Casein and casein micelle structures, functions and diversity in 20 species

Carl Holt

International Dairy Journal (2016)
DOI: 10.1016/j.idairyj.2016.01.004

Institute of Molecular, Cell and Systems Biology, School of Life Sciences, University of Glasgow, Glasgow G12 8QQ, UK.

Corresponding author: Carl Holt, RB413A Level B4, Joseph Black Building, University of Glasgow, Glasgow G12 8QQ, Tel. +441292560158. Email carl.holt@glasgow.ac.uk

ORCID ID 0000-0002-2087-1546

Primary structures of caseins from 20 species, including two monotremes and two marsupials, have been compared. Sequences of the mature proteins are very divergent whereas variation in amino acid composition is mostly restricted to a range of disorder-promoting residues. The number and size of clusters of phosphorylation sites in the caseins is variable, blurring the boundaries between them. Casein polar tract sequences were found in all caseins and are chiefly responsible for the weak and dynamic interactions among the tangled web of peptide chains in the matrix of casein micelles. The interactions take the predominant form of backbone-to-backbone contacts rather than sequence-specific side chain interactions of the hydrophobic effect. It is suggested that the dynamic casein micelle matrix be represented by an ensemble of interchanging structures with different types and degrees of inhomogeneity, influenced by solvent quality and other environmental factors.

1. Introduction

Caseins are pleiotropic proteins with an original and continuing function in biomineralisation and a better-known function as nutritional proteins. Recent research has revealed functions of the casein micelle relating to the needs of the mother rather than the neonate. In particular, the sequestration of calcium phosphate in a thermodynamically stable complex with caseins allows milk to be stored in the mammary gland without causing benign ectopic or pathological calcification in the cisternae or ducts of the gland (Holt & Carver, 2012; Holt, Carver, Ecroyd, & Thorn, 2013; Thorn, Ecroyd, Carver, & Holt, 2015). The same phenomenon is found in other biofluids including blood, extracellular fluid and saliva, but at very much lower concentrations of the sequestering protein, and is necessary to allow the easy coexistence of soft and mineralised tissues in the same organism (Holt, Sorensen, & Clegg, 2009; Holt, 2013; Holt, Lenton, Nylander, Sorensen, & Teixeira, 2014). For this purpose, the caseins are required to have clusters of phosphorylated residues and a disordered conformation but they must also have sticky sequences so that they can form clots or a gel in the stomach of the neonate. The sticky sequences in two of the individual bovine caseins contain certain smaller sub-sequences that are able to form amyloid fibrils at a rate that is facilitated by the disordered conformation of the whole protein (Leonil, et al., 2008; Thorn, et al., 2005; Thorn, Ecroyd, Sunde, Poon, & Carver, 2008; Treweek, Thorn, Price, & Carver, 2011). Fibril formation is, however, suppressed in milk by an alternative association pathway to produce the amorphous casein micelle. That casein genes affect three seemingly distinct phenotypic traits is not anomalous because the phenomenon of pleiotropism is common and forms a central concept in developmental biology (Hodgkin, 1998; Liberles, Tisdell, & Grahnen, 2011; Tokuriki, Stricher, Serrano, & Tawfik, 2008; Wang, Liao, & Zhang, 2010; Weinreich, Delaney, DePristo, & Hartl, 2006; Zeng & Gu, 2010). Indeed, the three functions of casein genes are essential to a successful reproductive strategy to such an extent that no single one should be described as the main function of caseins or the casein micelle. Notwithstanding this, there is an inevitable trade-off among the functions which affects the composition, structure and stability of the proteins.

Much of the scientific knowledge of caseins is derived from work done on farm animals, particularly the cow genus *Bos*. As a result, casein research has a boocentric bias. Capuco and Akers (Capuco & Akers, 2009) have pointed out that “Because no single species can provide an ample and sufficient model for the physiology of another, and because the potential gain in

knowledge from comparative studies is great, the research community should not be species-centric." Accordingly, an attempt is made here to extend earlier work (Ginger & Grigor, 1999; Martin, Cebo, & Miranda, 2013) by examining sequences from 20 species. This non-random sample includes a number of species closely related to the cow and is far from being representative of extant mammals but it does include two monotremes and three marsupials. This and further work may enable a more balanced, less-boocentric, perspective on the nature of caseins to be obtained.

2. Methods

2.1 Nomenclature

Casein gene and protein nomenclature is species-specific and potentially confusing so here the four bovine genes or protein names will be used wherever possible to identify the corresponding non-bovine orthologues. These are CSN1S1 encoding α_{S1}-, CSN1S2 encoding α_{S2}- , CSN2 encoding β- and CSN3 encoding κ-casein (Lefèvre, Sharp, & Nicholas, 2010; Rijnkels, Kooiman, de Boer, & Pieper, 1997)). A recent gene duplication in rodents and some other eutherians has generated the paralogous CSN1S2A and CSN1S2B-like genes. The CSN1S2A gene is orthologous to bovine CSN1S2. Previously the sequences coded by CSN1S2B-like genes were aligned with those of CSN1S2 orthologues (Kawasaki, Lafont, & Sire, 2011) but at the protein level the differences are quite marked, justifying them being regarded as a distinct casein. In the monotremes, the CSN2A gene is orthologous to the eutherian CSN2 but the CSN2B gene appears to be a chimera in which exons 2-6 can be aligned with exons from CSN2 and the following exons with CSN1S2 (Lefevre, Sharp, & Nicholas, 2009). Thus there are 6 casein genes known to science.

2.2. Sequences and sequence alignments

The sequences used include those from the work of Kawasaki *et al.* (Kawasaki, et al., 2011). Accession codes of additional sequences, mainly for the κ-casein alignments, are specified in Supplementary Data S1 and are for the default isoforms, including the signal sequence, in the UniprotKB database on the ExPASy Bioinformatics Resource Portal (http://www.expasy.org). Exon boundary conservation (Martin, et al., 2013) was used as a constraint throughout. Because of extensive cryptic duplications and divergence in the long exons encoding the casein polar tract sequences of β-casein, Kawasaki *et al.* (Kawasaki, et al., 2011) found that

no reliable multiple alignment was obtained across eutherian, marsupial, and monotreme sequences. In their work, separate alignments were made for each clade and were linked in their diagram without any further attempt at inter-clade alignment. A similar problem was encountered here in the alignment of 17 sequences encoded by the long exon 4 of κ-caseins. The problem was alleviated in both cases, although not eliminated, by using a unit scoring matrix to avoid giving undue weight to residues that are highly conserved in globular proteins but less-well conserved in disordered sequences (Brown, et al., 2002; Holt & Sawyer, 1993). Finally, alignments were edited manually to reduce the number of separate insertions or deletions. The resulting alignments of translated sequences, including signal sequences are shown in Supplementary Data S2-S6 for the 5 casein groups defined as (i) the α_{S1}-caseins (CSN1S1; Data S2), (ii) orthologues of the bovine α_{S2}-casein (CSN1S2; Data S3), (iii) miscellaneous other α_{S2}-type caseins (CSN1S2B-like + part monotreme CSN2B; Data S4), (iv) the β-caseins (CSN2 + part of monotreme CSN2B; Data S5) and (iv) the κ-caseins (CSN3; Data S6).

