Cartilage Acidic Protein 1 (CRTAC1), a new member of the beta-propeller protein family with amyloid propensity

Liliana Anjos [a,*], Isabel Morgado [a*#], Marta Guerreiro [a], João C. R. Cardoso [a], Eduardo P. Melo[b] and Deborah M. Power [a]

[*]These authors contributed equally to the work

[a]*Comparative Endocrinology and Integrative Biology Group (CEIB), Centro de Ciencias do Mar (CCMAR), University of Algarve, Campus de Gambelas, 8005-139 Faro, Portugal.*

[b]*Center for Biomedical Research, University of Algarve, Campus de Gambelas, 8005-139 Faro, Portugal.*

[#]*Current address: Department of Physiology and Biophysics, Boston University School of Medicine, 700 Albany Street, W329, Boston MA 02118-2526, USA*

Corresponding authors LA, IM & DMP at: Comparative Endocrinology and Integrative Biology Group (CEIB), CCMAR, University of Algarve, Campus de Gambelas, 8005-139 Faro, Portugal. Tel.: +351 289800958; fax: +351 289800069.

Email address: lanjos@ualg.pt; immorgado@ualg.pt; dpower@ualg.pt

Key words: cartilage acidic protein; extracellular matrix; teleost fish; protein evolution; amyloid

Abstract

Cartilage acidic protein1 (CRTAC1) is an extracellular matrix protein of chondrogenic tissue in humans and its presence in bacteria indicate it is of ancient origin. Structural modeling of piscine CRTAC1 reveals it belongs to the large family of beta-propeller proteins that in mammals have been associated with diseases, including amyloid diseases such as Alzheimer's. In order to characterize the structure/function evolution of this new member of the beta-propeller family we exploited the unique characteristics of piscine duplicate genes Crtac1a and Crtac1b and compared their structural and biochemical modifications with human recombinant CRTAC1. We demonstrate that CRTAC1 has a beta-propeller structure that has been conserved during evolution and easily forms high molecular weight thermo-stable aggregates. We reveal for the first time the propensity of CRTAC1 to form amyloid-like structures, and hypothesize that the aggregating property of CRTAC1 may be related to its disease-association. We further contribute to the general understating of CRTAC1's and beta-propeller family evolution and function.

1. Introduction

Cartilage acidic protein1 (CRTAC1) is an extracellular matrix protein of chondrogenic tissue in humans, which is also present in the brain, eye and nervous tissue [1-3]. In humans two splice variants occur, CRTAC1-A and CRTAC1-B that are the predominant forms in cartilage and brain, respectively [1]. The N-terminal domain of CRTAC1 includes seven phenylalanyl-glycyl-glycyl-alanyl-prolyl (FG-GAP) amino acid repeats forming a β-strand rich region that is associated with beta-propeller structures. This supports the notion that CRTAC1 has a seven-bladed beta-propeller structure as previously suggested [4]. Proteins with a beta-propeller fold are especially interesting because of their extreme sequence/function and phylogenetic diversity despite their similar three-dimensional fold [5,6]. The function of CRTAC1 remains to be established, although the presence of an N-terminal integrin α chain-like domain and a C-terminal EGF-like Ca binding motif suggest it is a calcium-binding protein with a role in cell-cell or cell-matrix interactions [7,8].

CRTAC1 genes have been described in vertebrates, none vertebrate eukaryotes and also in some prokaryotes and its retention during evolution suggests it has an important core function. In teleost fish, the teleost specific whole genome duplication [9] generated two homologue genes, *crtac1a* and *crtac1b*, that in the few teleost species examined so far have been retained. The two forms of teleost Crtac1 share high amino acid sequence conservation with the exception of the C-terminal EGF-Like calcium binding motif that is specific to Crtac1a [4]. The sea bream Crtac1b (saCrtac1b) protein is hyperthermostable and both recombinant and tissue-derived proteins have a high propensity to form large aggregates [10]. We previously hypothesized that the propensity of CRTAC1 to form amyloid-like aggregates may underlie several human pathologies. For example, concentrations of human CRTAC1-B are higher in the cerebrospinal fluid of multiple sclerosis

patients and in the plasma of acute bone fracture patients [11-13]. CRTAC1 is also associated with glomus tumours in neurofibromatosis type 1 disease [14] and in general with diseases of the human cardiovascular, haematological, neurological, respiratory and urinary systems [1]. In addition CRTAC1 has been proposed as a plasma biomarker for detecting/monitoring injury of cartilage [11]. The beta-propeller structures found in CRTAC1 are known in other proteins to play key biological roles and have been directly or indirectly associated with human diseases [5,15,16], including human retinal degeneration and glaucoma [15,17], Glanzmann thrombasthenia [18], and Kallmann syndrome [19]. In some of these diseases pathogenesis is proposed to be associated with the formation of amyloid fibrils and oligomers similar to what occurs in Alzheimer's disease [17,20-22]. Amyloids are toxic beta-sheet rich aggregates derived from the misfolding of proteins. They are well described in bacteria, fungi and mammals and in the latter are related to devastating diseases. Recent studies also point to a putative association between amyloidosis and multiple sclerosis [23]. Amyloids have also been attributed biological functions [24,25] and it is highly challenging to understand their origin, role and why evolution retained this toxic protein fold.

In the present study we update the evolutionary analysis of the *CRTAC1* gene [4] and explore the structural and biochemical modifications that occurred during CRTAC1 evolution and their influence on its aggregation potential [10]. Using a comparative approach, we explore the unique divergent character of teleost duplicate Crtac1's and human CRTAC1 and examine their stability and biochemical properties. The propensity of teleost Crtac1's to aggregate is evaluated from the perspective of their tendency to form amyloid fibrils. The study provides novel insight into the potential biological function of CRTAC 1 in vertebrates and its possible role in disease.

2. Material and Methods

2.1. Evolution of vertebrate CRTAC1

The genomes of 8 mammals, 1 bird (chicken), 3 reptiles, 1 amphibian and of 12 fishes including the lobe-finned fish the Coelacanth, 8 *Actinopterygii* (9 teleosts and the early divergent ray-finned fish, Spotted gar, available from ENSEMBL (http://www.ensembl.org)) were queried using the human (h)CRTAC1 protein (NP_060528) to find CRTAC1 homologues (supplementary Table S1).

The identity of the retrieved sequences was confirmed by reciprocal top BLAST hits against the NCBI human non-redundant protein database (Taxid: 9606) and against the MetaPhOrs database (a public repository of phylogeny-based orthology and paralogy predictions) [26].

The putative CRTAC deduced mature protein sequences were aligned (Fig. S2) using Clustal W (http://www.genome.jp/tools/clustalw/) and the alignment was edited in AliView (version 1.17.1) to select the conserved seven FG-GAP regions and ASPIC/UnbV domain that were concatenated and used to build the phylogenetic tree. The alignment was submitted to ProTest 2.4 server (http://darwin.uvigo.es/software/prottest2_server.html) to select the best model of protein evolution and the phylogenetic trees were constructed according to the Maximum Likelihood (ML) and Neighbour Joining (NJ) methods with the Jones–Taylor–Thornton (JTT) matrix based model. ML analysis was carried out using PhyML 3.0 implemented in ATGC (http://www.atgc-montpellier.fr/phyml/) with fixed I (0.098) and G (0.655) and branch support was analysed using 100 bootstrap replicates and the Bayes statistical methods (Fig. S1). The NJ tree was performed in MEGA 7 with 1000 bootstrap replicates (Fig. S3). Both methods generated similar tree topologies.

The gene environment of the vertebrate *CRTAC1* gene was characterized using BioMart (ENSEMBL, http://metazoa.ensembl.org/biomart/martview/) and sequence homology searches. Analysis was carried out using the human *CRTAC1* neighbouring genes as the query and

homologue genome regions identified in other species using ENSEMBL database annotations and sequence similarity searches.

2.2. CRTAC1 in silico biochemical characterisation

Multiple sequence alignment of the deduced amino acid sequences of CRTAC1 in terrestrial vertebrates (*Homo sapiens, Gallus gallus, Xenopus tropicalis*), fish (*Latimeria chalumnae, Lepisosteus oculatus, Dicentrarchus labrax, Sparus aurata, Oreochromis niloticus, Tetraodon nigroviridis, Danio rerio, Callorhinchus milii*), the invertebrate sea urchin (*Strongylocentrotus purpuratus*) and cyanobacteria (*Synechococcus sp*) were carried out using ClustalX v.2.0.11 and edited using GeneDoc v. 2.7.0 [27]. The multiple sequence alignment (for accession numbers see supplementary Table S1) was used to determine amino acid sequence similarity and to identify conserved motifs and domains by comparison with annotated CRTAC1 sequences and combining searches in NCBI using CD search (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi), Ensemble genomes (http://www.ensembl.org), Superfamily 1.75 (http://supfam.org/SUPERFAMILY/ cgi-bin/scop.cgi), SMART (http://smart.embl-heidelberg.de), Uniprot, (http://www.uniprot.org), Interpro (http://www.ebi.ac.uk/interpro/protein/), Pfam (http://pfam.xfam.org/family/), PROSITE (http://prosite.expasy.org/cgi-bin/prosite/ScanView.cgi) and for motif scanning Myhits (http://myhits.isb-sib.ch/cgi-bin/motif_scan) databases. CRTAC1 transmembrane topology prediction was investigated using the web-based application CCTOP (http://cctop.enzim.ttk.mta.hu/?_=/jobs/submit) [27] with a reliability ≥ 60% including TM filters. The consensus sequence for calcium (Ca)-binding sites were manually identified by sequence similarity for: a) Ca-binding β-hairpin loops characteristic of the integrin α chain like-domain [28,29]

(D/E-h-D/N-X-D/N-G-h-X-D/E (where "h" is hydrophobic and "X" is any residue), and b) epidermal growth factor (EGF)- associated with Ca-binding domain (D/N–x–D/N–E/Q–x_m–D/N*–x_n–Y/F) where m and n are variable and * indicates possible β-hydroxylation [30,31]. Post-translational modifications (PTMs) of CRTAC1 were scanned using ModPred software (http://montana.informatics.indiana.edu/ModPred/index.html) [32] with a cut-off threshold ≥ 0.5 (score between 0-1). CRTAC1 amino acid identity, physico-chemical properties (such as theoretical Mw, Ip, GRAVY and instability índex) were determined and putative disulphide bonds were identified combining sequence similarity with annotated CRTAC1 sequences (Uniprot) and using DISULFIND v.4 (http://disulfind.dsi.unifi.it, [33] and DIANNA v.1.1 (http://clavius.bc.edu/~clotelab/DiANNA/ [34]).

2.3. CRTAC1 homology modelling

Three-dimensional structural homology models of hCRTAC1, teleost dlCrtac1a and dlCrtac1b and sea bream (*Sparus aurata*, sa) saCrtac1b and also cyanobacter (*Synechococcus sp, sy*) syCRTAC1 were obtained using the online platform for protein structure prediction I-TASSER (http://zhanglab.ccmb.med.umich.edu/I-TASSER/) [35]. I-TASSER uses C-score (confidence score) to evaluate the quality of models which is typically in the range of [-5 to 2], where 2 represents the highest confidence. The predicted CRTAC1 models with the highest C-score values were retrieved and visualized using the software PyMOL v1.5.0.3 [36] and compared using the PyMOL structure-based alignment command *cealign* to obtain structural RMSD values. The aggregating propensity of human and piscine CRTAC1's was evaluated using the tools Zyggregator [37], a measure of intrinsic aggregation propensity and Amylpred [38], which identifies particularly aggregation-prone regions within a protein sequence.

2.4. Production of recombinant human and sea bass CRTAC1's

2.4.1. Cloning of sea bass crtac1 homologues

The nucleotide sequences of *dlcrtac1a* and *dlcrtac1b* homologues were extracted from the European sea bass genome as described in 2.1. Specific primers (supplementary Table S2) were designed using Primer Premier 5 software (Premier Biosoft International) to amplify the full length cDNA for *dlcrtac1a* and *dlcrtac1b*. PCR reactions were performed using cDNA synthesized from sea bass larvae total RNA (RNA extraction and cDNA synthesis were performed as reported in [39]). For each PCR reaction 12.5 pmol of each primer, 1.5 mM $MgCl_2$, 0.25 mM dNTP and 1 × Phusion GC PCR buffer were mixed in a final 25 µl reaction volume containing 0.5 unit of Phusion DNA polymerase (Thermo Scientific, US) and 5 µg of cDNA. The thermocycle consisted of: 98 °C for 30 sec followed by 35 cycles of 10 sec at 98 °C, 10 sec at 58 °C and 1 min at 72 °C, followed by a final cycle of 5 min at 72 °C. PCR products were cloned into pGEMT-easy vector (Promega, USA) and sequenced to confirm their identity and the reading frame.

2.4.2. Construction of recombinant expression plasmids for sea bass (dlCrtac1a1 and dlCrtac1b) and human (hCRTAC1) CRTAC1's

A clone containing the coding sequence of *hCRTAC1* was purchased as an ORFEXPRESS™ Gateway PLUS shuttle clone (GeneCopoeia™, USA) and ligated into a pDONR vector (Gateway PLUS shuttle clone). For tag-free expression human and piscine CRTAC1's was cloned into a pET11a vector (Novagen, USA) using specific primers (Table S2). The forward primer included an *NdeI* site upstream of the ATG and the reverse primer a *BamHI* site after the stop codon. PCR reactions were carried out in a 25 µl reaction volume (10 pmol of each primer, 1.5 mM $MgCl_2$, 0.2

mM dNTP, 1x Phusion GC buffer) containing 0.5 units of *Phusion* DNA polymerase (Thermo Scientific, US). The cycling conditions were: 98 °C for 30 s, followed by 30 cycles of 98 °C for 10 s, 62–64 °C for 30 s and 72 ºC for 1 min followed by a final extension of 5 min at 72ºC. The reaction products (27 ng) were subcloned into the pGEM-T easy vector (15ng, Promega, USA) and propagated in *E. coli DH5α* strain (Promega, USA). Restriction digest of the constructs (2 µg) with *NdeI* and *BamHI* (5 U, GE Healthcare, UK) liberated CRTAC1's which were ligated into pET11a (100 ng, Novagen, Germany) using 4 units of T4 ligase (Promega, USA). Recombinant vectors pET11a/CRTAC1's were used to transform one Shot *DH5α* competent *E. coli* using heat-shock. Plasmid DNA pET11a/CRTAC1's was isolated from individual clones using the alkaline lysis method and the authenticity of the vector inserts confirmed by PCR, restriction digest and sequencing.

2.4.3. Expression and purification of recombinant sea bass (dlCrtac1a and dlCrtac1b) and human (hCRTAC1) CRTAC1's

The three CRTAC1 expression constructs (10 ng) were used to transform BL21 (DE3) and Origami^{TM}2(DE3) (Novagen, USA) and single positive bacterial clones of BL21 (DE3)-hCRTAC1 and Origami^{TM}2(DE3)- dlCrtac1a and dlCrtac1b were used to inoculate LB media containing ampicillin (100 µg/ml) and cultured overnight at 37ºC and 250 rpm. Large scale cultures were established by inoculating (ratio 1:40) 2 L of LB enriched with 35 mM K_2HPO_4, 4 mM KH_2PO_4, 4 mM glucose and ampicillin (100 µg/ml) with the pre-culture of a single bacterial colony. Protein expression was induced with 0.8 mM IPTG (Sigma-Aldrich, Spain) when cultures attained an $OD_{600\,nm}$ = 0.6-0.8. Optimal expression conditions were found to be 6 h post-induction growth at 30 ºC for dlCrtac1a and dlCrtac1b and 12 h at 30 ºC for hCRTAC1. Bacterial pellets

were harvested by centrifugation for 10 min at 10 000 x g, 4 °C, and lysed with lysis buffer [200 mM Tris–HCl pH 8, 500 mM NaCl, 0.1 mg/ml lysozyme (Sigma, UK), 0.025 mg/ml DNase I (Sigma, USA), 4.2 mM MgCl2, 1 mM phenylmethanesulfonyl fluoride (Sigma, US)] at a ratio of 6:41 (v/m, buffer to bacterial pellet). Three freeze/thaw cycles and sonication (3 cycles of 30 sec and 1 cycle of 1 min) were used to maximize lysis. The insoluble protein fraction was collected by centrifugation for 30 min at 25 000 x g, 4 ºC and solubilized in denaturing buffer (8 M Urea, 50 m M Tris pH 8, 100 mM NaCl, 10 mM EDTA) at a ratio of 3: 1 (v/m, buffer : pellet) for 16h at 4ºC with gentle stirring. Soluble and insoluble protein expression was confirmed by analysis on SDS-PAGE using a 10% polyacrylamide gel.

