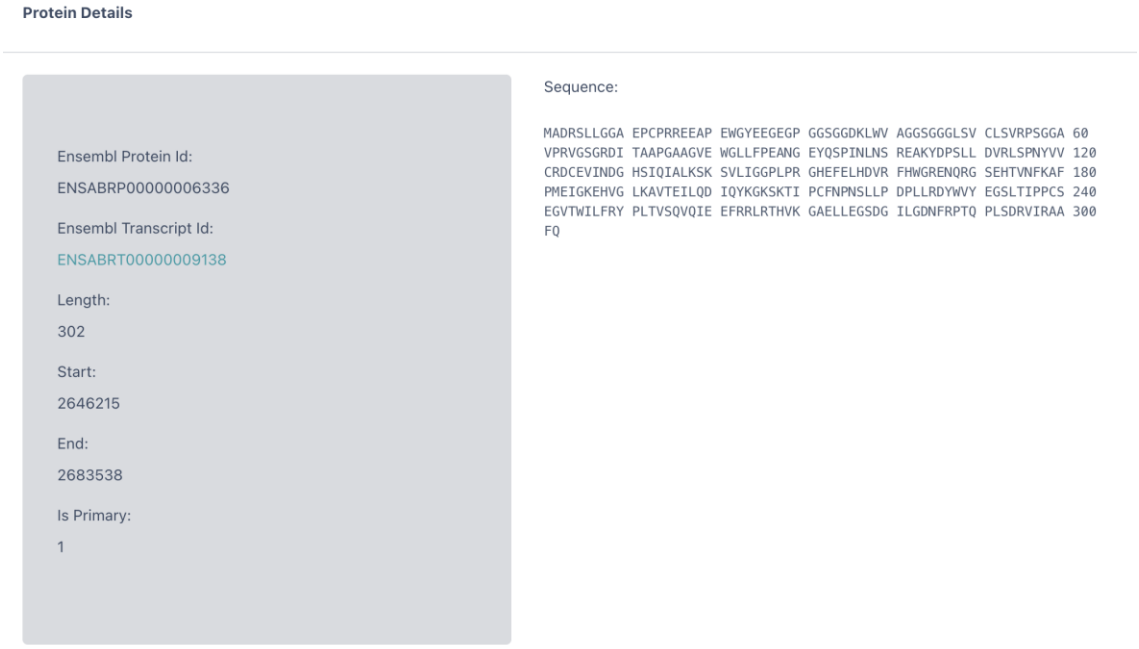
5.6.2 Genes Detailed View



5.6.3 Transcript General View

5.6.4 Transcripts Detailed view

5.6.5 Translation (Protein) Detailed View



5.6.6 Taxon detail view