2.3. Casein polar tracts

Polar tracts are sequences found in intrinsically disordered proteins that are capable, under certain circumstances, of forming a more condensed structure by interacting with themselves or similar sequences. They are deficient in charged, hydrophobic, and Pro residues and enriched in polar amino acids such as Asn, Gly, Gln, His, Ser, and Thr (Das, Ruff, & Pappu, 2015). Polar tract-type sequences that are enriched in Pro tend to form extended, water-rich structures such as gels, viscous mucus and slimes rather than condensed structures (Kay, Williamson, & Sudol, 2000; Williamson, 1994). Casein polar tracts are largely hydrophilic sequences in which Pro and Gln are the most common residues (P,Q-rich) and they are encoded by long exons. The partial exception is the long exon 4 of CSN3 (κ-casein) which encodes two different types of polar tract: a P,Q,Y-rich sequence in the N-terminal half and a P,S,T-rich sequence in the C-terminal half. Separating the two is an aspartate proteinase-type cleavage site which generates the macropeptide in the stomach of some neonates. Here the term macropeptide is applied generically to the P,S,T-rich sequences of κ-caseins whether or not they are cleavage products of an aspartate proteinase

2.4. Clusters of phosphorylation sites

The sequences were inspected for canonical sites of phosphorylation as defined by the work of Jean-Claude Mercier and his colleagues (Martin, et al., 2013; Mercier, 1981; Mercier & Vilotte, 1993) and by the observed specificity of the Golgi casein kinase Fam20C (Tagliabracci, et al., 2012). The canonical kinase recognition sequence is -S-X-E/pS-, where X is any residue and pS is phosphoserine. Actual degrees of phosphorylation may be less than predicted due to incomplete phosphorylation of the available sites but may be larger because of phosphorylation at non-canonical sites such as -T-X-E/pS/D- or -S-X-D-. A cluster of *n* phosphorylation sites results from complete phosphorylation of sequences such as $-S-(S)_{n-1}-E-E$ or $-S-X-(S)_{n-1}-E-E$ where there is a single initial kinase recognition site. In caseins, the initial kinase recognition site of a cluster of sites is often formed by exon splicing (Mercier, 1981; Mercier & Vilotte, 1993). An exception is the potential phosphate cluster in platypus κ-casein which is found near the N-terminus of the casein polar tract sequence encoded by exon 4 (-S-S-E-E-S-E-E-). It is therefore not formed by exon splicing and contains two initial kinase recognition sites. A smaller cluster is found at the same position in the echidna κ-casein.

A peptide containing a phosphate cluster may be capable of sequestering amorphous calcium phosphate to form a thermodynamically stable complex. The requisite number and nature of phosphorylated residues in such a cluster is not well defined experimentally but all known examples have had three or more pS residues close together with a single initial kinase recognition site (Clegg & Holt, 2009; Holt, Wahlgren, & Drakenberg, 1996; Holt, Timmins, Errington, & Leaver, 1998; Holt, et al., 2009; Little & Holt, 2004).

2.5. Other sequence analysis tools

Amyloid-forming zipper sequences were predicted by the method of Goldschmidt *et al.* (Goldschmidt, Teng, Riek, & Eisenberg, 2010) and edited to remove predictions that conflicted with likely post translational modifications of phosphorylation and disulphide bond formation (Holt & Carver, 2012).

Amino acid composition and the mole fractions of positively or negatively charged residues were calculated using the ProtParam web tool (Gasteiger, et al., 2005) (http://web.expasy.org/protparam/). Hydropathy values were calculated from the

hydrophobicity scale of Kyte and Doolittle (Kyte & Doolittle, 1982) and renormalized to a scale from zero (Arg) to one (Ile) (Uversky, Gillespie, & Fink, 2000). The disorder propensity scale was taken from the TOP-IDP analysis (Campen, et al., 2008).

3. Results

3.1. Structure of casein genes

The phylogenetic studies of Kawasaki *et al.* (Kawasaki & Weiss, 2003; Kawasaki, et al., 2011) showed that all caseins are members of a group of secreted, calcium (phosphate)-binding proteins. They are formed from short and long exons separated by phase zero introns and the first and last exons of casein genes are totally untranslated. Exon 2 encodes a signal sequence which is necessary for secretion. All casein genes evolved from the bone and tooth gene ODAM which encodes the ODontogenic AMeloblast-associated protein but the calcium-sensitive casein genes took a slightly different subsequent path to the κ-casein gene. Thorn *et al.* (Thorn, et al., 2015) attempted to capture the essential differences between the two classes of casein genes by means of a Bauplan for each. The relationship of the Bauplan to some actual CSN2 gene structures is illustrated in Figure 1

3.2. Casein composition

The composition of whole casein is known to a reasonable degree of completeness in only a handful of eutherian species (Holt, et al., 2013; Martin, et al., 2013). Even within this small group, the proportions of the individual caseins are widely variable. For all the caseins except κ-casein, the proportion may be zero in certain individuals of a species, or in all members of a species or group of related species or in whole lineages. This is in spite of the evolution of the four orthologues CSN1S1, CSN1S2, CSN2 and CSN3 before the radiation of mammals (Lefèvre, et al., 2010; Oftedal, 2012). From the limited data to hand at the present time, it appears that casein micelles are formed from a mixture of between three and five types of casein, depending on the species. The nutritional function of the casein micelle does not provide a compelling explanation for why a mixture of casein types is preferred since the caseins have rather similar amino acid compositions and none is rich in essential amino acids (Hambraeus & Lonnerdal, 2003). By contrast, an explanation in terms of the suppression of amyloid formation by individual caseins (Holt & Carver, 2012; Holt, et al., 2013) has experimental support (Thorn, et al., 2005; Thorn, et al., 2008; Treweek, et al., 2011). The

175 experiments show that in the bovine caseins the formation of amyloid fibrils by any
176 individual casein is suppressed by all the other casein types in the mixture, each of which acts
177 as a molecular chaperone. Thus, a mixture of different types of casein is preferable to one,
178 amyloid-forming, casein. What is still not clear is what determines the number of components
179 in the mixture and their proportions. Also, why is it not possible to naturally select caseins
180 without amyloid-forming tendencies? The motivation for the present work is that studies of
181 non-bovine caseins from as wide a range of species as possible should help to solve these and
182 related problems.

183 The mean mole % and standard deviation of residues in the mature translated but
184 unphosphorylated sequences of the CSN1S1, CSN1S2, CSN2 and CSN3 orthologues are
185 shown in Figure 2(a)-(d). Excluded from this analysis are the eutherian CSN1S2B-like
186 sequences and the monotreme CSN2B sequences because there are few examples of these
187 from which to calculate reliable means with small standard deviations.

188 The most common residues are very largely drawn from the group of non-essential amino
189 acids with a high disorder propensity (Figure 2e) and low hydropathy (Figure 2f).
190 Hydrophobic residues, normally associated with the core of globular proteins, occur
191 relatively infrequently in all members of the sampled group. Only one essential, hydrophilic,
192 amino acid residue occurs at high frequency, and only in the α_{S2}-caseins and this is Lys,
193 which has a high disorder propensity. Only one hydrophobic essential amino acid with a low
194 disorder propensity, Leu, occurs at high (in the β-caseins) or moderate (in the α_{S1}-caseins)
195 frequency. The essential amino acids Thr and Val are found at high-to-moderate frequency in
196 κ-caseins and Val also occurs at moderate frequency in β-caseins.

197 There are relatively few residues that can be used to discriminate among the casein
198 orthologues. These can be identified in the multiple bar chart of mean compositions shown in
199 Figure 3a. Lys in α_{S2}-, Leu in β- and Thr and Ala in κ-casein occur more frequently than in
200 the other orthologues. Thr in α_{S1}- and Tyr in β-caseins occur less commonly, on average,
201 than in the other orthologues. Figure 3b shows a 3-D scatter plot of the individual species
202 using the orthogonal axes of the mole % Lys (x), mole % Leu (y) and mole % Thr (z).
203 Inspection of this projection and of others showed that the orthologous groups were
204 reasonably well separated using only these three dimensions.