The untagged CRTAC1's were purified by continuous elution electrophoresis using a Model 491 Prep Cell (Bio-Rad, Portugal) as previously described [10]. Protein fractions (2.5 ml) were collected from the elution chamber, analyzed by SDS-PAGE, and fractions containing isolated protein, pooled and concentrated using Ultrafree-15 centrifugal filters (Millipore, Bedford, UK). After concentration to 1 mg/ml the proteins were mixed with 5 mM of DTT and 2 M of Urea and incubated at 4ºC with agitation for 2h. The proteins were refolded by dialysis (cut-off of dialysis tube 12kDa) against several buffers; 1) 0.5 M Urea, 100 mM Glycine, 100 mM Tris-HCl pH 8, 0.4 M L-arginine, 0.5 mM Glutathione (oxidized), 5 mM Glutathione (reduced); 2) 100 mM Glycine, 100 mM Tris-HCl pH 8, 0.2 M L-arginine, 0.5 mM Glutathione (oxidized), 5 mM Glutathione (reduced); 3) 100 mM Tris-HCl pH 8, 100 mM Glycine, 0.1 M L-arginine, 0.25 mM Glutathione (oxidized), 2.5 mM Glutathione (reduced); 4) 100 mM Tris-HCl pH 8, 100 mM Glycine; and 5) 100 mM Tris-HCl pH 8, 250 mM NaCl. Each dialysis step was performed for 16 h at 4 ºC with gentle stirring. Recovery of recombinant CRTAC1 was monitored at all steps of purification, protein concentration and refolding by measuring UV–vis absorption at 280 nm and determining

the molar extinction coefficient (ε_{280nm}=52,600 $M^{-1}cm^{-1}$, ε_{280nm}=58,050 $M^{-1}cm^{-1}$, ε_{280nm}=38,270 $M^{-1}cm^{-1}$ [40].

2.4.4. Mass spectrometry analysis of recombinant CRTAC1's

Purified recombinant CRTAC1 was fractionated by SDS-PAGE (10% polyacrylamide gel), stained with Coomassie Blue (0.1% w/v), distained with acetic acid (10% (v/v) and methanol (45% (v/v)), and protein bands with the predicted mass of CRTAC1's excised and placed in MilliQ water and analysed by mass spectrometry. An additional low molecular weight protein observed during dlCrtac1b expression was also analysed by mass spectrometry. Proteins were subject to in-gel trypsin digestion and analyzed by MALDI-TOF (Matrix Assisted Laser Desorption Ionization (MALDI) tandem Time-of-Flight (TOF) mass spectrophotometer) in the Centro de Genómica y Proteómica (Unidad de Proteómica, Faculdad de Farmacia, UCM, Spain). MALDI data was analysed using mgf archives for database searching and the MASCOT search engine (http://www.matrixscience.com). The following parameters were permitted during the searches: 2 missed cleavage sites as well as fixed and variable modifications; carbamidomethyl of cystein and oxidation of methionine.

2.5. CRTAC1's structure, stability and aggregation

2.5.1. Circular Dichroism

Far-UV spectra were recorded between 180-250 nm in a Spectropolarimeter Jasco J-815 model using a 1 mm path length quartz cell (Hellma). Pure tag-free CRTAC1 was dialyzed against 10mM Tris-HCl pH8 (or water for the thermal stability measurements) and used at a final concentration of 5 µM. Thermal scans were performed to evaluate and compare the thermal stability of CRTAC1 in water between 20 ºC and 90 ºC at 210 nm using a peltier unit. The temperature measurements

were obtained at a rate of 1 ºC/min. Spectra were corrected for the base line measured with buffer or water and a minimum of three thermal scans averaged for each measurement. Results are expressed in mean residue weight ellipticity ([θ] mrw).

2.5.2. Fluorescence

Intrinsic fluorescence was measured in a Fluoromax3 Horiba Scientific (Japan) using a rectangular quartz cell with a 10 mm path length and excitation at 280 nm. Emission spectrum were recorded between 290 and 420 nm using a 1 nm slit width. Protein samples were prepared in water to a final concentration of 1 μM. Thermal stability was assessed qualitatively by recording scans at 20 ºC and after heating the samples to 90 ºC. Chemical denaturation was tested by recording scans of CRTAC1 in water only, or after incubation with GnHCl 6 M. Samples were prepared at 5 μM in Tris 10 mM pH 8 containing 6 M GnHCl and diluted to 1 μM in H_2O immediately before measurement. Scans were normalized to the maximum emission wavelength for comparison.

2.5.3. Size exclusion chromatography

Chromatography was carried out using a Superdex 200 10/300 GL column (GE healthcare, UK) connected to an AKTA fast protein liquid chromatography (FPLC) system (GE healthcare, UK) following the manufacturer's instructions and guidelines. The column was pre-calibrated for molecular weight estimation by injection of a 500 μl sample of protein standards containing Thyroglobulin (670 kDa), γ-globulin (158 kDa), Ovalbulmin (44 kDa), Myoglobin (17 kDa) and Vitamin B12 (1.35 kDa). Isolated recombinant CRTAC1's were dialysed against chromatography buffer (100 mM Tris-HCl pH8, 250 mM NaCl) and then 1ml of hCRTAC1 (33 μM), dlCrtac1a

(61 μM) and dlCrtac1b (77 μM) analysed individually using a flow rate of 1ml/min with absorbance set at 208 nm and monitored using UNICORN software (GE healthcare, UK).

2.5.4. Native PAGE

Native PAGE (8% polyacrylamide) [41] under non-denaturing and non-reducing conditions was used to confirm the aggregates and monomeric species identified by size exclusion chromatography. Fractions containing CRTAC1 species were loaded on polyacrylamide gels run at a constant power (35 mA) for 1h and stained with Coomassie Blue (0.1% (w/v), 10% (v/v) acetic acid and 40% (v/v) methanol) and distained (10% v/v acetic acid, 40% (v/v) methanol) before analysis. The monomeric hCRTAC1, dlCrtac1a and 1b proteins were detected using a Silver Stain Plus™ Kit (Bio-Rad, USA).

2.5.5. Transmission electron microscopy (TEM)

Before TEM analysis hCRTAC1, dlCrtac1a1 and dlCrtac1b were incubated in Tris buffer 100 mM, 250 mM NaCl for 3 weeks at 37°C, samples were taken before incubation (T = 0) and after each week and immediately used to prepare TEM grids. Salts (50 mM and higher) are reported to increase the aggregation propensity of Alzheimer's related Aβ peptide [42]. Since CRTAC1 is a putative calcium-binding protein [10] we evaluated the impact of calcium on its aggregation propensity by incubating hCRTAC1, dlCrtac1a and dlCrtac1b for 1 week with 50 mM $CaCl_2$. TEM analysis was carried out using the negative stain method and a CM12 transmission electron microscope (Philips Electron Optics) operated at 80 kV. Samples were prepared by applying 4 μl

droplets of each CRTAC1 protein solution onto glow-discharged carbon formvar-coated copper grids and incubated for 30 sec or 1 min (dlCrtac1b), 4 min (hCRTAC1) or 3 min (dlCrtac1a), washed with a 4 μl water droplet for 1 min and counterstained with a 4 μl droplet of a freshly filtered solution of 2% (w/v) uranyl acetate. Grids were left to air dry for 3 min after staining. Protein concentrations were as follows: dlCrtac1b 2.7 mg/ml (39 μM); hCRTAC1 0.7 mg/ml (10 μM); dlCrtac1a 1.3 mg/ml (20 μM).

3. Results and discussion

3.1. Biochemical and molecular characteristics of CRTAC1 proteins during evolution

A comparative approach was taken to investigate the biochemistry, structure and aggregating behaviour of CRTAC1's during evolution. To put into context our comparative characterization of CRTAC1's and because information about CRTAC1 is still fairly recent and scarce, we performed a detailed *in silico* evaluation of CRTAC1 gene evolution also updating previous information [4] (Supplementary Fig. S1, Fig. S2, Fig. S3 and Fig. S4). We then used further *in silico* methods to obtain a biochemical, molecular and structural profile of CRTAC1 homologues from prokaryotes to vertebrates (Table 1 and Fig. 1, Fig. S2, Fig. S5). While providing novel important information, the *in silico* studies complemented and formed the basis of our further experimental analysis of CRTAC1 structure and aggregation.

3.1.1. Evolution of CRTAC1 in metazoan

Phylogenetic analysis confirmed that CRTAC1's are an ancient family of proteins since they are present from bacteria to man**.** The vertebrate *CRTAC1* gene complement shared a common

evolutionary origin and were duplicated early in the teleost expansion. The identification of multiple gene copies of *CRTAC1* in the genome of the early deuterostomes, amphioxus (*Branchiostoma floridae*) and sea urchin (*Strongylocentrotus purpuratus*) and its absence from others even though the chromosome region has been conserved (Fig. S4, eg: Ciona and insects and nematodes) indicates that during the metazoan radiation the ancestral *CRTAC1* gene was preserved in some and lost in other genomes. The duplicate crtac1's in sea bass (*crtac1a* and *crtac1b*) and the human *CRTAC1* were used as models in the present study to evaluate experimentally the stability and aggregation of CRTAC1's from evolutionary distant organisms.

3.1.2. Molecular organization of CRTAC1

The overall biochemical/biophysical characteristics of CRTAC1 have been conserved from bacteria to mammals (Table 1 and Fig. S5), suggesting they are important for structural (secondary/tertiary structure) topology maintenance and function. Although the molecular complexity of CRTAC1 increases from bacteria to mammals (Fig. S4 and S5, see legend and supplementary text), an N-terminal integrin α chain like-domain and an ASPIC and a UnbV domain are conserved in all the CRTAC1 sequences analyzed. A C-terminal calcium-binding epidermal growth factor domain (EGF-Ca) is present in all CRTAC1 forms with the exception of teleost Crtac1b and sea urchin and cyanobacteria CRTAC1. Five consensus Ca-binding β-hairpin loops are present in CRTAC1 from bacteria to mammals (as previously described [4]) along with an additional EGF-Ca binding site in the EGF-Ca domain (Fig. S2 and S5). Sea bream saCrtac1b lacks the EGF-Ca binding domain but still has high Ca binding affinity[10] suggesting this motif does not have a major role in Ca binding by the protein.

Overall, vertebrate CRTAC1 (including teleost Crtac1a) are cysteine rich containing approximately twelve conserved cysteines. Six of the cysteines are in the C-terminal EGF-Ca domain [43] and potentially form 3 disulfide bonds according to three-dimensional model predictions [44]. The homologues lacking this domain such as teleost Crtac1b have generally fewer cysteines (6 present in Crtac1b) (Fig. S2 and Table1).

3.2. Recombinant human and sea bass CRTAC1

In order to evaluate experimentally the biochemistry and structure of CRTAC1's we produced recombinant human and sea bass CRTAC1's based on the sequences available or extracted as described in section 2.4 and Table S1 legend.

Duplicate *Crtac1* transcripts from sea bass that were 1905 bp and 1644 bp in length for *dlCrtac1a* (deposited in *GenBank* with accession number: KX364268) and *dlCrtac1b* (deposited in *GenBank* with accession number: KX364269), respectively, were used as a template for recombinant protein expression. The biochemical and structural information obtained for CRTAC1 homologues *in silico* (section *3.1.2*) was used to assist the design of the recombinant CRTAC1 production strategy as these proteins proved to be difficult to express. hCRTAC1, dlCrtac1a and dlCrtac1b were expressed in the insoluble cellular fraction as inclusion bodies (Fig. S6A). The proteins were recovered from the inclusion bodies and successful purification verified by SDS-PAGE analysis (Fig. S6B and Table 1). The yield of purified CRTAC1's (pure refolded protein/protein after the first purification step) was 76.5% for hCRTAC1, 90.9% for dlCrtac1a and 63.5 % for dlCrtac1b. Loss of protein was highest during the concentration step as a result of insoluble aggregate formation. The yield of purified hCRTAC1, dlCrtac1a and dlCrtac1b was 11.5, 8.4 and 11.9 mg per gram of bacteria, respectively.

MALDI-TOF analysis was used to determine the fragmentation pattern of the recombinant proteins generated. Comparison of the recombinant protein fragmentation pattern with that predicted for deduced sea bass and human proteins confirmed protein purity and that the appropriate recombinant proteins was expressed. The MS fingerprint spectra of recombinant hCRTAC1, dlCrtac1a and dlCrtac1b were composed of 21, 20 and 24 tryptic peptides, respectively (Table S3). MASCOT blast search revealed the MS fingerprint of hCRTAC1 and dlCrtac1a best matched Cartilage Acidic protein 1 isoform X1 (*Homo sapiens*, gi:530393911) and Cartilage Acidic Protein 1 (*Tetraodon nigroviridis*, gi: 85838736), respectively. The sequence coverage, the nominal mass (Mr) and *Ip* was respectively, 39%, 68.6 kDa and 5.05 for hCRTAC1 and 29%, 69.6 kDa and 5.42 for dlCrtac1a. MASCOT blast searches against public databases (http://www.ncbi.nlm.nih.gov) failed to recover a match for the MS fingerprint of the two recombinant dlCrtac1b proteins (high and low molecular weight, Fig. S6A) identified by SDS-PAGE. This was unsurprising as blast searches with the full length cDNA of dlCrtac1b also failed to retrieve a match. MASCOT blast search against an in house database [45] retrieved dlCrtac1b. The sequence coverage, nominal mass (M_r) and *Ip* of the full length dlCrtac1b was 60%, 59.2 kDa and 5.03, respectively.

3.3. CRTAC1 structure and conformation

3.3.1. Three-dimensional structure models

Three dimensional models of human, teleosts and cyanobacter (*Synechococcus sp*, syCRTAC1, ancient CRTAC1) CRTAC1's were generated using I-TASSER [35] (Fig. 1). The models were built using the top scoring templates retrieved by the algorithm. C-scores of all models were between -1.21 and -1.98. The templates used for CRTAC1 model construction (chosen by the algorithm

based on fold or super-secondary structure similarity) were the Ca-binding integrin-like fungal (*Psathyrella velutina*) lectin (PDB:2BWR), a highly multispecific and multivalent protein and the lymphocyte receptor integrin α(4)β(7) (PDB:3V4P) which mediates both rolling and firm cell adhesion and is targeted by therapeutics approved for multiple sclerosis [46,47] and Crohn's disease. Similar roles and disease relevance have been suggested for human CRTAC1.

The I-TASSER models of the N-terminal region of hCRTAC1, dlCrtac1a, dl and saCrtac1b and syCRTAC1 in the present study coincided with the previously modelled structure of saCrtac1b [4] (Fig. 1). All the proteins analysed had a conserved beta-propeller structure and five integrin-like Ca-binding loops rich in β-sheet structure (Fig. 1). Remarkably, despite the noticeable differences in sequence similarity between syCRTAC1, fish and human CRTAC1's (Table S4A), RMSD values support an overall high resemblance between fish and human CRTAC1's (Fig. 1B). This evidence supports the notion that this protein is under strong evolutionary pressure for structural conservation (regardless of sequence identity) possibly related with a determinant role of the beta-propeller, the most conserved region (hCRTAC1 beta-propeller isolated region shares 85 - 87% aa similarity with teleosts Crtac1a and 1b and 62% with syCRTAC1 and 61 - 64% of aa similarity are shared between syCRTAC1 and teleosts Crtac1a and 1b respectively, Table S4B). Curiously there is a higher resemblance (lower RMSD values) between fish CRTAC1's and their homologues in the extreme evolutionary ends (syCRTAC1 vs saCrtac1b and hCRTAC1 vs dlCrtac1a) than between the fish CRTAC1's despite their higher sequence identity (60-90%). The discrepancy between the sequence and structure is particularly noticeable with dlCrtac1b and even taking into consideration the low confidence score of the structural model seems to suggest functional divergence occurred between piscine CRTAC1's. This could hypothetically be related with the generation of conformational diversity arising from the teleost genome duplication.

3.3.2. Secondary structure and stability

The secondary structure of human and teleost CRTAC1's (Fig. 2A) was determined using far-UV CD spectra measurements and the algorithm CONTIN. The UV spectra of hCRTAC1 had a negative peak (-1.0853x10^3 deg cm2/dmol) at 211.2 nm and was estimated to contain 10.3% α-helix, 38% β-sheet and 51.6% disordered structure. dlCrtac1a had a negative peak (-9.5017x10^3 deg cm2/dmol) at 210.1nm and was estimated to contain 10.3% α-helix, 38% β-sheet and 51.6% disordered structure. The UV spectra of dlCrtac1b had a negative peak (-1.00619x10^3 deg cm2/dmol) at 207.6 nm and contained 10.4% α-helix, 38.1% β-sheet and 51.5% disordered structure. As expected from observation of the three-dimensional models (Fig.1A), all CRTAC1's had high β-sheet content (Fig. 2A) and hCRTAC1 and dlCrtac1a with an ellepticity minimum at 211nm shared the greatest structural similarity. The dlCrtac1b had an ellepticity minimum (207.6 nm) close to that reported for saCrtac1b (202.5 nm) [10].

Thermal unfolding (thermal state transition), identified as a change in the ellepticity minimum (210 nm) was not completely achieved by CRTAC1 when heated from 20 ºC to 95 ºC suggesting high thermostability (Fig. 2B). There was only a small increase in the ellepticity minimum, indicative of a loss of secondary structure, which started around 50 ºC and was much more pronounced for hCRTAC1 that changed from 48.4% at 20 ºC to 39.6% at 70 ºC. The ellepticity minimum for dlCrtac1a and dlCrtac1b only started to increase at 70 ºC. The high thermostability of CRTAC1's was further confirmed using intrinsic fluorescence spectroscopy (Fig. 2C). Spectra were measured at 20 ºC and 90 ºC and no appreciable emission maximum shift, suggestive of unfolding, was detected after heating (Fig. 2C). CRTAC1's was thermostable and retained tertiary structure up to 90ºC. The results of fluorescence spectroscopy and Far-UV CD spectra, were

coincident and indicated that CRTAC1's were thermally stable with the piscine forms being more stable than human CRTAC1. In contrast, the outcome of exposure of CRTAC1's to Guanidine Hydrochloride (6M) was indicative of tertiary structure unfolding and a strong maximum emission shift occurred from, 338 to 356 nm for hCRTAC1, 343 to 353 nm for dlCrtac1a1 and 344 to 355 nm for dlCrtac1b (Fig. 2D) and indicated the proteins were susceptible to chemical unfolding. As predicted CRTAC1's possessed a high aliphatic index (Fig. S5, hCRTAC1-74.67%, dlCrtac1a-76.56%, dlCrtac1b-78.89%) which is an indicative factor of high thermostability [48] and proteins with both, a beta-propeller fold and ion-binding functions have been described with thermostable functions and structures [49]. Since CRTAC1 forms high molecular weight aggregates spontaneously in solution[10], the observed thermostability may also be a consequence of a high thermoresistance of the formed aggregates. The association between thermostability and protein aggregate formation have already been reported for Fe-superoxide dismutase [50] and α-crystalline [51].

3.4. CRTAC1's aggregation

3.4.1. The aggregating propensity of CRTAC1

CRTAC1's are acidic proteins with a high aliphatic index (Table 1, Fig. S5) (71.3 - 82.7%) that with few exceptions (teleost Crtac1b) contain three main globular domains (Glob. I, II and III, Fig. S2). The main globular regions are separated by disordered or unstructured regions and the C-terminal region of CRTAC1's, including the EGF-Ca domain are rich in disordered structure. Human and sea urchin CRTAC1 are predicted to be more disordered than the protein from cyanobacteria, elephant shark and Crtac1b of some teleosts. These observations are intriguing and may suggest that acquisition by CRTAC1 of structural disorder may be an advantageous adaptation. Disordered regions in globular proteins can provided structural flexibility, improve

molecular interactions and even help to prevent aggregation[52]. Proteins with more intrinsically disordered regions have much lower packing density, which is typically very high in amyloid proteins. Monte Carlo simulations reveal that hydrophobic peptides with disordered flanks become more stable and less-aggregation prone[53]. Future experimental studies will be essential to determine the structural and functional consequences of disordered regions in CRTAC1.

The aggregation propensity of CRTAC1 was evaluated with Zyggregator. Results suggest that hCRTAC1 has a lower aggregation propensity than piscine Crtac1, which had values close to the high aggregation threshold 1. Curiously, according to Amylpred predictions, hCRTAC1 possesses slightly more aggregation-prone segments (Table 1). This may indicate a structural conformation/organization that confer higher stability despite the presence of additional "hot spots" in the EGF-Ca domain which is absent in dlCrtac1b. In human and piscine CRTAC1's more than 75% of the aggregation "hot spots" were located in the beta-propeller region, which accounts for 57%, 60% and 69% of the length of hCRTAC1, dlCrtac1a and dlCrtac1b, respectively and suggests that this region may determine their aggregation propensity into amyloids. This is in line with observation of other proteins belonging to the beta-propeller family that have associated this structure with aggregation and amyloid formation [15,17,20,22].

3.4.2 CRTAC1's form high molecular weight amyloid-like aggregates

Size-exclusion chromatography (SEC) and PAGE revealed that purified human and piscine recombinant CRTAC1's spontaneously form high molecular weight species as previously found for saCrtac1b[10]. hCRTAC1, dlCrtac1a and dlCrtac1b were found predominantly as high molecular weight species, > 670kDa (Fig. 3A). However, smaller peaks with higher retention volumes of approximately 14-15 ml, which could correspond to approximately (qualitative rough estimation)

40-70kDa were also identified for hCRTAC1, dlCrtac1a and dlCrtac1b and likely correspond to the monomeric form of the protein. Lesser amounts of monomeric dlCrtac1a and dlCrtac1b were detected compared to hCRTAC1 suggesting that the piscine proteins are mostly aggregated and less stable or less soluble than hCRTAC1. This result supported the previous observations that hCRTAC1 may be less amyloid-aggregation prone (Table 1) and therefore less likely to form thermostable aggregates. Native PAGE confirmed the presence of CRTAC1 aggregates and monomers with low and higher retention volumes, respectively in the SEC (Fig. 3B). However, for dlCrtac1b, PAGE revealed that the protein in the higher retention volume fraction contained a mixture of both aggregates and monomers suggesting that monomeric dlCrtac1b was highly unstable and aggregated immediately after isolation. Additional analysis in semi-native PAGE (where samples were fractionated in a 10% SDS-PAGE without the stacking gel and in the presence or absence of DTT and without thermal denaturation) revealed that 50 mM DDT disaggregates CRTAC1's suggesting that disulphide bonds may be involved in self-association of CRTAC1 monomers (Fig. S7).

In order to evaluate the nature of CRTAC1 aggregates we used electron microscopy and found that CRTAC1's can form amyloid structures *in vitro*. Samples were analyzed over time to follow the progression of aggregation and the morphology of the aggregates (Fig. 4). Before incubation (T = 0) the piscine CRTAC1's already formed small aggregation nuclei mixed with small oligomeric structures, which appeared larger in the case of dlCrtac1b. Such structures were much smaller or absent in hCRTAC1 before incubation. After 1 week at 37ºC both human and piscine CRTAC1 solutions contained elongated oligomeric structures that formed dense clusters of aggregates. In some cases, instead of clusters, long and branched fibrils were observed. In particular, compared to dlCrtac1b both dlCrtac1a and hCRTAC1 formed larger and more branched

aggregates. After 2 weeks at 37ºC dlCrtac1a and hCRTAC1 samples had fewer aggregation clusters and more disperse smaller aggregation units while dlCrtac1b contained very well defined curvilinear protofibrils. No significant change in the aggregation patterns of the CRTAC1 proteins analysed was identified after 3 weeks incubation (Fig. S8). In general aggregates of hCRTAC1 and dlCrtac1a formed faster and had less defined morphology than those of dlCrtac1b (lacking the EGF-Ca domain).

Because saCrtac1b is a high affinity calcium-binding protein [10] and several calcium binding sites were predicted to be present in all CRTAC1 homologues (Fig. S5), the influence of 50 mM calcium on the formation of amyloid aggregates was also evaluated. Different concentrations of salts (50 mM upwards) have previously been reported to accelerate amyloid aggregation for Aβ peptide [42]. After incubation at 37ºC in the presence of 50 mM calcium for one week both isoforms of piscine CRTAC1's had fallen out of solution. A white dense precipitate was visible at the bottom of the tube but no protein aggregates could be detected by TEM. In contrast hCRTAC1 did not show signs of aggregation and revealed a heterogeneous distribution of small nuclei and slightly elongated and curvilinear oligomers. This could reflect a higher stability of monomeric hCRTAC1 (also suggested by *in silico* analysis, section 3.1.2) that allows calcium in solution to reach the calcium-binding pockets of the beta-propeller. We hypothesize that in the rapidly aggregating piscine protein the calcium-binding pockets are probably less available, which maintains high calcium levels in solution and this promotes further aggregation.

Proteins with beta-propeller structures have been shown to form amyloid aggregates suggesting this is a common event associated with this particular fold. For example Vitronectin (VTNC), a four bladed beta-propeller and multifunctional protein was found to form spherical oligomers and

elongated amyloid fibrils *in vitro* upon incubation for several days in PBS buffer at room temperature [20]. Moreover, VTNC oligomers were found to be toxic to neuroblastoma cells and amyloid deposits were detected in Alzheimer's disease brain plaques. VTNC like CRTAC1 is an extracellular matrix protein potentially involved in cell adhesion and both are associated with nervous tissues. Similarly, myocilin (olfactomedin domain of myocilin, myoc-OLF) is a calcium-binding protein that contains a five-bladed beta-propeller domain which forms amyloid fibrils *in vitro* under physiological pH and buffering conditions[17]. myoc-OLF is associated with glaucoma hypothetically through aggregation in the region of the anterior eye and hCRTAC1 is also expressed in the eye [2,3] although its function is not clear. We speculate similarities in the structural and functional properties between CRTAC1, VTNC and myoc-OLF may suggest CRTAC1-derived amyloids *in vivo* may be equally toxic and disease relevant.

4. Conclusion

CRTAC1 proteins are poorly described and their function remains unclear even though the encoding genes have been maintained in the genome from bacteria to humans. Several studies predict that in humans CRTAC1 is involved in chondrocyte function [1,54] but additional functions are likely as this protein is also present in the brain (isoform CRTAC1-B) and nervous tissue [1]. A significant interest in CRTAC1 arises from its putative but unclear association with several diseases [12,13].

Our results confirm the ancient character of CRTAC1 and indicate that in metazoans it is preserved in some lineages but deleted from others and that in teleosts *crtac1* duplicates (*crtac1a* and *crtac1b*) emerged during the teleost specific whole genome duplication event. Molecular details of CRTAC1's domains/motifs, structural and biochemical features provide a pathway to future

structure-function studies. The amino acid sequence and structural domains of CRTAC1 are strongly conserved across species but acquisition of novel domains, such as the TM regions, RGD tripeptide or GPI-lipid anchor, by some lineages indicate diversification of protein function during evolution. Teleosts are the only organisms in which duplicate *crtac1* genes coexist and the loss and gain of functional domains (e.g. EGF-Ca binding motif, GPI-lipid anchor and nuclear exporting signal) suggests neo-functionalization has occurred.

The comparative approach taken, using human and the duplicate piscine Crtac1a and Crtac1b provided insight into the protein features that influence the aggregation propensity of CRTAC1 in vertebrates. hCRTAC1 had decreased aggregation propensity compared to the piscine homologues as suggested by: 1) high percentage of disordered structure (sequence-based *in silico* prediction), 2) lower amyloidogenic index (sequence-based *in silico* prediction), 3) formation of less thermostable aggregates, 4) higher percentage of monomer in solution, and 5) slower rate of amyloid formation. The sequence or structural basis of the higher stability and lower aggregation propensity of hCRTAC1 was not identified in the present study but may reflect adaptive evolutionary mechanisms linked to the acquisition of thermoregulation and maintenance of a core temperature of 37º C in order to avoid or slow down amyloid formation. Amyloids have been attributed biological functions [24,25] and it is highly challenging to understand their origin, role and why evolution retained this toxic protein fold. It will be important to further ascertain whether CRTAC1's aggregates can be found *in vivo* and more importantly whether they are disease-related. Nevertheless, here we provide the basis for future studies to understand CRTAC1 function, evolution and putative association with diseases.

Acknowledgments

The authors thank Dr. Rute Martins for invaluable help with *in silico* prediction of sea bass Crtac1 sequences and the availability of the sea bass larvae cDNA and also Marlene Trindade for help with the CRTAC1 evolutionary studies.

Funding: This work was supported by the European Regional Development Fund through COMPETE and the Portuguese Foundation for Science and Technology (FCT) [PTDC/MAR/122296/2010] and national funds through FCT project CMAR/Multi/04326/2013. IM was funded by a Post-doctoral fellowship supported by these projects and a European Commission MSC International Outgoing Fellowship (IOF) 628077; LA was in receipt of a Post-doctoral fellowship [SFRH/BPD/79105/2011] from FCT, the Ministry of Science and Higher Education.

References

1. Steck E, Braun J, Pelttari K, Kadel S, Kalbacher H, Richter W. Chondrocyte secreted CRTAC1: a glycosylated extracellular matrix molecule of human articular cartilage. Matrix Biol 2007;26(1):30-41.
2. Rabinowitz YS, Dong L, Wistow G. Gene expression profile studies of human keratoconus cornea for NEIBank: a novel cornea-expressed gene and the absence of transcripts for Aquaporin 5. Invest Ophthalmol Vis Sci 2005;46(4):1239-1246.
3. Turner HC, Budak MT, Akinci MAM, Wolosin JM. Comparative analysis of human conjunctival and corneal epithelial gene expression with oligonucleotide microarrays. Invest Ophthalmol Vis Sci 2007;48(5):2050-2061.
4. Redruello B, Louro B, Anjos L, Silva N, Greenwell RS, Canario AV, Power DM. CRTAC1 homolog proteins are conserved from cyanobacteria to man and secreted by the teleost fish pituitary gland. Gene 2010;456(1-2):1-14.

5. Chen CKM, Chan N-L, Wang AHJ. The many blades of the beta-propeller proteins: conserved but versatile. Trends Biochem Sci 2011;36(10):553-561.
6. Paoli M. Protein folds propelled by diversity. Prog Biophys Mol Biol 2001;76(1-2):103-130.
7. Hughes AL. Evolution of the integrin alpha and beta protein families. J Mol Evol. 2001 52(1):63-72.
8. Cioci G, Mitchell EP, Chazalet V, Debray H, Oscarson S, Lahmann M, Gautier C, Breton C, Perez S, Imberty A. Beta-Propeller crystal structure of psathyrella velutina Lectin: an integrin-like fungal protein interacting with monosaccharides and calcium. J Mol Biol 2006;357(5):1575-1591.
9. Nelson J. Fishes of the world. John Wiley and Sons; 2006.
10. Anjos L, Gomes AS, Melo EP, Canario AV, Power DM. Cartilage Acidic Protein 2 a hyperthermostable, high affinity calcium-binding protein. Biochim Biophys Acta 2013;1834(3):642-50.
11. Grgurevic L, Macek B, Durdevic D, Erjavec I, Pandzic M, Mann M, al. e. Novel biomarkers in the plasma of patients with a bone fracture. Calcif Tissue Int 2008;82:S122-23.
12. Hammack BN, Fung KY, Hunsucker SW, Duncan MW, Burgoon MP, Owens GP, Gilden DH. Proteomic analysis of multiple sclerosis cerebrospinal fluid. Mult Scler 2004;10((3)):245-60.
13. Veenstra TD, Conrads TP, Hood BL, Avellino AM, Ellenbogen RG, Morrison RS. Biomarkers: mining the biofluid proteome. Mol Cell Proteomics 2005;4(4):409-18.
14. Brems H, Park C, Maertens Ol, Pemov A, Messiaen L, Upadhyaya M, Claes K, Beert E, Peeters K, Mautner V and others. Glomus tumors in neurofibromatosis type 1: genetic, functional, and clinical evidence of a novel association. Cancer Res 2009;69(18):7393-7401.
15. Pons T, Gomez R, Chinea G, Valencia A. Beta-propellers: associated functions and their role in human diseases. Curr Med Chem 2003;10(6):505-24.
16. Szeltner Z, Polgar L. Structure, function and biological relevance of prolyl oligopeptidase. Curr Protein Pept Sci 2008;9(1):96-107.