At this point, any explanations for the compositional differences among the caseins are still highly speculative. The higher frequency of Thr residues in the κ-caseins may be due in part to a need for actual or potential sites for *O*-glycosylation in the macropeptide sequences. The Leu and Val residues in β-caseins occur in the casein polar tract sequences and may serve to supplement the backbone-to-backbone interactions that predominate in these regions with hydrophobic interactions of the side chains, either stabilising intermolecular binding or reinforcing a range of preferred intramolecular conformations. The higher frequency of Lys residues in the α_{S2}-casein A orthologues may be to reduce the net negative charge arising from phosphorylated residues but the positively charged residues are concentrated towards the C-terminus and hence that region could help to stabilise a range of dynamic conformations in which it interacts electrostatically with more central regions of net negative charge.

3.3. Multiple sequence alignment

The alignments shown in Supplementary Data S2-S6 are aligned in boxes corresponding to translated exon sequences such that all entries in the same box are considered to be orthologous (Kawasaki, et al., 2011). Fully conserved residues (identities) are those that appear in every species at a given position in the alignment. There were very few (63) positions of complete conservation in the 5 alignments and 24 of these were in the CSN1S2B-like group where diversity is low because there are only a few aligned sequences. Of the remaining 39 identities, most (22) were in sequences encoded by exon 2, 7 were in polar tracts and 10 were in sequences encoded by short exons, such as phosphate clusters (7). The signal sequences of the κ-caseins appear to be less well conserved than in other caseins. Residues such as Cys and Trp are normally better conserved in globular proteins than Ser or Glu but in the caseins the reverse is true, as was previously noted in an alignment of a smaller number of eutherian caseins (Holt & Sawyer, 1993). Cys residues in caseins differ from most Cys residues in globular proteins in that they are almost exclusively involved in intermolecular disulphide bridges but their role, if any, in the maintenance of casein micelle structure remains an enigma(Thorn, et al., 2015).

Fully conserved exons are those that are found in every species in an alignment. The numbers of fully conserved translated exons were 5 (13%), 9 (47%), 7 (41%), 2 (20%) and 3 (100%) in the alignments of Supplementary Data S2-S6, respectively. Here, the numbers in

parentheses are the fraction of fully conserved exons as a percentage of the total number of exons in the alignment. Not surprisingly, the number of conserved positions and exons is considerably higher when only eutherian species are considered. For example, among the eutherian β-caseins, 5 of the 6 translated exons are fully conserved compared to only two among the 9 translated exons of all the β-caseins and the β-casein-like part of monotreme CSN2B. The number of conserved residues and exons depend, to a degree, on the assumptions and methods used in the alignment and the particular sequences chosen. A more diverse and fully representative group of mammalian casein genes would almost certainly have produced even fewer examples of complete conservation but the chosen sample does demonstrate clearly that apart from the signal peptides of the calcium-sensitive caseins, casein sequences are very variable. This conclusion reinforces previously expressed views based on smaller groups of eutherian species (Martin, et al., 2013; Mercier & Vilotte, 1993) that the mature caseins show very low levels of sequence conservation.

3.4. Occurrence and variation of phosphorylation site clusters

The clustering of phosphorylation sites has been analysed by dividing the sequences into 4 groups. These groups are the sequences from Supplementary Data S2 (α_{S1}-caseins) , Supplementary Data S3 combined with Supplementary Data S4 (α_{S2}-type caseins), Supplementary Data S5 (β-caseins) and Supplementary Data S6 (κ-caseins). A histogram showing the frequency of occurrence of different cluster sizes in the individual casein groups and for all caseins is shown in Figures 4a and 4b, respectively.

Whereas single sites were the most common cluster size in κ-caseins (Fig. 4a) and the most common in the aggregate of all calcium-sensitive caseins (Figure 4b), the most common size in the individual calcium-sensitive caseins was 3 (α_{S2}-like), 4 (β-) or 5 (α_{S1}-) and the largest size was 7 in guinea pig, platypus and echidna α_{S1}-caseins. Figures 4c and 4d show the number of clusters of size $n \geq 2$ or $n \geq 3$ in individual sequences. The number varied from zero in most κ-caseins to 3 ($n \geq 3$) in some of the α_S-caseins. For a cluster size of 2 or greater the maximum number of clusters in a sequence was 5 in guinea pig α_{S2}-casein.

A minimum cluster size of $n = 3$ is required for their cross-linking action in casein micelles according to the findings of Aoki and co-workers (Aoki, Umeda, & Kako, 1992; Umeda &

Aoki, 2002). In the formation of calcium phosphate nanoclusters using pure phosphopeptides, thermodynamically stable complexes have only been observed if $n \geq 3$ (Holt, et al., 1996; Holt, et al., 1998; Little & Holt, 2004). Nevertheless, in mixtures of caseins or casein and osteopontin phosphopeptides, smaller clusters of $n = 2$ appear to enter the sequestering shell of peptides provided larger cluster sizes are also present (Clegg & Holt, 2009; Holt, et al., 2009). The theory of amorphous calcium phosphate sequestration by phosphopeptides has been developed using the term "phosphate centre", defined as a short sequence containing three or more phosphorylated residues. Ideally, this simplifying concept, although it has been very useful, will be replaced one day by a theory that accounts explicitly for the effects of the phosphate cluster size distribution. In eight eutherian species for which casein composition is well-enough established, the average number of phosphate centres per average mole of caseins has been calculated and shown to vary only within close bounds (Holt & Carver, 2012). If confirmed for a broader range of species it would indicate that the main factor in the interspecific variation of sequestered calcium and phosphate concentrations is the concentration of casein rather than its composition or the number and distribution of potential sites of phosphorylation.

When the data in Figure 4 are examined as a whole, it is clear that while there are differences in the average cluster size and number of clusters among the groups, the distributions overlap considerably, blurring the boundaries between caseins in their ability to sequester amorphous calcium phosphate.

3.5. Conservation of casein polar tracts

All the sequences in the sample contain at least one casein polar tract (coloured green in Supplementary Data S2-S6) and in the α_{S2}- casein orthologues and in the rabbit and rat α_{S2}-casein B sequences there are two because of an intragenic duplication (Stewart, et al., 1987). This level of conservation is second only to the conservation of the signal peptide sequences in caseins. The variation in the lengths of the polar tracts has been investigated by dividing the sequences into the same 4 groups used in the phosphorylation site cluster analysis. Among the four groups there are clear differences in the average lengths of the casein polar tract sequences with only limited overlaps in the number distribution histograms (Figure 5). For example, the average number of residues in polar tract sequences in the 4 groups are 57.75±12.76 in α_{S1}-caseins), 40.93±2.30 and 40.79±2.08 in the α_{S2}-like caseins, 88.68±5.20

in the κ-caeins and 170.43±28.62 in the β-caseins. Thus, the β-casein orthologues contain up to approximately twice as many residues in polar tract sequences as in all other casein types.

3.6. Occurrence of amyloid zipper sequences in caseins

There is a limited amount of evidence that amyloid protofibrils are part of the structure of bovine casein micelles (Lencki, 2007). Casein sequences were therefore examined to see whether predicted amyloid zipper sequences are conserved, as they would be if the protofibrils are important for casein micelle formation. Amyloid zipper predictions are abundant in the casein polar tract sequences (Holt & Carver, 2012) but because of the uncertain alignment of the casein polar tracts in the β-caseins, the sequences of choice to test the hypothesis are in the shorter polar tracts of κ- and α_S- caseins. Amyloid zipper predictions have been validated in experimental studies with other proteins (Goldschmidt, et al., 2010; Nelson, et al., 2005; Rodriguez, et al., 2015; Sawaya, et al., 2007; Thompson, et al., 2006; Wiltzius, et al., 2008). They were found in the regions of bovine α_{S2}- and κ-caseins where amyloid formation has been demonstrated experimentally (Ecroyd, et al., 2008; Ecroyd, Thorn, Liu, & Carver, 2010; H. M. Farrell, Cooke, Wickham, Piotrowski, & Hoagland, 2003; Niewold, Murphy, Hulskamp-Koch, Tooten, & Gruys, 1999). Nevertheless, neither these nor any other predicted amyloid zipper sequences were fully conserved in the aligned casein sequences. This is illustrated in Figure 6 for the κ-caseins where the predicted amyloid zipper sequences are highlighted blue in the multiple sequence alignment.