17. Hill SE, Donegan RK, Lieberman RL. The glaucoma-associated olfactomedin domain of myocilin forms polymorphic fibrils that are constrained by partial unfolding and peptide sequence. J Mol Biol 2014;426(4):921-935.
18. Nurden AT. Glanzmann thrombasthenia. Orphanet Journal of Rare Diseases 2006;1:10-10.
19. Hefner J, Csef H, Seufert J. Kallmann-Syndrom. Der Nervenarzt 2009;80(10):1169-1175.
20. Shin T, Isas J, Hsieh C-L, Kayed R, Glabe C, Langen R, Chen J. Formation of soluble amyloid oligomers and amyloid fibrils by the multifunctional protein vitronectin. Mol Neurodegener 2008;3(1):1-12.
21. Xu D, Baburaj K, Peterson CB, Xu Y. Model for the three-dimensional structure of vitronectin: predictions for the multi-domain protein from threading and docking. Proteins. 2001 Aug 15;44(3):312-20. 2001.
22. Orwig SD, Perry CW, Kim LY, Turnage KC, Zhang R, Vollrath D, Schmidt-Krey I, Lieberman RL. Amyloid fibril formation by the glaucoma-associated olfactomedin domain of myocilin. J Mol Biol 2012;421(2-3):242-255.
23. Kang SJ, Yi JH, Hong HS, Jang SH, Park MH, Kim HJ, Lee KY, Lee YJ, Han SW, Koh SH. Secondary amyloidosis associated with multiple sclerosis. J Clin Neurol 2009;5(3):146-8.
24. Fowler DM, Koulov AV, Balch WE, Kelly JW. Functional amyloid Aβ from bacteria to humans. Trends Biochem Sci 2007;32(5):217-224.
25. Chiti F, Dobson CM. Protein misfolding, functional amyloid, and human disease. Annu Rev Biochem 2006;75(1):333-66.
26. Pryszcz LP, Huerta-Cepas J, Gabaldon T. MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. Nucleic Acids Res 2011;39(5):e32.
27. Nicholas KB, Nicholas HB, Deerfield DW. GeneDoc: analysis and visualization of genetic variation. EMBNEW. NEWS 1997;4:14.
28. Xiong J-P, Stehle T, Diefenbach B, Zhang R, Dunker R, Scott DL, Joachimiak A, Goodman SL, Arnaout MA. Crystal structure of the extracellular segment of integrin αV β3. Science 2001;294(5541):339-345.

29. Zhang K, Chen J. The regulation of integrin function by divalent cations. Cell Adh Migr 2012;6(1):20-29.
30. Downing AK, Knott V, Werner JM, Cardy CM, Campbell ID, Handford PA. Solution structure of a pair of calcium-binding epidermal growth factor-like domains: implications for the Marfan syndrome and other genetic disorders. Cell 1996;85(4):597-605.
31. Handford PA, Mayhew M, Baron M, Winship PR, Campbell ID, Brownlee GG. Key residues involved in calcium-binding motifs in EGF-like domains. Nature 1991;351(6322):164-167.
32. Pejaver V, Hsu W-L, Xin F, Dunker AK, Uversky VN, Radivojac P. The structural and functional signatures of proteins that undergo multiple events of post-translational modification. Protein Sci 2014;23(8):1077-1093.
33. Ceroni A, Passerini A, Vullo A, Frasconi P. DISULFIND: a disulfide bonding state and cysteine connectivity prediction server. Nucleic Acids Res 2006;34(suppl 2):W177-W181.
34. Ferrè F, Clote P. DiANNA: a web server for disulfide connectivity prediction. Nucleic Acids Res 2005;33(suppl 2):W230-W232.
35. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: protein structure and function prediction. Nat Meth 2015;12(1):7-8.
36. Schrödinger L. The PyMOL molecular graphics system, Version 1.5.0.3 2012.
37. Tartaglia GG, Vendruscolo M. The Zyggregator method for predicting protein aggregation propensities. Chem Soc Rev 2008;37(7):1395-1401.
38. Tsolis AC, Papandreou NC, Iconomidou VA, Hamodrakas SJ. A consensus method for the prediction of 'aggregation-prone' peptides in globular proteins. PLoS ONE 2013;8(1):e54175.
39. Martins RST, Pinto PIS, Guerreiro PM, Zanuy S, Carrillo M, Canário AVM. Novel galanin receptors in teleost fish: Identification, expression and regulation by sex steroids. Gen Comp Endocrinol 2014;205:109-120.
40. Gill SC, von Hippel PH. Calculation of protein extinction coefficients from amino acid sequence data. Anal Biochem 1989;182(2):319-26.
41. Ornstein L. Disc Electrophoresis-I Background and Theory. Annals of the New York Academy of Sciences 1964;121(2):321-349.

42. Klement K, Wieligmann K, Meinhardt J, Hortschansky P, Richter W, Fändrich M. Effect of different salt ions on the propensity of aggregation and on the structure of Alzheimer's abeta(1-40) amyloid fibrils. J Mol Biol 2007;373(5):1321-1333.
43. Huang LH, Cheng H, Pardi A, Tam JP, Sweeney WV. Sequence-specific 1H NMR assignments, secondary structure, and location of the calcium binding site in the first epidermal growth factor like domain of blood coagulation factor IX. Biochemistry 1991;30(30):7402-7409.
44. Cooke RM, Wilkinson AJ, Baron M, Pastore A, Tappin MJ, Campbell ID, Gregory H, Sheard B. The solution structure of human epidermal growth factor. Nature 1987;327(6120):339-341.
45. Louro B, Power DM, Canario AVM. Advances in European sea bass genomics and future perspectives. Marine Genomics 2014;18, Part A:71-75.
46. Rice GP, Hartung HP, Calabresi PA. Anti-alpha4 integrin therapy for multiple sclerosis: mechanisms and rationale. Neurology 2005;64(8):1336-42.
47. Kawamoto E, Nakahashi S, Okamoto T, H. I, Shimaoka M. Anti-Integrin therapy for Multiple Sclerosis. Autoimmune Dis 2012;2012:1-6.
48. Ikai A. Thermostability and aliphatic index of globular proteins. J Biochem 1980;88(6):1895-1898.
49. Reddy CS, Achary VMM, Manna M, Singh J, Kaul T, Reddy MK. Isolation and molecular characterization of thermostable phytase from Bacillus subtilis (BSPhyARRMK33). Appl Biochem Biotechnol 2015;175(6):3058-3067.
50. Wang S, Dong Z-Y, Yan Y-B. Formation of high-order oligomers by a hyperthemostable Fe-superoxide dismutase (tcSOD). PLoS ONE 2014;9(10):e109657.
51. Srinivas PN, Patil MA, Reddy GB. Temperature-dependent coaggregation of eye lens α B- and β -crystallins. Biochem Biophys Res Commun 2011;405(3):486-490.
52. Liu Z, Huang Y. Advantages of proteins being disordered. Protein Sci 2014;23(5):539-50.
53. Abeln S, Frenkel D. Disordered flanks prevent peptide aggregation. PLoS Comput Biol 2008;4(12):e1000241.

54. Steck E, Benz K, Lorenz H, Loew M, Gress T, Richter W. Chondrocyte expressed protein-68 (CEP-68), a novel human marker gene for cultured chondrocytes. Biochem J 2001;15(353(Pt 2)):169-74.
55. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol 1982;157(1):105-132.
56. Gasteiger E, Hoogland C, Gattiker A, Duvaud Se, Wilkins M, Appel R, Bairoch A. Protein identification and analysis tools in the ExPASy server. In: Walker J, editor. The Proteomics Protocols Handbook. Totowa, NJ: Humana Press; 2005. p 571-607.
57. Guruprasad K, Reddy BVB, Pandit MW. Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. Protein Eng 1990;4(2):155-161.

Figure Legends

Fig. 1. A- Structural models obtained using I-TASSER for the cyanobacter (*Synechococcus sp*) (syCRTAC1), sea bream (saCrtac1b), sea bass (dlCrtac1a and dlCrtac1b), human (hCRTAC1) mature CRTAC1 proteins. The seven beta-propeller blades are shown in different colours and the black areas represent the Ca-binding loops. Grey strands in hCRTAC1and dlCrtac1a represent the EGF-like domain missing in the remaining proteins. B- Percentage of sequence identity and root mean square deviation (RMSD) between predicted models, showing in brackets the number of equivalent carbon-α atoms for each value.

Fig 2. A- Far-UV spectra of recombinant human and piscine CRTAC1's (5uM) in 10mM Tris-HCl pH8 at room temperature (25ºC) scanned from 250 to 180nm. B- Temperature scans of human and piscine CRTAC1's (5uM) ellepticity from 20 to 90ºC at 210 nm in water. C- Intrinsic fluorescence emission spectra of human and piscine CRTAC1's at 20 and 90ºC. D- Intrinsic fluorescence emission spectra of human and piscine CRTAC1's in water with and without pre-incubation in GnHCl 6M.

Fig. 3. Analysis of human and piscine CRTAC1 aggregation. **A-** Size exclusion chromatographs (SEC) of purified recombinant dlCrtac1a, dlCrtac1b and hCRTAC1 in Tris 50 mM pH8, NaCl 250 mM. Two main peaks were evident: a high molecular weight species that corresponded to protein aggregates (7.7ml for hCRTAC1, 7.9ml for dlCrtac11 and 7.6ml for dlCrtac1b) and a lower molecular weight species that corresponded to monomeric CRTAC1's (14.6 ml for hCRTAC1, 14.7 ml for dlCrtac1aand 15.2 ml for dlCrtac1b). Molecular weight markers used for calibration in the first panel of A were: Thyroglobulin (670 kDa), γ-globulin (158 kDa), Ovalbulmin (44 kDa), Myoglobin (17 kDa), Vitamin B12 (1.35 kDa). **B-** Validation of SEC peak samples for hCRTAC1, dlCrtac1a1 and dlCrtac1b using discontinuous Ornstein-Davis 8% polyacrylamide gels (Native-PAGE). Samples (10 μg) of the 1st and 2nd peaks of each CRTAC1 collected during SEC were mixed with an equal volume of 0.125 M Tris-HCl (pH 6.8), 10% glycerol and 0.01% of bromophenol blue and resolved on a polyacrylamide gel. The gel was stained with Coomassie blue for aggregates (1st peak) and double stained with silver nitrate for monomers (2nd peak).

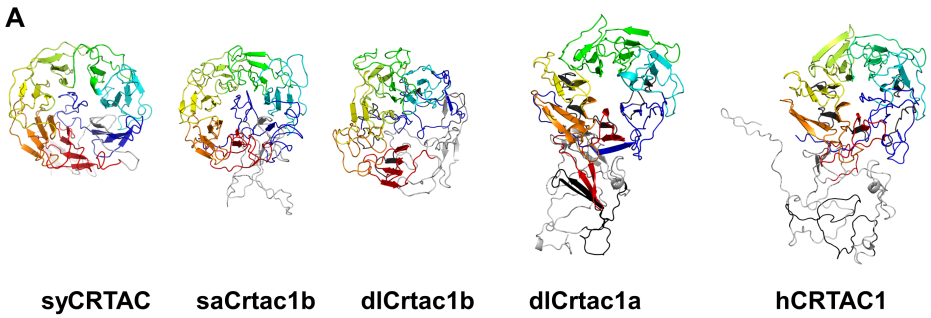
Fig. 4. Characterization of CRTAC1 aggregates by A- TEM: transmission electron micrographs of hCRTAC1, dlCrtac1a and dlCrtac1b in 100 mM Tris pH8, 250 mM NaCl before (T=0) and after

incubation at 37ºC for one and two weeks and for one week at 37ºC in the presence of 50 mM $CaCl_2$ as indicated in the figure.

Table legend

Table 1. Predicted physico-chemical properties and aggregation propensity of human CRTAC1 and teleost dlCrtac1a and dlCrtac1b (mature protein).

Fig.1



B

	hCRTAC1	dlCrtac1a	dlCrtac1b	saCrtac1b	syCRTAC
hCRTAC1		65%	57%	56%	35%
dlCrtac1a	3.54 (504)		60%	60%	36%
dlCrtac1b	5.70 (192)	5.63 (168)		90%	41%
saCrtac1b	5.52 (184)	5.51 (184)	5.71 (208)		42%
syCRTAC	5.60 (184)	4.99 (176)	5.53 (200)	3.43 (400)	

Fig. 2

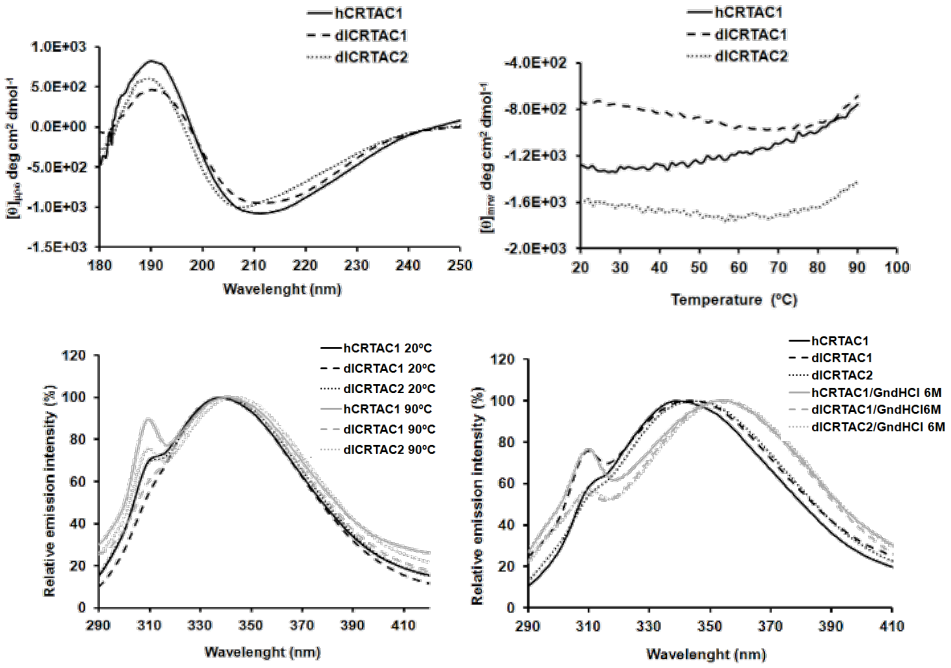

hCRTAC1
dICRTAC1
dICRTAC2
Wavelenght (nm)
Temperature (°C)
hCRTAC1 20°C
dICRTAC1 20°C
dICRTAC2 20°C
hCRTAC1 90°C
dICRTAC1 90°C
dICRTAC2 90°C
Relative emission intensity (%)
hCRTAC1/GndHCl 6M
dICRTAC1/GndHCl6M
dICRTAC2/GndHCl 6M

Fig.3

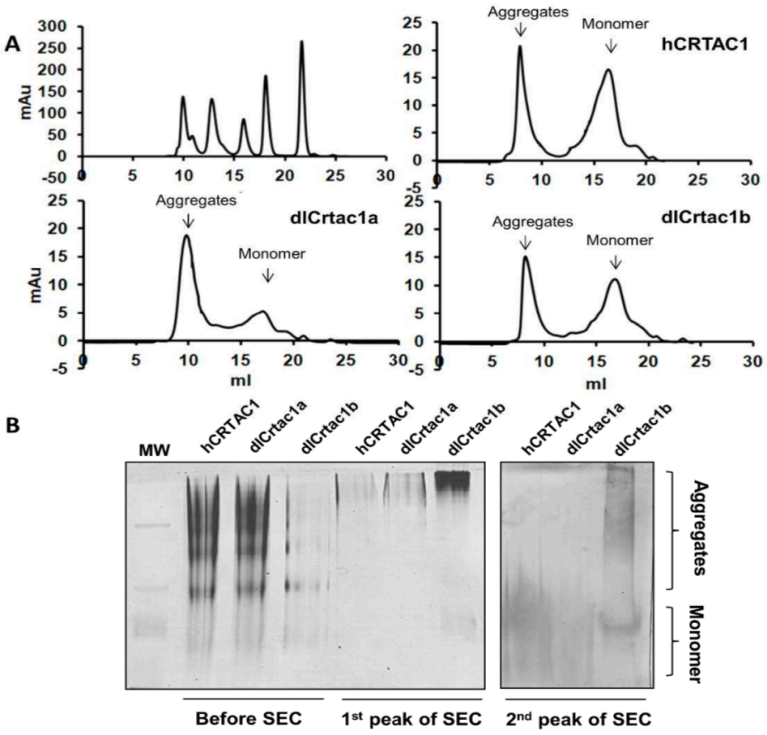

A
mAu
ml
Aggregates
Monomer
hCRTAC1
dICrtac1a
dICrtac1b
B
MW
Aggregates
Monomer
Before SEC
1st peak of SEC
2nd peak of SEC