The complete absence of experimental evidence of amyloid fibril formation in non-bovine caseins is not evidence of absence. It is important to gather more data to help explain why potentially cytotoxic, amyloid-forming sequences, are tolerated in a vital food.

3.7. Condensation of casein polar tracts and casein micelle structure

Casein micelles have been found in all milks so far examined but a striking conclusion from interspecific studies (Martin, et al., 2013) is that they can be made in a large number of distinct ways using a mixture of different caseins in variable proportions.

The quality of the solvent and its structure in the solvation sheath around the backbone are factors that are thought to be important in the condensation of polar tracts (Das, et al., 2015; R. V. Pappu, Srinivasan, & Rose, 2000; Rohit V. Pappu, Wang, Vitalis, & Crick, 2008; Tran,

Mao, & Pappu, 2008) and for the conformational preferences of intrinsically disordered proteins, especially for transitions within the conformational space enclosing the poly-*L*-proline helix and the more extended β-strand structure (Ilawe, Raeber, Schweitzer-Stenner, Toal, & Wong, 2015; Meral, Toal, Schweitzer-Stenner, & Urbanc, 2015). Thus, changes in water structure are important in both polar tract interactions and the hydrophobic effect between side chains.

The arguments for using the term "casein polar tract" rather than "hydrophobic tail" to describe the P,Q-rich casein sequences were set out previously (Holt, et al., 2013). The finding that caseins act on each other to control amyloid fibril formation and micelle size and act on a broad range of denatured globular proteins to limit their aggregation is one of the strongest arguments that the interactions have a low sequence specificity and are therefore mainly backbone-to-backbone.

Casein polar tracts contain many Pro residues which favour the PP-II conformation, prevent β-strands from forming and inhibit amyloid structures. In salivary Pro-rich proteins, the side chain of Pro can stack with polyphenolic molecules such as tannins to form an insoluble complex (Bennick, 2002; Charlton, Haslam, & Williamson, 2002; Luck, et al., 1994; Murray, Williamson, Lilley, & Haslam, 1994; Pascal, Pate, Cheynier, & Delsuc, 2009; Williamson, 1994). Stacking (also called π–π stacking) refers to attractive, noncovalent interactions between aromatic rings, since they contain π bonding electrons. Despite intense experimental and theoretical interest, there is no unified description of the factors that contribute to stacking interactions. The interaction of the major polycyclic flavan-3-ol from green tea with the salivary proline-rich protein IB5 appears to favour the PP-II conformation (Pascal, et al., 2009). In studies by NMR employing time-averaged nuclear Overhauser measurements of a model peptide Q-G-R-P-P-Q-G with the polyphenol (-)-epigallocatechin gallate (Charlton, et al., 2002), a range of possible conformations was generated in which there were stacking interactions of the Pro residues with the A and C rings of the polyphenol. The structures appear to show how a precipitate can grow from the complex by further stacking interactions. In the interaction of casein micelles with the same polyphenol, it was found that up to a million molecules of the tannin could be incorporated in the micelles but without forming a precipitate (Shukla, Narayanan, & Zanchi, 2009).

Although Pro residues exhibit the stacking interaction with polycyclic phenols it is unclear whether the interaction has a significant contribution from the hydrophobic effect. Pro residues are not classified as hydrophobic because they hold tightly onto their solvating water molecules, partly because the backbone carbonyl is a good hydrogen bond acceptor (Theillet, et al., 2013). As a result, the presence of Pro residues in polar tracts restricts the condensation process so that they form water-rich structures such as gels, mucus and slimes (Williamson, 1994). The water-rich matrix of casein micelles may not be fully homogeneous. Between most of the Pro residues in caseins are short, conventional polar tract sequences, on average comprising 5-6 residues , that could interact with other, similar, sequences and form more compact or condensed substructures within the matrix. To explain the amplitude of a feature in the small-angle scattering of casein micelles, de Kruif et al. (de Kruif, Huppertz, Urban, & Petukhov, 2012) proposed that condensed protein structures on a scale of 2 nm size are present in the matrix as a result of hydrophobic interactions. Hydrophobic interactions are not needed for the explanation, however. In the interaction of casein polar tracts through backbone-to-backbone interactions there is nevertheless a dependence on solvent quality through the amino acid composition (not sequence) of the tract. Under any given condition of solvent quality, fluctuations about the mean density of the matrix in casein micelles could occur through alternative and nearly equivalent polar tract interactions. Other studies have provided evidence of voids or channels in the protein matrix and of distortions of micelle shape at a surface or in ice (Bouchoux, Gésan-Guiziou, Pérez, & Cabane, 2010; Bouchoux, et al., 2015; Dalgleish, Spagnuolo, & Goff, 2004; Gebhardt & Kulozik, 2014; Ouanezar, Guyomarc'h, & Bouchoux, 2012; Trejo, Dokland, Jurat-Fuentes, & Harte, 2011). Many older electron microscopy techniques produced images of larger-scale substructures called submicelles and a debate continues on the extent to which these submicelles are artefacts of the sample preparation methods (Farrell, Malin, Brown, & Qi, 2006; McMahon & McManus, 1998; McMahon & Oommen, 2012). Drying changes the small-angle X-ray scattering of casein micelles, including effects on the length scale attributed to submicelles or the average distance between nanoclusters (Mata, Udabage, & Gilbert, 2011).

Irrespective of whether any particular structure is an artefact, as some undoubtedly are, a generalisation can be made from these various observations which is that the structure of the casein micelle is both fragile and dynamic and its internal structure can be perturbed in a

variety of different ways so that internal fluctuations in matrix density become more or less important.

An intrinsically disordered protein can be described by its average size and shape. However, such proteins are dynamic and explore a huge number of alternative conformations, some of which are preferred but most are transient. The description of all possible conformations is impracticable but an improvement over the description of the average size and shape is the use of an ensemble of structures derived from scattering and spectroscopic measurements (Bernado, Bertoncini, Griesinger, Zweckstetter, & Blackledge, 2005; Bernado, Blanchard, et al., 2005; Jensen, Salmon, Nodet, & Blackledge, 2010). The ensemble describes both the average structure and of excursions from the average structure experienced by a single molecule over time or, equivalently, by a population of molecules at an instant of time. The ensemble is a collection of static structures that conveys the dynamic or mutable nature of the molecule they represent. The weighting given to a particular member of the ensemble can be considered to be the size of the subpopulation having that conformation. Thus, the effect of a change of environment such as solvent quality can be represented by shifts in the population among ensemble members.

The dynamic nature of casein micelles means that at any one moment a micelle can have a substructure which departs in some way from the average. The ensemble hypothesis for describing native casein micelles is that the total range of substructures can be represented by a much smaller ensemble of substructures such that the average substructure and the distribution of alternative substructures is obtained as a weighted sum over all members of the ensemble. In addition to a distribution of substructures, casein micelles exist as a dynamic distribution of sizes, also responsive to environmental change, although the rate of change may be slow (Dewan, Chudgar, Mead, Bloomfield.V, & Morr, 1974; Huppertz, Kelly, & de Kruif, 2006; Jackson & McGillivray, 2011; Lin, Dewan, Bloomfield, & Morr, 1971)

Four members are suggested as an ensemble of alternative casein micelle substructures (Figure 7a). These are all nanocluster models having various types or degrees of disproportionation of the protein matrix density. These are (7b) a more-or-less homogeneous matrix, (7c) a matrix with void spaces, (7d) a matrix with condensed protein structures and

(7e) a matrix with both void spaces and condensed protein structures. This basis set can be modified or added to in the future as the need arises.

Experiments on native casein micelles diluted with their own ultrafiltrate and studied by small-angle X-ray or neutron scattering suggest that the protein matrix is relatively, though not completely, homogeneous (de Kruif, et al., 2012; de Kruif, 2014; Holt, de Kruif, Tuinier, & Timmins, 2003; Ingham, et al., 2015; Marchin, Putaux, Pignon, & Léonil, 2007; Shukla, et al., 2009). Nevertheless, it very readily disproportionates to form less homogeneous structures under a wide range of conditions including heat treatment, drying, freezing, the reduction of water activity, adsorption at a surface, filtration forces, addition of ethanol or other poor solvent and addition of polyphenols. The fact that disproportionated matrices are readily observed is an indication that such structures are already present in an ensemble of native states but perhaps are not highly populated.

3.8. Pleiotropy of caseins

The first casein evolved in some stem amniote, before the great divergence into the sauropsid and synapsid lineages, about 50 million years before the emergence of lactation in the mammal-like reptiles (Kawasaki, et al., 2011). Caseins are therefore pleiotropic proteins in which an antecedent function, probably in the control of some aspect of biomineralisation, is closely related to one of the current functions of sequestering amorphous calcium phosphate in the form of thermodynamically stable nanoclusters, but unrelated to an additional current function, namely, neonatal nutrition. The additional function has required a great increase in expression level and introduced conflicting pressures on the composition and structure of the caseins. In biofluids other than milk, such as blood, extracellular fluid, saliva and urine, the sequestering function of phosphoproteins is exercised at concentrations typically three orders of magnitude smaller than in milk (Holt, et al., 2014). Such an increase in concentration has brought with it problems that were of minor or of no importance in the antecedent function of casein.

An unfolded, intrinsically disordered, conformation appears to be a characteristic functional feature of many of the phosphoproteins involved in the control of biomineralisation (Holt, et al., 2009; Kalmar, Homola, Varga, & Tompa, 2012) and was recognised to be advantageous for the sequestration of amorphous calcium phosphate by caseins and other phosphoproteins

such as osteopontin (Holt & Sawyer, 1993; Holt, 2013). However the subset of amino acids that favour disordered conformations do not overlap well with the subset of essential amino acids, as demonstrated in Figure 2e. Indeed, the amino acid composition of caseins, with few exceptions, is commensurate with the sequestering function and the delivery of high concentrations of calcium and phosphate for bone growth but it does not deliver high concentrations of the essential amino acids to the neonate.

The effect of pleiotropy on the evolution of proteins has received a great deal of theoretical and experimental attention (Delgado, et al., 2001; Hodgkin, 1998; Liberles, et al., 2011; Wang, et al., 2010). Because most mutational changes to a globular protein cause it to lose stability, the translational selection hypothesis is that protein evolution rate is controlled by protein stability and the need to avoid the formation of cytotoxic misfolding products such as amyloid (Zeng & Gu, 2010). For example, in the adaptation of a bacterial β-lactamase to a new antibiotic, a total of 5 residue substitutions produced a substantial increase in resistance to the drug but the experimental generation of the 5! = 120 trajectories produced mostly destabilised, amyloid-forming intermediates (Weinreich, et al., 2006). In a recent impressive study of the bacterial signalling kinase, PhoQ, a total of 160,000 mutations were generated at 5 residue positions affecting the binding to its substrate, PhoP (Podgornaia & Laub, 2015). Of these 1659 were functional in that they would still recognise PhoP. Podgornaia and Laub (2015) suggest that not all the functional variants are found in nature because of further context-sensitive constraints. Thus, trajectories that somehow sustain neutral or maladaptive intermediates on the way to an adaptive new function must exist, even for globular proteins.

The problem of maladaptive intermediates is no less severe for pleiotropic intrinsically disordered proteins than for globular proteins because the generation of amyloid-forming or other types of cytotoxic sequences cannot be hidden in a hydrophobic core. The problem of amyloid fibril formation by caseins is made more severe by the pressure to increase the protein concentration to fulfil the nutritional function. In considering the evolution of caseins and milk, Holt and Carver (2012) proposed that pleiotropy favoured an increase in casein gene complexity through an epistatic mechanism. The increased complexity at the casein locus took the form of an increased number of similar casein genes. Thus, when a number of casein gene products are expressed together, each has been shown to act as a molecular chaperone to inhibit the formation of amyloid fibrils by all of the other caseins. The

alternative pathway of aggregation produces an amorphous casein micelle rather than highly ordered amyloid fibrils. In consequence, the total casein concentration could be increased safely from μM to mM concentrations (Holt, et al., 2013) without causing amyloid to form. In other words, the casein micelle has sustained a trajectory of evolutionary change producing increased levels of expression and mutational changes in caseins that would otherwise be maladaptive.

Among existing orthologues studied here there is none that has an amino acid composition rich in all the essential amino acids, but as the studies of Podgornaia and Laub (2015) suggest, the orthologues of caseins that exist today and all that have ever existed may still only be a fraction of the total number of functional sequences. Among the latter may be sequences that are better suited, in terms of essential amino acid composition, to the current nutritional role of caseins.

Negative pleiotropy has been invoked in relation to the adaptation of signalling proteins and domains such as CH2, SH3 and WW (Liberles, et al., 2011; Zarrinpar, Park, & Lim, 2003). Such domains can respond in a signalling cascade to a range of ligands (Uversky, Oldfield, & Dunker, 2005) but in a positive pleiotropic adaption the danger is that they will also acquire the ability to bind a new ligand producing a potentially cytotoxic response. Negative pleiotropy is a process in which undesirable adaptations are eliminated while new functions are acquired. The ability of signalling domains to respond to different ligands is sometimes called promiscuous ligand binding. Promiscuous binding of target proteins is an essential attribute of molecular chaperones in preventing misfolded proteins from aggregating or forming amyloid fibrils (Barral, Broadley, Schaffar, & Hartl, 2004; Ecroyd & Carver, 2009; Westerheide, Raynes, Powell, Xue, & Uversky, 2012). Promiscuous interactions have been considered important in the pleiotropy of viral proteins (Habchi & Longhi, 2012). The interaction of co-secreted caseins leading to the formation of the casein micelle also reduces the possibility of very high concentrations of caseins binding promiscuously to unintended targets in the intracellular secretion pathway, the cisterns and ducts of the mammary gland and the stomach of the neonate.

In summary, pleiotropy provides us with a potential explanation for the manifest diversity in casein sequences and an increase over evolutionary time in the complexity of the casein gene

locus. Co-secretion of a mixture of caseins to produce the casein micelle reduced their individual ability to interact promiscuously with non-casein proteins and has neutralised what would otherwise be maladaptive mutations on a trajectory towards the newly acquired nutritional function.

4. Conclusions

Caseins are pleiotropic proteins in which the antecedent function in the control of some aspect of biomineralisation is related to their current function in neonatal nutrition where they sequester amorphous calcium phosphate. Potentially pathological consequences of a very large increase in expression level have been neutralised by increasing the number of co-expressed casein genes so that they bind to each other and form the casein micelle. The content of essential amino acids has remained at a low level, probably because higher levels would be incompatible with the unfolded conformation needed for caseins to form a thermodynamically stable complex with amorphous calcium phosphate. The casein micelle is a fragile and dynamic structure which can therefore be represented better by an ensemble of interconverting states than by an average state.