Fig.4

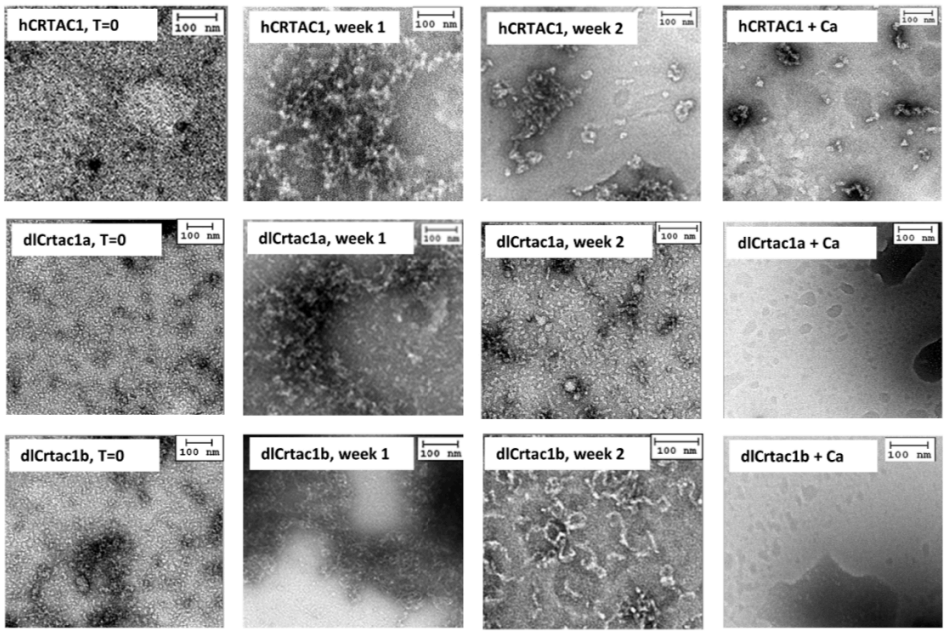

hCRTAC1, T=0
100 nm
hCRTAC1, week 1
100 nm
hCRTAC1, week 2
100 nm
hCRTAC1 + Ca
100 nm
dlCrtac1a, T=0
100 nm
dlCrtac1a, week 1
100 nm
dlCrtac1a, week 2
100 nm
dlCrtac1a + Ca
100 nm
dlCrtac1b, T=0
100 nm
dlCrtac1b, week 1
100 nm
dlCrtac1b, week 2
100 nm
dlCrtac1b + Ca
100 nm

Table 1

	hCRTAC1 *(Homo sapiens)*	**dlCrtac1a** *(Dicentrarchus labrax)*	**dlCrtac1b** *(Dicentrarchus labrax)*
Amino acid identity % with	dlCRTAC1:68%	dlCRTAC2:61%	hCRTAC1:59%
Molecular weight (kDa)	68.6	68	56.9
Amino acid number and composition	635/ 14 Cys; 4 Trp; 22 Tyr; 27 Phe	620/ 12 Cys; 5 Trp; 22 Tyr; 27 Phe	529/ 6 Cys; 3 Trp; 22 Tyr; 24 Phe
Disulfide bridges	>5	>5	3
Theoretical pI (+/- charge)	4.97 (60/83)	5.20 (64/83)	4.94 (48/66)
GRAVY*	-0.349	-0.448	-0.331
Instability index**	35.29-Stable	36.20 Stable	24.51 Stable
Glycosylation sites[v]	1 N-Glyc 3 O-Glyc	2 N-Glyc 1 O-Glyc	2 N-Glyc 1 O-Glyc
Kinase specific phosphorylation sites	5 PKC 3 PKA 1 CDK5	6 PKC 2 PKA	3 PKC
Aggregation propensity (Zyggregator)¶	0.697	0.804	0.801
N° aggregation "hot spots" (Amylpred)¥	18	16	14
Aggregation "hot spots" in the N-term beta-propeller	15	12	12

* *GRAVY*, Grand average of hydropathicity, according to [55] and [56] in http://web.expasy.org/protparam/

** *Instability index* – according to [56,57] in http://web.expasy.org/protparam/ a protein with an instability index smaller than 40 is stable, a value above 40 predicts that the protein may be unstable. Aggregation threshold = 1[37];

¶ Aggregation threshold=1[37]

¥ [38]

Supplementary Material

Fig.S1

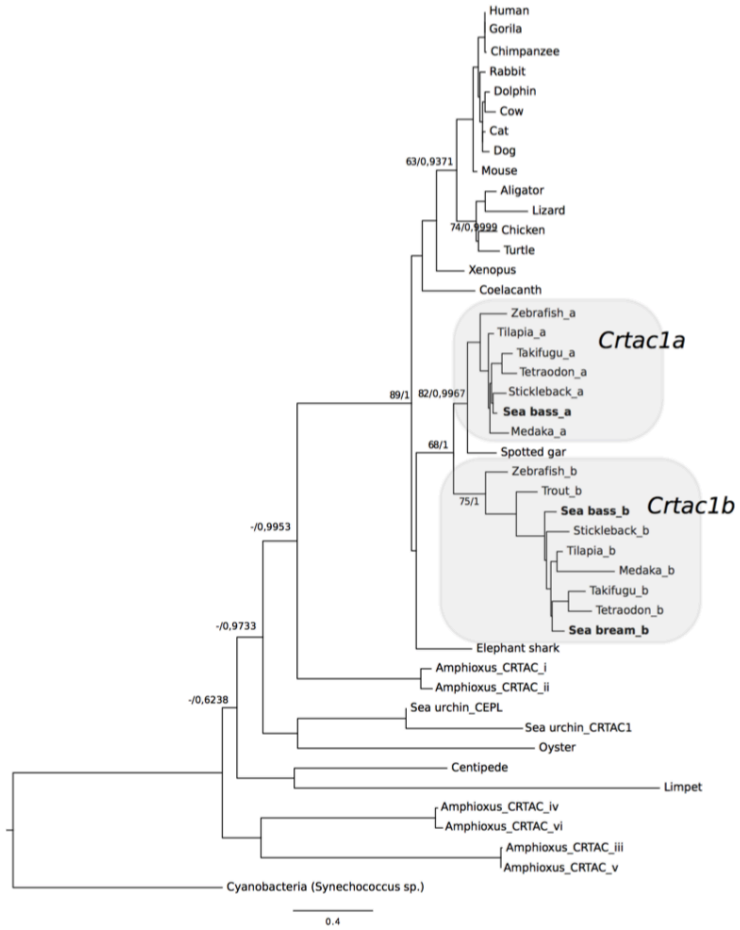


Fig. S1. Phylogenetic analysis of CRTAC1 in metazoans. The phylogenetic tree was constructed using the maximum likelihood (ML) method and the reliability of internal branching was assessed using bootstrap/aBayes methods. Only statistical branch support for the main vertebrate clades is represented. The duplicate teleost genes were named *crtac1a* and *crtac1b*. Duplicate *CRTAC1* genes were identified in early deuterostomes, amphioxus (*Branchiostoma floridae*) and sea urchin (*Strongylocentrotus purpuratus*) genomes. In amphioxus CRTAC1_i and CRTAC1_ii are identical although they map to different chromosome regions and the same occurs for the paralogues CRTAC1_iii and CRTAC1_v and also for CRTAC1_iv and CRTAC1_vi (it remains to be established if they are all *bona fide* or result from genome assembly errors). Lamprey CRTAC1 was not included in the phylogenetic analysis as the sequence was very incomplete. The

phylogenetic tree was rooted with cyanobacteria (*Synechococcus sp.*) CRTAC1. The accession numbers of the sequences used for structural analysis are given in Table S1.

Fig.S2

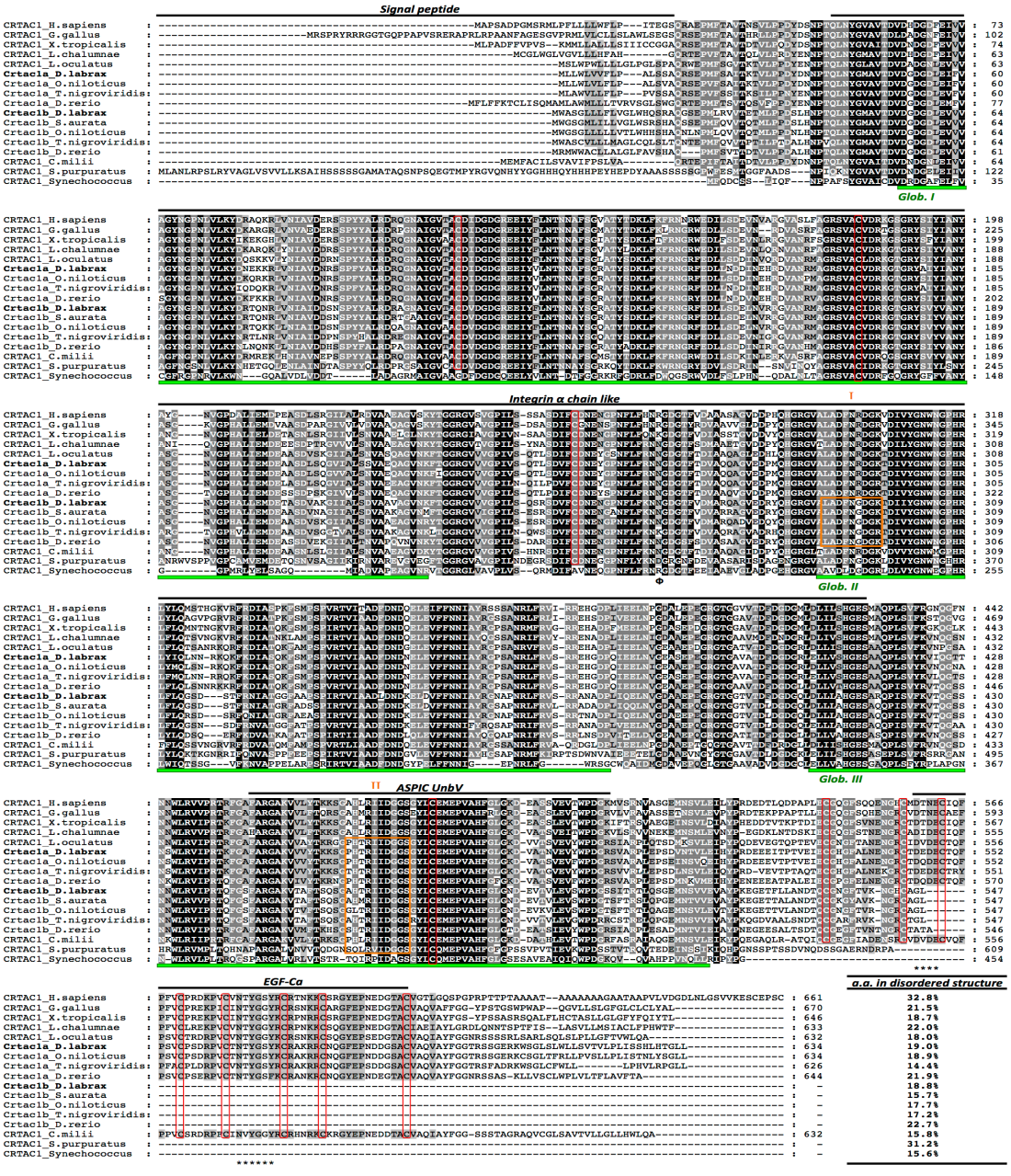


Fig. S2. Amino acid sequence alignments of CRTAC1 family members from prokaryotes to superior vertebrates. The conserved amino acid residues are shaded and conserved cysteine (C)

residues are visible in red boxes. The numbers on the right hand side indicate the position of the amino acid residues. The predicted signal peptide (except for sea urchin (*S. purpuratus*) and chicken (*G. gallus*) for which the cleavage site for signal peptide was not possible the prediction), integrin α chain like, ASPIC/UnbV and EGF-Ca binding domains are indicated above each alignment block. The EGF-Ca binding site consensus sequence (D/N–x–D/N–E/Q–x_m–D/N*–x_n–Y/F, where "x" is any residue, m and n are variable number and * indicates possible β-hydroxylation) in the EGF-Ca domain is indicated by an * below the aligned sequences. Ordered and disordered regions were predicted using GlobPlot software and the green horizontal bars below the alignment blocks delimit the consensus regions for the main globular domains (marked as Glob. I, II, III) shared between all CRTAC1 analyzed sequences. An exception was for Glob II region that was not predicted in *D. labrax*, *S. aurata* and *T. nigroviridis* Crtac1b. Orange boxes, delimits the potential disordered regions exclusively found in teleosts Crtac1b (I) or in all fish species (except *L. chalumnae and S. aurata*) and sea urchin (*S. purpuratus*) CRTAC1 (II) sequences. On the right side, in bottom is tabled the amino acid percentage (%) involved in disordered structure for each CRTAC1 sequence. The accession numbers of all the sequences used in this alignment are shown in supplementary Table S1.

Fig.S3.

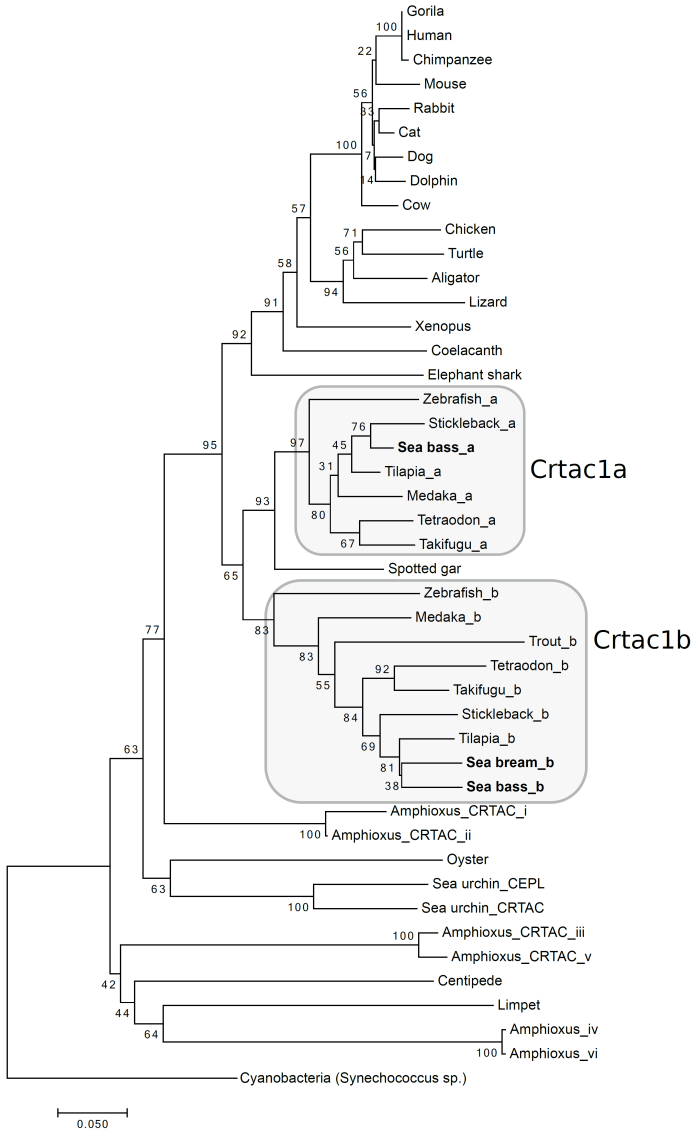


Fig. S3. Phylogenetic tree of CRTAC1 in metazoans obtained with the Neighbor Joining method. Evolutionary analysis was conducted in MEGA7 and the evolutionary distances were computed using the p-distance method. The rate variation among sites was modeled with a gamma distribution (shape parameter = 0.6). The percentages of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches.