5. Acknowledgements

John Carver and David Thorn of the Australian National University , Heath Ecroyd of the University of Wollongong and Jared Raynes of CSIRO, Werribee, Australia all played important roles in the development of the ideas in this paper.

6. References

References

Aoki, T., Umeda, T., & Kako, Y. (1992). The least number of phosphate groups for cross-linking of casein by colloidal calcium phosphate. *Journal of Dairy Science, 75*, 971-975.

Barral, J. M., Broadley, S. A., Schaffar, G., & Hartl, F. U. (2004). Roles of molecular chaperones in protein misfolding diseases. *Seminars in Cell & Developmental Biology, 15*, 17-29.

Bennick, A. (2002). Interaction of plant polyphenols with salivary proteins. *Critical Reviews in Oral Biology & Medicine, 13*, 184-196.

Bernado, P., Bertoncini, C. W., Griesinger, C., Zweckstetter, M., & Blackledge, M. (2005). Defining long-range order and local disorder in native alpha-synuclein using residual dipolar couplings. *Journal of the American Chemical Society, 127*, 17968-17969.

Bernado, P., Blanchard, L., Timmins, P., Marion, D., Ruigrok, R. W. H., & Blackledge, M. (2005). A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering. *Proceedings of the National Academy of Sciences of the United States of America, 102*, 17002-17007.

Bouchoux, A., Gésan-Guiziou, G., Pérez, J., & Cabane, B. (2010). How to Squeeze a Sponge: Casein Micelles under Osmotic Stress, a SAXS Study. *Biophysical Journal, 99*, 3754-3762.

Bouchoux, A., Ventureira, J., Gesan-Guiziou, G., Garnier-Lambrouin, F., Qu, P., Pasquier, C., Pezennec, S., Schweins, R., & Cabane, B. (2015). Structural heterogeneity of milk casein micelles: A SANS contrast variation study. *Soft Matter, 11*, 389-399.

Brown, C. J., Takayama, S., Campen, A. M., Vise, P., Marshall, T. W., Oldfield, C. J., Williams, C. J., & Dunker, A. K. (2002). Evolutionary rate heterogeneity in proteins with long disordered regions. *Journal of Molecular Evolution, 55*, 104-110.

Campen, A., Williams, R. M., Brown, C. J., Meng, J. W., Uversky, V. N., & Dunker, A. K. (2008). TOP-IDP-scale: A new amino acid scale measuring propensity for intrinsic disorder. *Protein and Peptide Letters, 15*, 956-963.

Capuco, A. V., & Akers, R. M. (2009). The origin and evolution of lactation. *Journal of Biology (London), 8*, Article No.: 37.

Charlton, A. J., Haslam, E., & Williamson, M. P. (2002). Multiple conformations of the proline-rich protein/epigallocatechin gallate complex determined by time-averaged nuclear Overhauser effects. *Journal of the American Chemical Society, 124*, 9899-9905.

Clegg, R. A., & Holt, C. (2009). An E. coli over-expression system for multiply-phosphorylated proteins and its use in a study of calcium phosphate sequestration by novel recombinant phosphopeptides. *Protein Expression and Purification, 67*, 23-34.

Dalgleish, D. G., Spagnuolo, P. A., & Goff, H. D. (2004). A possible structure of the casein micelle based on high-resolution field-emission scanning electron microscopy. *International Dairy Journal, 14*, 1025-1031.

Das, R. K., Ruff, K. M., & Pappu, R. V. (2015). Relating sequence encoded information to form and function of intrinsically disordered proteins. *Current Opinion in Structural Biology, 32*, 102-112.

de Kruif, C. G., Huppertz, T., Urban, V. S., & Petukhov, A. V. (2012). Casein micelles and their internal structure. *Advances in Colloid and Interface Science, 171*, 36-52.

de Kruif, C. G. (2014). The structure of casein micelles: a review of small-angle scattering data. *Journal of Applied Crystallography, 47*, 1479-1489.

Delgado, S., Casane, D., Bonnaud, L., Laurin, M., Sire, J.-Y., & Girondot, M. (2001). Molecular Evidence for Precambrian Origin of Amelogenin, the Major Protein of Vertebrate Enamel. *Molecular Biology and Evolution, 18*, 2146-2153.

Dewan, R. K., Chudgar, A., Mead, R., Bloomfield.V, A., & Morr, C. V. (1974). Molecular weight and size distribution of bovine milk casein micelles. *Biochimica Et Biophysica Acta, 342*, 313-321.

Ecroyd, H., Koudelka, T., Thorn, D. C., Williams, D. M., Devlin, G., Hoffmann, P., & Carver, J. A. (2008). Dissociation from the oligomeric state is the rate-limiting step in fibril formation by kappa-casein. *Journal of Biological Chemistry, 283*, 9012-9022.

Ecroyd, H., & Carver, J. A. (2009). Crystallin proteins and amyloid fibrils. *Cellular and Molecular Life Sciences, 66*, 62-81.

Ecroyd, H., Thorn, D. C., Liu, Y., & Carver, J. A. (2010). The dissociated form of κ-casein is the precursor to its amyloid fibril formation. *Biochemical Journal, 429*, 251-260.

Farrell, Malin, E. L., Brown, C. J., & Qi, P. X. (2006). Casein micelle structure: What can be learned from milk synthesis and structural biology? *Current Opinion in Colloid & Interface Science, 11*, 135-147.

Farrell, H. M., Cooke, P. H., Wickham, E. D., Piotrowski, E. G., & Hoagland, P. D. (2003). Environmental Influences on Bovine κ-Casein: Reduction and Conversion to Fibrillar (Amyloid) Structures. *Journal of Protein Chemistry, 22*, 259-273.

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D., & Bairoch, A. (2005). Protein Identification and Analysis Tools on the ExPASy Server. In J. M. Walker (Ed.), *The Proteomics Protocols Handbook* (pp. 571-607). New York, NY, USA: Humana.

Gebhardt, R., & Kulozik, U. (2014). Simulation of the shape and size of casein micelles in a film state. *Food & Function, 5*, 780-785.

Ginger, M. R., & Grigor, M. R. (1999). Comparative aspects of milk caseins. *Comparative Biochemistry and Physiology B-Biochemistry & Molecular Biology, 124*, 133-145.

Goldschmidt, L., Teng, P. K., Riek, R., & Eisenberg, D. (2010). Identifying the amylome, proteins capable of forming amyloid-like fibrils. *Proceedings of the National Academy of Sciences, 107*, 3487-3492.

Habchi, J., & Longhi, S. (2012). Structural disorder within paramyxovirus nucleoproteins and phosphoproteins. *Molecular Biosystems, 8*, 69-81.

Hambraeus, L., & Lonnerdal, B. (2003). Nutritional aspects of milk proteins. In P. F. Fox & P. L. H. McSweeney (Eds.), *Advanced Dairy Chemistry* (Vol. 1 Part B, pp. 605-645). New York, NY, USA: Kluwer Academic/Plenum.

Hodgkin, J. (1998). Seven types of pleiotropy. *International Journal of Developmental Biology, 42*, 501-505.

Holt, C., & Sawyer, L. (1993). Caseins as Rheomorphic Proteins - Interpretation of Primary and Secondary Structures of the Alpha-S1-Caseins, Beta-Caseins and Kappa-Caseins. *Journal of the Chemical Society-Faraday Transactions, 89*, 2683-2692.

Holt, C., Wahlgren, N. M., & Drakenberg, T. (1996). Ability of a beta-casein phosphopeptide to modulate the precipitation of calcium phosphate by forming amorphous dicalcium phosphate nanoclusters. *Biochemical Journal, 314*, 1035-1039.