Fig. S4.

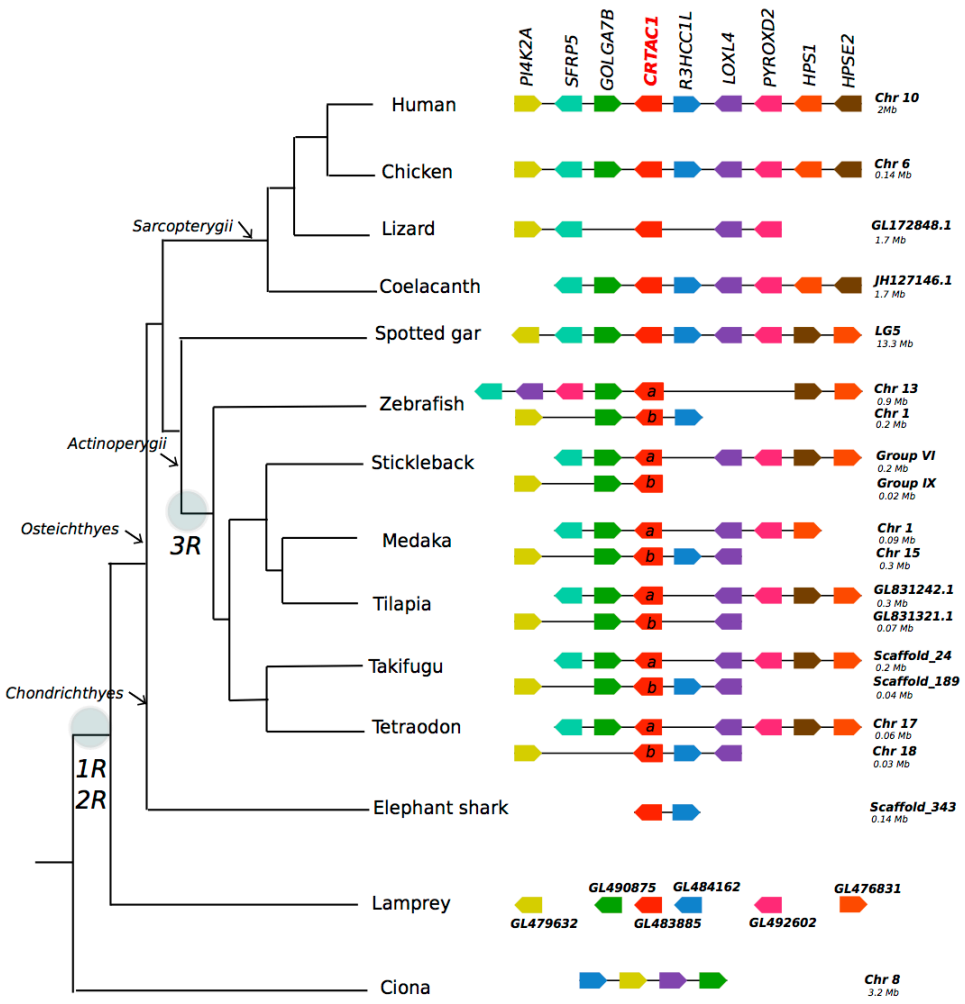


Fig. S4. Gene synteny of *CRTAC1's* in vertebrates and *Ciona* genomes. The neighboring gene environment of *CRTAC1* in the human, chicken, *Xenopus* and coelacanth and 7 *Actinopterygii* species (Zebrafish, Stickleback, Takifugu, Tetraodon, Medaka, Tilapia and Spotted gar) and the cartilaginous Elephant shark and the jawless lamprey is compared. The urochordate Ciona (*Ciona intestinallis*) homologue region is also represented but the *CRTAC1* gene has been lost. In the amphioxus (*Branchiostoma floridae*) multiple *CRTAC1*-like genes were retrieved but no conserved gene synteny with vertebrates was found. In the figure, genes are represented according to their order in the genome and gene homologues are represented in the same color. Arrowheads indicate the direction of gene transcription and the lines represents chromosomes. Chromosomes

and the relative length of the genome fragment (Mb) analyzed are indicated. Gene names: Phosphatidylinositol 4-kinase type 2-alfa (*PI4K2A*), Secreted frizzled-related protein 5 (*SFRP5*), golgin A7 family, member B (*GOLGA7B*), R3H domain and coiled-coil containing 1(*R3HCC1*), Lysyl oxidases lysyl oxidase like (*LOXL*), Pyridine nucleotide-disulphide oxidoreductase domain 2 (*PYROXD2*), Hermansky-Pudlak Syndrome 1(*HPS1*), Heparanase (*HPSE*).

Fig.S5.

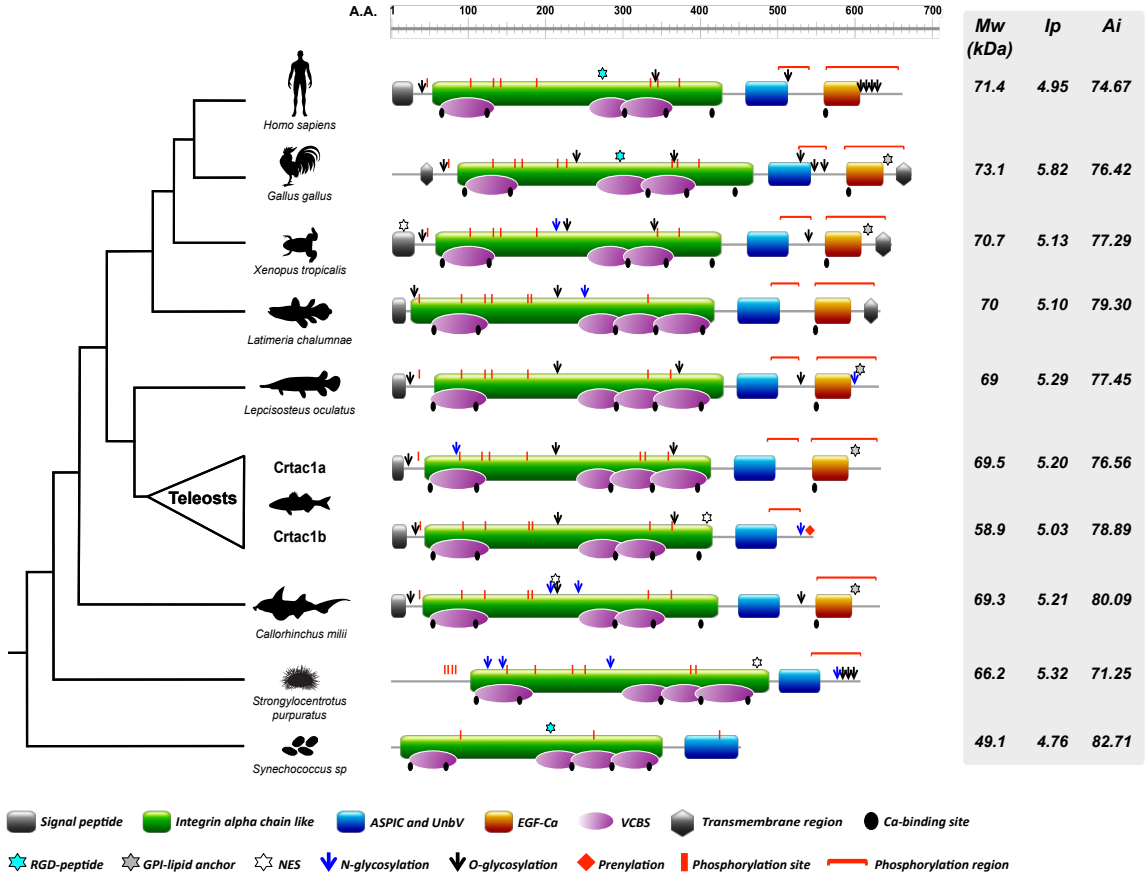


Fig. S5. Dendogram comparing the predicted structural and biochemical features of the CRTAC1 protein family. Conserved domains/motifs in the amino acid sequences of CRTAC1 from representative organisms of prokaryotes to vertebrates are indicated as coloured blocks: the signal peptide (black); N-terminal integrin α chain like-domain (green); ASPIC and UnbV domain (pfam07593, blue); EGF-Ca, calcium-binding epidermal growth factor domain (pfam07645,

orange); α-helical transmembrane region (black hexagon); VCBS, repeat domain in Vibrio, Colwellia, Bradyrhizobium and Shewanella (pfam13517, pink elliptic forms). The consensus regions for small functional motifs are shown in blue and white stars indicate RGD (Arg-Gly-Asp) and NES (leucine-rich nuclear export signals) and calcium (Ca)- binding sites are denoted by black circles. Sites of post-translational modifications (PTMs) were specifically investigated for: a) *N*-linked (GlcNAc, D-N-acetylglucosamine) and *O*-linked glycosylation (GalNAc, D-N-acetylgalactosamine) using NetNGlyc v.1.0 and NetOGlyc v.4.0 software (http://www.cbs.dtu.dk/ [1, 2]; b) generic phosphorylation using NetPhos v.2.0 or NetPhosBac v.1.0 servers (http://www.cbs.dtu.dk/ [1, 3]; c) prenylation (http://mendel.imp.ac.at/PrePS/, [4]; d) glycosylphosphatidylinositol (GPI)-lipid anchoring [2] modification using big predictor (http://mendel.imp.ac.at/sat/gpi/gpi_server.html); e) *Net*NES v1.1 server (http://www.cbs. dtu.dk/ [5] was used to identify the leucine-rich nuclear exporting signal (NES, LxxLxL or LxxxLxL, where L can be either L, I, V, F or M) setting the minimum cut-off value at 0.7 (on a scale of 0 and 2). PTMs of CRTAC1 were scanned using ModPred software (http://montana.informatics.indiana.edu/ModPred/index.html, [6] with a cut-off threshold ≥ 0.5 (score between 0-1) and are indicated and described in the figure footnote: phosphorylation sites or regions (red vertical/horizontal bars); N-linked and O-linked glycosylation (blue and black arrows); prenylation (red square) and Glycosylphosphatidylinositol (GPI) lipid anchoring (grey stars).

Signal peptide, molecular weight (Mw), isoelectric point (Ip) and aliphatic index (Ai) were determined using SignalP v.4.1 (http://www.cbs.dtu.dk/services/SignalP/, [7] and ProtParam (http://web.expasy.org/protparam/, [8]. Predictions of globularity and disordered regions within the CRTAC1 protein were established with GlobPlot v.2.3 (http://globplot.embl.de, [9]. PROSITE MyDomains image creator software (http://prosite.expasy.org/prosite.html) was used to build the CRTAC1 structural outlines.

The scale on the top indicates the number of amino acids (A.A). The molecular weight (Mw, in kDa), isolelectric point (Ip) and aliphatic index (Ai, in %) are indicated on the right hand side for all CRTAC1 amino acid sequences. Dicentrarchus labrax Crtac1a and b modular architecture is indicated and is representative of teleost Crtac1a and b such as *Oreochromis niloticus*, *Danio rerio*, *Tetraodon nigroviridis* and *Sparus aurata*.

Note: the Mw/Ip/Ai for *Oreochromis niloticus*, *Danio rerio*, *Tetraodon nigroviridis* and *Sparus aurata* (not given in the figure) were respectively for Crtac1a, 69.4/5.25/76.72, 71/5.21/73.09, 69/5.42/78 (not available for Sparus aurata) and for Crtac1b, 59.4/4.88/79.3, 59/4.83/78.97, 59/5.33/80.18, 59/5.03/77.37, following the same order of species in brackets. The accession numbers of the sequences used for the analysis are given in Table S1.

Supplementary text

CRTAC1 protein evolution

Several consensus VCBS domains (multiple copies found in proteins from Vibrio, Colwellia, Bradyrhizobium and Shewanella) were distributed along the N-terminal integrin α chain like-domain in each CRTAC1 protein sequence (Fig. S5). The number of VCBS repeats varied between 3 in terrestrial vertebrates, teleosts (Crtac1b) and elephant shark CRTAC1 and 4 in the remaining organisms (teleost Crtac1a, non-teleosts, sea urchin and cyanobacteria). The role of VCBS is unclear, but it has been described in proteins with FG-GAP repeat domains that are known to be involved in colony development and biofilm formation such as the biofilm-associated extracellular matrix protein (Bap1) in bacteria and supports the proposed function of CRTAC1 in cell adhesion. Hydrophobic amino acid sequences characteristic of α-helical transmembrane regions were predicted in chicken, xenopus and coelacanth (*Latimeria chalumnae*) CRTAC1 (Fig. S5) to be localized mainly in the divergent C-terminal region. A transmembrane region exists in the mouse brain–specific CRTAC1B (named as LOTUS) and is a potent blocker of the Nogo receptor-1. Several post-translational modifications (PTMs) were predicted in vertebrate, invertebrate and prokaryote CRTAC1 (Fig. S5). Notable PTMs include *O*-linked glycosylation sites (2 to 7, with the exception of CRTAC1 in cyanobacteria) and accordingly human chondrocytes were found to secrete glycosylated CRTAC1. CRTAC1 *in vivo* is likely to be highly phosphorylated since *in silico* predictions revealed 20 phosphorylation sites in spotted gar CRTAC1, 15 – 13 sites in sea urchin, human, chicken, xenopus and teleost (Crtac1a) CRTAC1, 9 sites in coelacanth, elephant shark and teleost Crtac1b and 3 sites in cyanobacteria CRTAC1.

Fig.S6.

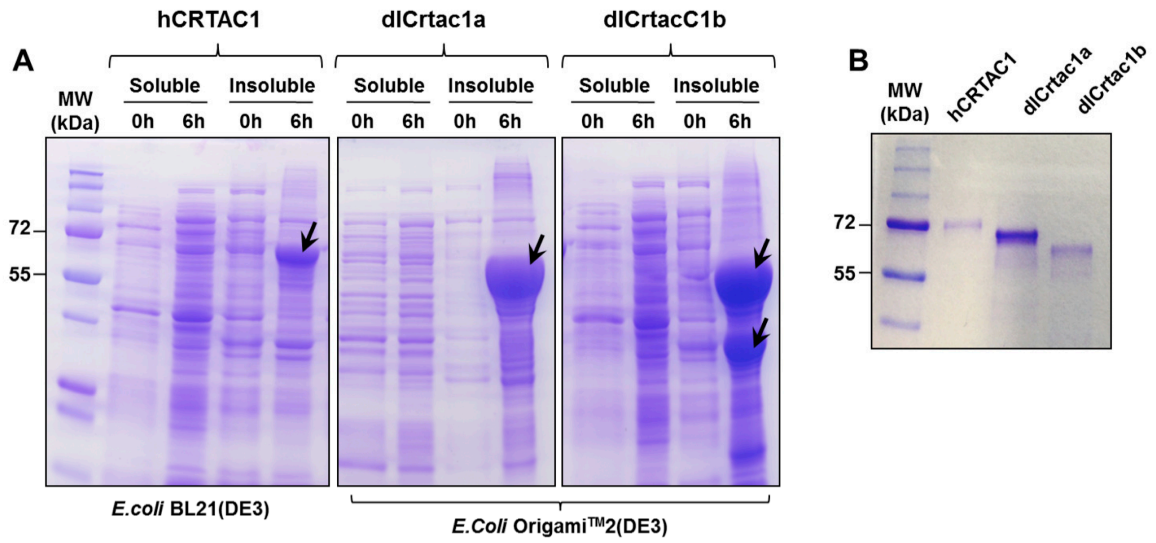


Fig. S6. Analysis of expression and purification of recombinant proteins hCRTAC1, dlCRTAC1a and dlCRTAC1b by Coomassie blue stained SDS-PAGE (10% polyacrylamide, reducing conditions). **A-** Soluble and insoluble protein bacterial fractions recovered after 0 and 8 h induction with 0.8 mM IPTG. Note the abundant CRTAC1s (55-70 kDa, black arrows) in the insoluble bacterial fraction. **B-** Recombinant CRTAC1s (0.5 or 1 μg) purified by continuous elution electrophoresis. Isolated hCRTAC1, dlCRTAC1a and dlCRTAC1b recombinant proteins with a molecular weight of 68.6, 68 and 56.9 kDa. MW – molecular weight markers (kDa).