Holt, C., Timmins, P. A., Errington, N., & Leaver, J. (1998). A core-shell model of calcium phosphate nanoclusters stabilized by beta-casein phosphopeptides, derived from sedimentation equilibrium and small-angle X-ray and neutron-scattering measurements. *European Journal of Biochemistry, 252*, 73-78.

Holt, C., de Kruif, C. G., Tuinier, R., & Timmins, P. A. (2003). Substructure of bovine casein micelles by small-angle X-ray and neutron scattering. *Colloids and Surfaces a-Physicochemical and Engineering Aspects, 213*, 275-284.

Holt, C., Sorensen, E. S., & Clegg, R. A. (2009). Role of calcium phosphate nanoclusters in the control of calcification. *Febs Journal, 276*, 2308-2323.

Holt, C., & Carver, J. A. (2012). Darwinian transformation of a 'scarcely nutritious fluid' into milk. *Journal of Evolutionary Biology, 25*, 1253-1263.

Holt, C. (2013). Unfolded phosphoproteins enable soft and hard tissues to coexist in the same organism with relative ease. *Current Opinion in Structural Biology, 23*, 420-425.

Holt, C., Carver, J. A., Ecroyd, H., & Thorn, D. C. (2013). Caseins and the casein micelle: their biological functions, structures and behaviour in foods. *Journal of Dairy Science, 96*, 6127–6146.

Holt, C., Lenton, S., Nylander, T., Sorensen, E. S., & Teixeira, S. C. M. (2014). Mineralisation of soft and hard tissues and the stability of biofluids. *Journal of Structural Biology, 185*, 383-396.

Huppertz, T., Kelly, A. L., & de Kruif, C. G. (2006). Disruption and reassociation of casein micelles under high pressure. *Journal of Dairy Research, 73*, 294-298.

Ilawe, N. V., Raeber, A. E., Schweitzer-Stenner, R., Toal, S. E., & Wong, B. M. (2015). Assessing backbone solvation effects in the conformational propensities of amino acid residues in unfolded peptides. *Physical Chemistry Chemical Physics, 17*, 24917-24924.

Ingham, B., Erlangga, G. D., Smialowska, A., Kirby, N. M., Wang, C., Matia-Merino, L., Haverkamp, R. G., & Carr, A. J. (2015). Solving the mystery of the internal structure of casein micelles. *Soft Matter, 11*, 2723-2725.

Jackson, A. J., & McGillivray, D. J. (2011). Protein aggregate structure under high pressure. *Chemical Communications, 47*, 487-489.

Jensen, M. R., Salmon, L., Nodet, G., & Blackledge, M. (2010). Defining Conformational Ensembles of Intrinsically Disordered and Partially Folded Proteins Directly from Chemical Shifts. *Journal of the American Chemical Society, 132*, 1270-1272.

Kalmar, L., Homola, D., Varga, G., & Tompa, P. (2012). Structural disorder in proteins brings order to crystal growth in biomineralization. *Bone, 51*, 528-534.

Kawasaki, K., & Weiss, K. M. (2003). Mineralized tissue and vertebrate evolution: The secretory calcium-binding phosphoprotein gene cluster. *Proceedings of the National Academy of Sciences of the United States of America, 100*, 4060-4065.

Kawasaki, K., Lafont, A.-G., & Sire, J.-Y. (2011). The evolution of casein genes from tooth genes before the origin of mammals. *Molecular Biology and Evolution, 28*, 2053-2061.

Kay, B. K., Williamson, M. P., & Sudol, P. (2000). The importance of being proline: the interaction of proline-rich motifs in signaling proteins with their cognate domains. *Faseb Journal, 14*, 231-241.

Kyte, J., & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology, 157*, 105-132.

Lefevre, C. M., Sharp, J. A., & Nicholas, K. R. (2009). Characterisation of monotreme caseins reveals lineage-specific expansion of an ancestral casein locus in mammals. *Reproduction Fertility and Development, 21*, 1015-1027.

Lefèvre, C. M., Sharp, J. A., & Nicholas, K. R. (2010). Evolution of lactation: ancient origin and extreme adaptations of the lactation system. *Annual Review of Genomics and Human Genetics, 11*, 219-238.

Lencki, R. W. (2007). Evidence for fibril-like structure in bovine casein micelles. *Journal of Dairy Science, 90*, 75-89.

Leonil, J., Henry, G., Jouanneau, D., Delage, M.-M., Forge, V., & Putaux, J.-L. (2008). Kinetics of Fibril Formation of Bovine κ-Casein Indicate a Conformational Rearrangement as a Critical Step in the Process. *Journal of Molecular Biology, 381*, 1267-1280.

Liberles, D. A., Tisdell, M. D. M., & Grahnen, J. A. (2011). Binding constraints on the evolution of enzymes and signalling proteins: the important role of negative pleiotropy. *Proceedings of the Royal Society B-Biological Sciences, 278*, 1930-1935.

Lin, S. H. C., Dewan, R. K., Bloomfield, V. A., & Morr, C. V. (1971). Inelastic light-scattering study of the size distribution of bovine milk casein micelles. *Biochemistry, 10*, 4788-4793.

Little, E. M., & Holt, C. (2004). An equilibrium thermodynamic model of the sequestration of calcium phosphate by casein phosphopeptides. *European Biophysics Journal with Biophysics Letters, 33*, 435-447.

Luck, G., Liao, H., Murray, N. J., Grimmer, H. R., Warminski, E. E., Williamson, M. P., Lilley, T. H., & Haslam, E. (1994). Polyphenols, astringency and proline-rich proteins. *Phytochemistry, 37*, 357-371.

Marchin, S., Putaux, J.-L., Pignon, F., & Léonil, J. (2007). Effects of the environmental factors on the casein micelle structure studied by cryo transmission electron microscopy and small-angle x-ray scattering/ultrasmall-angle x-ray scattering. *J Chem Phys, 126*, 045101.

Martin, P., Cebo, C., & Miranda, G. (2013). Interspecies comparison of milk proteins: quantitative variability and molecular diversity. In P. L. H. McSweeney & P. F. Fox (Eds.), *Advanced Dairy Chemistry* (4th ed., Vol. 1A. Proteins: Basic Aspects, pp. 386-429). New York, NY, USA: Springer.

Mata, J. P., Udabage, P., & Gilbert, E. P. (2011). Structure of casein micelles in milk protein concentrate powders via small angle X-ray scattering. *Soft Matter, 7*, 3837-3843.

McMahon, D. J., & McManus, W. R. (1998). Rethinking casein micelle structure using electron microscopy. *Journal of Dairy Science, 81*, 2985-2993.

McMahon, D. J., & Oommen, B. S. (2012). Casein micelle structure, functions and interactions. In P. F. Fox & P. L. H. McSweeney (Eds.), *Advanced Dairy Chemistry* (4th ed., Vol. 1A Proteins: Basic Aspects, pp. 185-210). New York, NY, USA: Springer.

Meral, D., Toal, S. E., Schweitzer-Stenner, R., & Urbanc, B. (2015). Water-Centered Interpretation of Intrinsic pPII Propensities of Amino Acid Residues: In Vitro-Driven Molecular Dynamics Study. *The Journal of Physical Chemistry B, 119*, 13237-13251.

Mercier, J. C. (1981). Phosphorylation of caseins, present evidence for an amino-acid triplet code post-translationally recognized by specific kinases. *Biochimie, 63*, 1-17.

Mercier, J. C., & Vilotte, J. L. (1993). Structure and function of milk protein genes. *Journal of Dairy Science, 76*, 3079-3098.