Fig.S7.

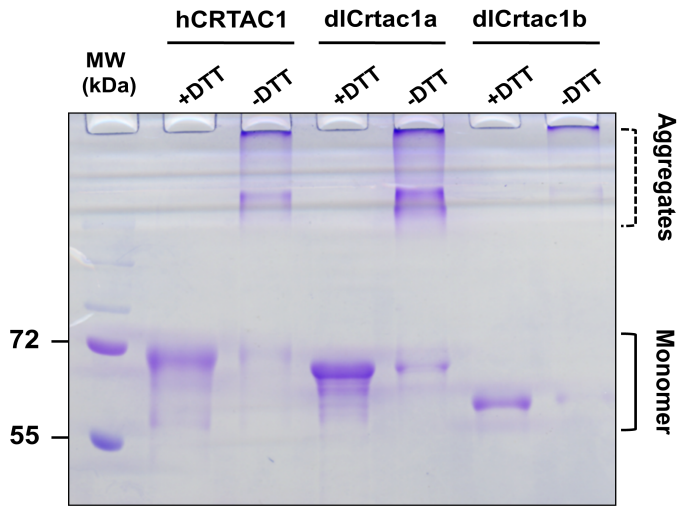


Fig. S7. Analysis of the effect of a reducing agent (DTT) in CRTAC1 aggregate formation. Purified, refolded recombinant hCRTAC1, dlCrtac1a and 1b (5 µg, before separation by size exclusion chromatography) were fractionated in a 10% SDS-polyacrilamide gel without stacking gel under reducing and non-reducing conditions in the presence or absence of DTT (50 mM) and without thermal denaturation). Note monomeric (filled arrows) and aggregated (dashed arrows) forms of all CRTAC1´s in the presence and absence of DTT respectively.

Fig.S8.

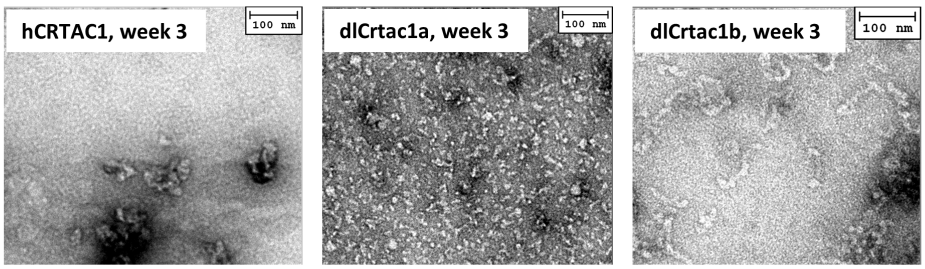


Fig. S8. **Characterization of CRTAC1's aggregates by TEM**: transmission electron micrographs of hCRTAC1, dlCrtac1a and dlCrtac1b in 100 mM Tris pH8, 250 mM NaCl after incubation at 37ºC for three weeks.

Table S1. List of the CRTAC1 sequences used in the evolutionary analysis.

Genomes of 8 mammals were queried using the human CRTAC1 protein (NP_060528) to find CRTAC1 homologues. The sea bass (*Dicentrarchus labrax*, dl) *crtac1* genes were extracted by blast search from the European sea bass genome assembly (http://seabass.mpipz.de/cgibin/hgGateway?org=European+seabass&db=dicLab1) but using the pufferfish (*Tetraodon nigroviridis*) *crtac1* nucleotide sequences. The genomic sequences were extracted (from linkage group LG11 and LG7 respectively) and Spidey software (http://www.ncbi.nlm.nih.gov/spidey/), BCM search launcher (http://searchlauncher.bcm. tmc.edu), BlastX (www.ncbi.nih.gov\BLAST\) and ClustalX v.1.8 [10] used to obtain the open reading frame (ORF). Similar searches were also performed in the genomes of two cartilaginous fishes, the Elephant shark (*Callorhinchus milii*, www. esharkgenome.imcb.a-star.edu.sg) and the little skate (*Leucoraja erinacea*, http://skatebase.org) and also of two jawless fishes, the Marine lamprey (*Petromyzon marinus,* www. ensembl.org) and the Japanese lamprey (*Lethenteron japonicum*, http://jlampreygenome.imcb.a-star.edu.sg). The search for *CRTAC1* was further extended to invertebrates and to the assembled genomes of the early deuterostomes, the urochordate (*Ciona intestinalis*, www.ensembl.org), the cephalochordate (*Branchiostoma floridae,* http://genome.jgi.doe.gov) and the hemichordate (*Strongylocentrotus purpuratus*, www.ensembl.genomes.org) and also to the protostomes (molluscs, annelids, arthropods and nematodes) available from ENSEMBL genomes (www.ensemblgenomes.org). Similar searches against the NCBI database (http://www.ncbi.nlm.nih.gov) were performed to confirm gene predictions and identify other vertebrate CRTAC1 family members. The status of the genome projects (according to Genomes online database, https://gold.jgi.doe.gov) and genome assemblies searched are indicated.

Species	Symbol	Gene ID (genome project status, assembly)	Protein ID
MAMMALS			
Human (*Homo sapiens)*	*CRTAC1*	ENSG00000095713 (Complete and published, GRCh38.p7)	NP_060528
Chimpanzee (*Pan troglodytes*)	*crtac1*	ENSPTRG00000002821 (Incomplete, CHIMP2.1.4)	XP_507967
Gorilla	*crtac1*	ENSGGOG00000016011	XP_004049964

(*Gorilla gorilla gorilla*)		(Permanent draft, gorGor3.1)	
Mouse (*Mus musculus*)	*crtac1*	ENSMUSG00000042401 (Complete and published, GRCm38.p4)	NP_660105
Rabbit (*Oryctolagus cuniculus*)	*crtac1*	ENSOCUG00000004978 (Complete and published, OtyCun2.0)	XP_008268532
Domestic cat (*Felis catus*)	*cratc1*	ENSFCAG00000013435 (Permanent draft, Felis_catus_6.2)	XP_006938122
Dog (*Canis lupus familiaris*)	*crtac1*	ENSCAFG00000009271 (Permanent draft, CanFam3.1)	XP_851054
Cow (*Bos taurus*)	*crtac1*	ENSBTAG00000008102 (Permanent draft, UMD3.1)	NP_001192254
Dolphin (*Tursiops truncatus*)	*crtac1*	ENSTTRG00000009905 (Permanent draft, turTru1)	ni
BIRDS/REPTILES			
Chicken (*Gallus gallus*)	*crtac1*	ENSGALG00000017378 (Permanent draft, v4.0)	NP_001073680
Anole lizard (*Anolis carolinensis*)	*crtac1*	ENSACAG00000003141 (Permanent draft, v2.0)	XP_008104723
Painted turtle (*Chrysemys picta bellii*)	*crtac1*	–	XP_008178051
Chinese aligator (*Alligator sinensis*)	*crtac1*	–	XP_006033800
AMPHIBIAN			
Xenopus (*Xenopus tropicalis*)	*crtac1*	ENSXETG00000030508 (Permanent draft, v4.2)	NP_001072373
COELACANTH			
Coelacanth (*Latimeria chalumnae*)	*crtac1*	ENSLACG00000011568 (Permanent draft, v1.0)	XP_006000251
TELEOST			
Takifugu (*Takifugu rubripes*)	*crtac1a*	ENSTRUG00000006008 (Permanent draft, v4.0)	NP_001092134
	crtac1b	ENSTRUG00000006661 (Permanent draft, v4.0)	NP_001092133
Medaka (*Oryzias latipes*)	*crtac1a*	ENSORLG00000011469 (Permanent draft, HdrR)	NP_001098377
	crtac1b	ENSORLG00000007646 (Permanent draft, HdrR)	XP_004077428
Tetraodon (*Tetraodon nigroviridis*)	*crtac1a*	ENSTNIG00000008683 (Permanent draft, v8.0)	ABC86213
	crtac1b	ENSTNIG00000014083 (Permanent draft, v8.0)	ABC86207
Stickleback (*Gasterosteus aculeatus*)	*crtac1a*	ENSGACG00000007647 (Permanent draft, BROAD S1)	ni
	crtac1b	ENSGACG00000018906 (Permanent draft, BROAD S1)	NP_001254573
Tilapia (*Oreochromis niloticus*)	*crtac1a*	ENSONIG00000018358 (Permanent draft, v1.0)	XP_013130985
	crtac1b	ENSONIG00000017670	XP_005460577

		(Permanent draft, v1.0)	
Sea bream (*Sparus aurata*)	*crtac1b*	–	ABD37673
Sea bass (*Dicentrarchus labrax*)	*crtac1a*	LG11:10210771-10228651* (Permanent draft, dicLab1)	ni
	crtac1b	LG7:10413305-10419561* (Permanent draft, dicLab1)	ni
Zebrafish (*Danio rerio*)	*crtac1a*	ENSDARG00000102517 (Complete and published, GRCz10)	XP_693888
	crtac1b	ENSDARG00000059826 (Complete and published, GRCz10)	NP_001073647
Trout (*Oncorhynchus mykiss)*	*crtac1b*	–	ABC86205
Spotted gar (*Lepisosteus oculatus*)	*crtac1*	ENSLOCG00000008921 (Complete and published, LepOcu1)	XP_006630619
CARTILAGINOUS FISH			
Elephant shark (*Callorhinchus milii*)	*crtac1*	SINCAMG00000011662 (Complete and published, ESHARK1)	XP_007908947
AGNATHA			
Lamprey (*Petromyzon marinus*)	*crtac1*	ENSPMAG00000008276 (Incomplete, v7.0)	ni
EARLY DEUTEROSTOMES			
Amphioxus (*Branchiostoma floridae*)	CRTACi	Brafl1|scaffold_349:79838-81394* (Permanent draft, v1.0)	XP_002604669
	CRTACii	Brafl1|scaffold_258:850585-852099* (Permanent draft, v1.0)	XP_002604669
	CRTACiii	Brafl1|scaffold_16:1885434-1939801* (Permanent draft, v1.0)	XP_002600738
	CRTACiv	Brafl1|scaffold_16:1885554-1940140* (Permanent draft, v1.0)	XP_002600740
	CRTACv	Brafl1|scaffold_121:1090526-1141607 * (Permanent draft, v1.0)	XP_002600738
	CRTACvi	Brafl1|scaffold_121:1090646-1141676* (Permanent draft, v1.0)	XP_002600740
Purple Sea urchin (*Strongylocentrotus purpuratus)*	SP-CEPL	SPU_028181 (Incomplete, v3.1)	NP_001074439
	SP-CRTAC1	SPU_024761 (Incomplete, v3.1)	ni
PROTOSTOMES			
Pacific oyster (Crassostrea gigas)		CGI_10021335 (Permanent draft, oyster_v9)	EKC40451
Owl limpet (*Lottia gigantea)*		LotgiG129677 (Permanent draft, lotgi1)	XP_009063122
Coastal European centípede (*Strigamia maritima*)		SMAR002215 (Permanent draft, smar1)	ni
CYANOBACTERIA			

Cyanobacteria bacterium Yellowstone A-Prime *(Synechoccus sp. JA-3-3Ab)*		CYA_0924 (Complete and published, ASM1320v1)	WP_011429809

* genome regions; ni: not identified; -: not available

Table S2. Primers used for amplification of original sea bass *crtac1a* and *1b* sequences and cloning of human and sea bass DNA sequences into the expression vector pET11a. Human CRTAC1 sequence was amplified by PCR but purchased and as a shuttle clone.

Proteins	Primers for amplification of full lenght sequence	Primers for cloning into pET11a vector
hCRTAC1 *(Homo sapiens)*	————————	Fw-5'CATATGTCC CAGCGGGCTGAACC3‘ Rev-5'GGATCCCTAGCAGCTGGGCTCGCAG3
dlCrtac1a *(Dicentrarchus labrax)*	Fw-5′ATGTTGTTGTGGCTGGTTGTGTTCC3′ Rev-5′TACAGTAGTCCAGTGTGGAGATGA3	Fw-5’CATATGCAGCGATCGGAGCCCATGT3‘ Rev-5'GGAT CCCTACAGTAGTCCAGTGTGG3’
dlCrtac1b *(Dicentrarchus labrax)*	Fw-5′ATGTGGGCTTCAGGTCTGTTGCTC3′ Rev-5′TCACAGACCTGCACAGTGTCCATTC3′	Fw-5'CATATGCAAGGCTCTGAGCCTATG3‘ Rev-5'GATCCTCACAGACCTGCACAGTG3

Table S3. Mass spectrometry analysis of recombinant CRTAC1’s. Peptides resultant from trypsin cleavage of putative recombinant hCRTAC1, dlCrtac1a and 1b (panel A, B and C respectively) analysed by MALDI-TOF, matched sequences predicted for CRTAC1´s by searching against databases via MASCOT search engine (http://www.matrixscience.com). The amino acid sequence of peptides resulting from trypsin cleavage of CRTAC1 proteins that match the sequences identified are shown in red.

A- hCRTAC1

Start - End	Mr Obs. (Da)	Mr expt. (Da)	Mr calc. (Da)	ppm	Miss	Peptide sequence
92 - 100	1028.5742	1027.5669	1027.5662	1	0	R.LVNIAVDER.S
101 - 108	956.4808	955.4735	955.4763	-3	0	R.SSPYYALR.D
101 - 110	1227.6362	1226.6289	1226.6044	20	1	R.SSPYYALRDR.Q
154 - 169	1929.9602	1928.9529	1928.9340	10	1	R.NNRWEDILSDEVNVAR.G
157 - 169	1545.7609	1544.7536	1544.7471	4	0	R.WEDILSDEVNVAR.G
170 - 178	877.4852	876.4779	876.4818	-4	0	R.GVASLFAGR.S
191 - 220	3307.6096	3306.6023	3306.5390	19	0	R.YSIYIANYAYGNVGPDALIEMDPEASDLSR.G
191 - 220	3323.6301	3322.6228	3322.5339	27	0	R.YSIYIANYAYGNVGPDALIEMDPEASDLSR.G
242 - 271	3250.5759	3249.5686	3249.5149	17	0	R.GVSVGPILSSSASDIFCDNENGPNFLFHNR.G
272 - 293	2180.0024	2178.9951	2178.9679	12	0	R.GDGTFVDAAASAGVDDPHQHGR.G
294 - 302	962.5046	961.4973	961.4981	-1	0	R.GVALADFNR.D
303 - 318	1826.9207	1825.9134	1825.8860	15	1	R.DGKVDIVYGNWNGPHR.L
306 - 318	1526.7922	1525.7849	1525.7426	28	0	K.VDIVYGNWNGPHR.L
333 - 346	1531.8479	1530.8406	1530.7864	35	1	R.DIASPKFSMPSPVR.T

339 - 346	920.4571	919.4498	919.4586	-10	0	K.FSMPSPVR.T
347 - 369	2748.3567	2747.3494	2747.3079	15	0	R.TVITADFDNDQELEIFFNNIAYR.S
382 - 403	2443.2300	2442.2227	2442.1775	19	1	R.REHGDPLIEELNPGDALEPEGR.G
383 - 403	2287.1155	2286.1082	2286.0764	14	0	R.EHGDPLIEELNPGDALEPEGR.G
437 - 447	1319.6290	1318.6217	1318.6167	4	0	R.GNQGFNNNWLR.V
452 - 459	925.4929	924.4856	924.4930	-8	1	R.TRFGAFAR.G
516 - 532	1891.9790	1890.9717	1890.9509	11	0	R.NVASGEMNSVLEILYPR.D

Putative hCRTAC1 - Match to: gi|530393911 cartilage acidic protein 1 isoform X1 [Homo sapiens]; Score: 169; Sequence Coverage: 39%; Nominal mass (M_r): 68627; Calculated pI value: 5.05

```
  1 MAPSADPGMS RMLPFLLLLW FLPITEGSQR AEPMFTAVTN SVLPPDYDSN
 51 PTQLNYGVAV TDVDHDGDFE IVVAGYNGPN LVLKYDRAQK RLVNIAVDER
101 SSPYYALRDR QGNAIGVTAC DIDGDGREEI YFLNTNNAFS GVATYTDKLF
151 KFRNNRWEDI LSDEVNVARG VASLFAGRSV ACVDRKGSGR YSIYIANYAY
201 GNVGPDALIE MDPEASDLSR GILALRDVAA EAGVSKYTGG RGVSVGPILS
251 SSASDIFCDN ENGPNFLFHN RGDGTFVDAA ASAGVDDPHQ HGRGVALADF
301 NRDGKVDIVY GNWNGPHRLY LQMSTHGKVR FRDIASPKFS MPSPVRTVIT
351 ADFDNDQELE IFFNNIAYRS SSANRLFRVI RREHGDPLIE ELNPGDALEP
401 EGRGTGGVVT DFDGDGMLDL ILSHGESMAQ PLSVFRGNQG FNNNWLRVVP
451 RTRFGAFARG AKVVLYTKKS GAHLRIIDGG SGYLCEMEPV AHFGLGKDEA
501 SSVEVTWPDG KMVSRNVASG EMNSVLEILY PRDEDTLQDP APLECGQGFS
551 QQENGHCMDT NECIQFPFVC PRDKPVCVNT YGSYRCRTNK KCSRGYEPNE
601 DGTACVERTL LLGLCNLLGK
```

B- dlCrtac1a

Start - End	Mr Obs. (Da)	Mr expt. (Da)	Mr calc. (Da)	ppm	Miss	Peptide sequence
78 - 87	1169.6958	1168.6885	1168.6676	18	1	K.RLVNIAVDNR.S
79 - 87	1013.5831	1012.5758	1012.5665	9	0	R.LVNIAVDNR.S
88 - 95	940.4978	939.4905	939.4814	10	0	R.SSPFYALR.D
115 - 129	1726.8467	1725.8394	1725.8322	4	0	R.EEIYVLNTNNAFSGR.A
115 - 135	2392.2529	2391.2456	2391.1342	47	1	R.EEIYVLNTNNAFSGRATYSDK.L
130 - 138	1072.5818	1071.5745	1071.5600	14	1	R.ATYSDKLFK.F
141 - 156	1956.9180	1955.9107	1955.9085	1	1	R.NGRFEDLLNDDINEHR.D
144 - 156	1629.7428	1628.7355	1628.7430	-5	0	R.FEDLLNDDINEHR.D
144 - 161	2185.0198	2184.0125	2184.0195	-3	1	R.FEDLLNDDINEHRDVANR.M
257 - 280	2560.0674	2559.0601	2559.1045	-17	0	R.NNGDGTFTDVAQQAGVEDPMQHGR.G
281 - 289	962.5063	961.4990	961.4981	1	0	R.GVALADFNR.D
281 - 292	1290.6930	1289.6857	1289.6476	30	1	R.GVALADFNRDGR.T
293 - 305	1528.7273	1527.7200	1527.7219	-1	0	R.TDIVYGNWNGPHR.L
325 - 332	920.4769	919.4696	919.4586	12	0	K.FSMPSPVR.T
325 - 332	936.4655	935.4582	935.4535	5	0	K.FSMPSPVR.T
368 - 389	2448.1040	2447.0967	2447.1313	-14	1	R.REHGDPQIEELNVGEASEPEGR.G
369 - 389	2292.0190	2291.0117	2291.0302	-8	0	R.EHGDPQIEELNVGEASEPEGR.G
390 - 403	1338.5929	1337.5856	1337.5848	1	0	R.GTGAVATDFDGDGR.L
390 - 422	3360.6174	3359.6101	3359.6634	-16	1	R.GTGAVATDFDGDGRLELLVSHGESAAQPLSVYK
438 - 445	897.5070	896.4997	896.4868	14	1	R.TKFGAFAR.G

Putative dlCrtac1a - Match to: gi|85838736 cartilage acidic protein 1 [Tetraodon nigroviridis]; Score: 154; Sequence Coverage: 29%; Nominal mass (M_r): 69652; Calculated pI value: 5.42

```
  1 MLAWVLLFLP PVSSAQRSEP VFSSITKSIL PPNYENNPTQ LNYGVAVTDV
 51 DGDGDLEVFV AGYNGPNLVL KYIQDQKRLV NIAVDNRSSP FYALRDRQGN
101 AIGVTACDID GDGREEIYVL NTNNAFSGRA TYSDKLFKFR NGRFEDLLND
151 DINEHRDVAN RMAGRSVACV DRKGTGRYAI YIANYASGNV GPHALIEMDE
201 LASDLSQGVI ALSNVAEEAG VNKFTGGRGV VVGPILNQIL PDVFCDNEYG
251 PNFLFRNNGD GTFTDVAQQA GVEDPMQHGR GVALADFNRD GRTDIVYGNW
301 NGPHRLFMQL NNRRQKFKDI ASQKFSMPSP VRTVIAADFD NDNELEVFFN
351 NIAYRGPSAN RLFRVSREH GDPQIEELNV GEASEPEGRG TGAVATDFDG
401 DGRLELLVSH GESAAQPLSV YKVLQGTSNS WLRVIPRTKF GAFARGAKVV
451 VYTKKSGTHT RIIDGGSGYL CEMEPVAHFG LGKDVATGVE VYWPDGRSVV
501 RLLEPSDLNS VLEIQYPRDV EVTPTAQTEC GHGFALNEKG RCTDEDECTR
551 YPFACPLDRP VCVNTYGSYR CRAKRRCNQG FEPSDDGSAC VGQVAYFGGT
601 RSFADRKWSG LCFWLLLPHV LRPGLL
```

C- dlCrtac1b

Start - End	Mr Obs. (Da)	Mr expt. (Da)	Mr calc. (Da)	ppm	Miss	Peptide sequence
83 - 99	1923.9750	1922.9677	1922.9738	-3	0	R.LVNIAIDDSNSPYYALR.D
83 - 101	2195.1028	2194.0955	2194.1018	-3	1	R.LVNIAIDDSNSPYYALRDR.A

102 - 133	3419.5569	3418.5496	3418.5484	0	1	R.AGNAIGVTACDVDGDGREEIYFLNTNNAYSGR.A
119 - 133	1790.8420	1789.8347	1789.8271	4	0	R.EEIYFLNTNNAYSGR.A
145 - 159	1776.8790	1775.8717	1775.8802	-5	1	R.NGRFEDLLSDELNVR.R
148 - 159	1449.7141	1448.7068	1448.7147	-5	0	R.FEDLLSDELNVR.R
182 - 211	3171.5063	3170.4990	3170.4866	4	0	R.YSVYVANYASGNVGPHALLEMDETASDVAK.G
212 - 232	2016.1317	2015.1244	2015.1164	4	1	K.GIIALSDVAAVAGVNKFTGGR.G
233 - 244	1211.6948	1210.6875	1210.7034	-13	0	R.GVVVGPILSQSR.S
245 - 260	1902.8413	1901.8340	1901.8254	5	0	R.SDVFCDNENGPNFLFK.N
245 - 272	3196.3804	3195.3731	3195.3662	2	1	R.SDVFCDNENGPNFLFKNNGDGTFVDMAR.Q
273 - 284	1415.6659	1414.6586	1414.6702	-8	1	R.QAGVEDRYQHGR.G
285 - 309	2687.2954	2686.2881	2686.2888	-0	1	R.GVALADFNGDGKTDIIYGNWNGPHR.L
297 - 309	1542.7275	1541.7202	1541.7375	-11	0	K.TDIIYGNWNGPHR.L
310 - 320	1270.6307	1269.6234	1269.6354	-9	0	R.LFLQGSDSTFR.N
321 - 334	1371.7269	1370.7196	1370.7306	-8	0	R.NIATGGFAAPSPIR.T
335 - 357	2656.3267	2655.3194	2655.3180	1	1	R.TVIAADLDNDKELDVFFNNIAYR.G
370 - 391	2335.1636	2334.1563	2334.1564	-0	1	R.RANADPLIQELNVGDAAEPEGR.G
371 - 391	2179.0588	2178.0515	2178.0553	-2	0	R.ANADPLIQELNVGDAAEPEGR.G
392 - 417	2602.2400	2601.2327	2601.2307	1	0	R.GTGGTVTDFDGDGQLDLLLAHGESAR.Q
392 - 424	3401.7026	3400.6953	3400.6900	2	1	R.GTGGTVTDFDGDGQLDLLLAHGESARQPISVFK.V
425 - 435	1261.6139	1260.6066	1260.6211	-11	0	K.VTQGSSNNWLR.V
440 - 447	913.4307	912.4234	912.4454	-24	0	R.TQFGSFAR.G
451 - 463	1362.6683	1361.6610	1361.6688	-6	0	K.VTAFTSQSGAHTR.I

Putative dlCrtac1b - Match to own sequence; Score: 248; Sequence Coverage: 60%; Nominal mass (M_r): 59216; Calculated pI value: 5.03

```
  1 MWASGLLLFL VGLWHQSRAQ GSEPMLRVVT ETMLPPDSLH NPTQLNYGMA
 51 VTDVDGDGDL EVVVAGYNGP NLVLKYDRTQ NRLVNIAIDD SNSPYYALRD
101 RAGNAIGVTA CDVDGDGREE IYFLNTNNAY SGRATYSDKL FKFRNGRFED
151 LLSDELNVRR GVANRMAGRS VACIDRKGTG RYSVYVANYA SGNVGPHALL
201 EMDETASDVA KGIIALSDVA AVAGVNKFTG GRGVVVGPIL SQSRSDVFCD
251 NENGPNFLFK NNGDGTFVDM ARQAGVEDRY QHGRGVALAD FNGDGKTDII
301 YGNWNGPHRL FLQGSDSTFR NIATGGFAAP SPIRTVIAAD LDNDKELDVF
351 FNNIAYRGNA PNRLFRVSRR ANADPLIQEL NVGDAAEPEG RGTGGTVTDF
401 DGDGQLDLLL AHGESARQPI SVFKVTQGSS NNWLRVIPRT QFGSFARGAK
451 VTAFTSQSGA HTRIIDGGSG YLCEMEPVAH FGLGNDEVTV LEVSWPDGSS
501 ITRTLQSGEM NSVVEVAYPK EGETFLLAND TQCGNGFTVK NGHCAGL
```

Table S4. Comparison of the amino acid sequence similarity of full length CRTAC1's (A) and the isolated beta-propeller region (bp) of CRTAC1's (bpCRTAC1) (B) between vertebrates, the piscine duplicates (Crtac1a and 1b), invertebrates and cyanobacteria. The sequence similarities are shown in percentages. In A, the sequences with the highest similarity with sea bass dlCrtac1a and 1b are highlighted in bold. In **B**, the highest similarity values between the isolated region bpCRTAC1 of the resolved three dimensional models [indicated in Fig. 1 for human, teleosts (*Dicentrarchus labrax* and *Sparus aurata*) and cyanobacter (Synechococcus sp.)] are shown in bold. Sequence similarity values were obtained through the multiple sequence alignment of the deduced amino acid sequences (carried out as described in section 2.2 of the manuscript), but in **B** considering only the bpCRTAC1 region for each protein that correspond to the hCRTAC1 homologue regions in between Tyr_{56} and Gln_{439}.

Abbreviations: Hs- *Homo sapiens,* Gg- *Gallus gallus.* Sa- *Sparus aurata*, Xt- *Xenopus tropicalis,* Lc- *Latimeria chalumnae,* Lo- *Lepisosteus oculatus*, Dl- *Dicentrarchus labrax*, Sa- *Sparus aurata,*

On- *Oreochromis niloticus,* Tn- *Tetraodon nigroviridis,* Dr- *Danio rerio*, Cm- *Callorhinchus milii,* Sp- *Strongylocentrotus purpuratu*s and Sy- *Synechococcus sp.*.

A

		CRTAC1					**Crtac1a**				**Crtac1b**					**CRTAC1**		
		Hs	Gg	Xt	Lc	Lo	**Dl**	On	Tn	Dr	**Dl**	Sa	On	Tn	Dr	Cm	Sp	Sy
CRTAC1	Hs	0	79	83	79	76	77	77	76	75	68	67	67	65	68	78	52	46
	Gg		0	82	79	76	75	75	73	76	65	65	65	63	65	76	54	45
	Xt			0	83	80	80	80	79	79	70	70	69	67	70	83	54	47
	Lc				0	82	81	81	79	80	70	70	70	67	69	84	55	49
	Lo					0	87	87	85	86	74	72	73	71	74	82	54	48
Crtac1a	**Dl**						0	**98**	94	89	72	72	72	70	73	80	54	48
	On							0	93	89	72	72	72	70	73	80	54	48
	Tn								0	87	73	72	72	71	73	78	55	49
	Dr									0	71	70	70	69	71	79	54	48
Crtac1b	**Dl**										0	**95**	**95**	91	87	70	61	56
	Sa											0	94	91	86	70	62	55
	On												0	88	86	70	61	56
	Tn													0	83	68	60	56
	Dr														0	69	62	56
CRTAC1	Cm															0	54	48
	Sp																0	51
	Sy																	0

B

		bpCRTAC1					**bpCrtac1a**				**bpCrtac1b**					**bpCRTAC1**		
		Hs	Gg	Xt	Lc	Lo	**Dl**	On	Tn	Dr	**Dl**	Sa	On	Tn	Dr	Cm	Sp	Sy
bpCRTAC1	Hs	0	90	92	90	80	**86**	87	86	86	**86**	**85**	86	85	85	76	76	**62**
	Gg		0	90	90	87	85	85	84	88	85	85	86	84	84	75	76	62
	Xt			0	90	89	87	86	87	88	88	86	88	85	86	77	78	62
	Lc				0	91	88	88	88	89	87	86	87	85	86	79	77	63
	Lo					0	94	94	93	93	90	89	91	88	89	78	77	62

bpCrtac1a	**Dl**	0	98	98	96	**89**	**88**	90	87	89	76	77	**63**
	On		0	97	96	89	88	90	87	89	76	76	63
	Tn			0	96	89	88	89	87	88	75	76	63
	Dr				0	88	87	88	86	87	75	76	64
bpCrtac1b	**Dl**					0	**96**	96	95	92	76	77	**62**
	Sa						0	96	93	91	75	78	**61**
	On							0	93	91	77	78	62
	Tn								0	88	74	76	63
	Dr									0	74	78	62
bpCRTAC1	Cm										0	67	53
	Sp											0	61
	Sy												0

1. Blom N, Sicheritz-Pontén T, Gupta R, Gammeltoft S, Brunak S: **Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence**. *Proreomics* 2004, **4**(6):1633-1649.
2. Steentoft C, Vakhrushev SY, Joshi HJ, Kong Y, Vester - Christensen MB, Schjoldager KTBG, Lavrsen K, Dabelsteen S, Pedersen NB, Marcos - Silva L *et al*: **Precision mapping of the human O - GalNAc glycoproteome through SimpleCell technology**. *EMBO J* 2013, **32**(10):1478-1488.
3. Miller ML, Soufi B, Jers C, Blom N, Macek B, Mijakovic I: **NetPhosBac – A predictor for Ser/Thr phosphorylation sites in bacterial proteins**. *Proteomics* 2009, **9**(1):116-125.
4. Maurer-Stroh S, Eisenhaber F: **Refinement and prediction of protein prenylation motifs**. *Genome Biol* 2005, **6**(6):1-15.
5. la Cour T, Kiemer L, Mølgaard A, Gupta R, Skriver K, Brunak S: **Analysis and prediction of leucine-rich nuclear export signals**. *Protein Eng Des Sel* 2004, **17**(6):527-536.

6. Pejaver V, Hsu W-L, Xin F, Dunker AK, Uversky VN, Radivojac P: **The structural and functional signatures of proteins that undergo multiple events of post-translational modification**. *Protein Sci* 2014, **23**(8):1077-1093.
7. Petersen TN, Brunak S, von Heijne G, Nielsen H: **SignalP 4.0: discriminating signal peptides from transmembrane regions**. *Nat Meth* 2011, **8**(10):785-786.
8. Gasteiger E, Hoogland C, Gattiker A, Duvaud Se, Wilkins M, Appel R, Bairoch A: **Protein identification and analysis tools in the ExPASy server**. In: *The Proteomics Protocols Handbook.* Edited by Walker J. Totowa, NJ: Humana Press; 2005: 571-607.
9. Linding R, Russell RB, Neduva V, Gibson TJ: **GlobPlot: exploring protein sequences for globularity and disorder**. *Nucleic Acids Res* 2003, **31**(13):3701-3708.
10. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice**. *Nucleic Acids Research* 1994, **22**(22):4673-4680.