Murray, N. J., Williamson, M. P., Lilley, T. H., & Haslam, E. (1994). Study of the interaction between salivary proline-rich proteins and a polyphenol by ^{1}H-NMR spectroscopy. *European Journal of Biochemistry, 219*, 923-935.

Nelson, R., Sawaya, M. R., Balbirnie, M., Madsen, A. O., Riekel, C., Grothe, R., & Eisenberg, D. (2005). Structure of the cross-[beta] spine of amyloid-like fibrils. *Nature, 435*, 773-778.

Niewold, T. A., Murphy, C. L., Hulskamp-Koch, C. A. M., Tooten, P. C. J., & Gruys, E. (1999). Casein related amyloid, characterization of a new and unique amyloid protein isolated from bovine corpora amylacea. *Amyloid-International Journal of Experimental and Clinical Investigation, 6*, 244-249.

Oftedal, O. T. (2012). The evolution of milk secretion and its ancient origins. *Animal, 6*, 355-368.

Ouanezar, M., Guyomarc'h, F., & Bouchoux, A. (2012). AFM Imaging of milk casein micelles: Evidence for structural rearrangement upon acidification. *Langmuir, 28*, 4915-4919.

Pappu, R. V., Srinivasan, R., & Rose, G. D. (2000). The Flory isolated-pair hypothesis is not valid for polypeptide chains: Implications for protein folding. *Proceedings of the National Academy of Sciences of the United States of America, 97*, 12565-12570.

Pappu, R. V., Wang, X., Vitalis, A., & Crick, S. L. (2008). A polymer physics perspective on driving forces and mechanisms for protein aggregation. *Archives of Biochemistry and Biophysics, 469*, 132-141.

Pascal, C., Pate, F., Cheynier, V., & Delsuc, M.-A. (2009). Study of the Interactions Between a Proline-Rich Protein and a Flavan-3-ol by NMR: Residual Structures in the Natively Unfolded Protein Provides Anchorage Points for the Ligands. *Biopolymers, 91*, 745-756.

Podgornaia, A. I., & Laub, M. T. (2015). Pervasive degeneracy and epistasis in a protein-protein interface. *Science, 347*, 673-677.

Rijnkels, M., Kooiman, P. M., de Boer, H. A., & Pieper, F. R. (1997). Organization of the bovine casein gene locus. *Mammalian Genome, 8*, 148-152.

Rodriguez, J. A., Ivanova, M. I., Sawaya, M. R., Cascio, D., Reyes, F. E., Shi, D., Sangwan, S., Guenther, E. L., Johnson, L. M., Zhang, M., Jiang, L., Arbing, M. A., Nannenga, B. L., Hattne, J., Whitelegge, J., Brewster, A. S., Messerschmidt, M., Boutet, S., Sauter, N. K., Gonen, T., & Eisenberg, D. S. (2015). Structure of the toxic core of [agr]-synuclein from invisible crystals. *Nature, 525*, 486-490.

Sawaya, M. R., Sambashivan, S., Nelson, R., Ivanova, M. I., Sievers, S. A., Apostol, M. I., Thompson, M. J., Balbirnie, M., Wiltzius, J. J. W., McFarlane, H. T., Madsen, A. O., Riekel, C., & Eisenberg, D. (2007). Atomic structures of amyloid cross-beta spines reveal varied steric zippers. *Nature, 447*, 453-457.

Shukla, A., Narayanan, T., & Zanchi, D. (2009). Structure of casein micelles and their complexation with tannins. *Soft Matter, 5*, 2884-2888.

Stewart, A. F., Bonsing, J., Beattie, C. W., Shah, F., Willis, I. M., & Mackinlay, A. G. (1987). Complete nucleotide-sequences of bovine alpha-s2-case in and beta-casein cDNAs - comparisons with related sequences in other species. *Molecular Biology and Evolution, 4*, 231-241.

Tagliabracci, V. S., Engel, J. L., Wen, J., Wiley, S. E., Worby, C. A., Kinch, L. N., Xiao, J., Grishin, N. V., & Dixon, J. E. (2012). Secreted Kinase Phosphorylates Extracellular Proteins That Regulate Biomineralization. *Science, 336*, 1150-1153.

Theillet, F.-X., Kalmar, L., Tompa, P., Han, K.-H., Selenko, P., Dunker, A. K., Daughdrill, G. W., & Uversky, V. N. (2013). The alphabet of intrinsic disorder. *Intrinsically Disordered Proteins, 1*, e24360.

Thompson, M. J., Sievers, S. A., Karanicolas, J., Ivanova, M. I., Baker, D., & Eisenberg, D. (2006). The 3D profile method for identifying fibril-forming segments of proteins. *Proceedings of the National Academy of Sciences of the United States of America, 103*, 4074-4078.

Thorn, D. C., Meehan, S., Sunde, M., Rekas, A., Gras, S. L., MacPhee, C. E., Dobson, C. M., Wilson, M. R., & Carver, J. A. (2005). Amyloid fibril formation by bovine milk kappa-casein and its inhibition by the molecular chaperones alpha(s-) and beta-casein. *Biochemistry, 44*, 17027-17036.

Thorn, D. C., Ecroyd, H., Sunde, M., Poon, S., & Carver, J. A. (2008). Amyloid fibril formation by bovine milk alpha(s2)-casein occurs under physiological conditions yet is prevented by its natural counterpart, alpha(s1)-casein. *Biochemistry, 47*, 3926-3936.

Thorn, D. C., Ecroyd, H., Carver, J. A., & Holt, C. (2015). Casein structures in the context of unfolded proteins. *International Dairy Journal, 46*, 2-11.

Tokuriki, N., Stricher, F., Serrano, L., & Tawfik, D. S. (2008). How Protein Stability and New Functions Trade Off. *Plos Computational Biology, 4*, Article Number: e1000002

Tran, H. T., Mao, A., & Pappu, R. V. (2008). Role of Backbone−Solvent Interactions in Determining Conformational Equilibria of Intrinsically Disordered Proteins. *Journal of the American Chemical Society, 130*, 7380-7392.

Trejo, R., Dokland, T., Jurat-Fuentes, J., & Harte, F. (2011). Cryo-transmission electron tomography of native casein micelles from bovine milk. *Journal of Dairy Science, 94*, 5770-5775.

Treweek, T. M., Thorn, D. C., Price, W. E., & Carver, J. A. (2011). The chaperone action of bovine milk alpha(S1)- and alpha(S2)-caseins and their associated form alpha(S)-casein. *Archives of Biochemistry and Biophysics, 510*, 42-52.

Umeda, T., & Aoki, T. (2002). Relation between micelle size and micellar calcium phosphate. *Milchwissenschaft-Milk Science International, 57*, 131-133.

Uversky, V. N., Gillespie, J. R., & Fink, A. L. (2000). Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins-Structure Function and Genetics, 41*, 415-427.

Uversky, V. N., Oldfield, C. J., & Dunker, A. K. (2005). Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. *Journal of Molecular Recognition, 18*, 343-384.

Wang, Z., Liao, B.-Y., & Zhang, J. (2010). Genomic patterns of pleiotropy and the evolution of complexity. *Proceedings of the National Academy of Sciences of the United States of America, 107*, 18034-18039.

Weinreich, D. M., Delaney, N. F., DePristo, M. A., & Hartl, D. L. (2006). Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science, 312*, 111-114.

Westerheide, S. D., Raynes, R., Powell, C., Xue, B., & Uversky, V. N. (2012). HSF Transcription Factor Family, Heat Shock Response, and Protein Intrinsic Disorder. *Current Protein & Peptide Science, 13*, 86-103.

Williamson, M. P. (1994). The structure and function of proline-rich regions in proteins. *Biochemical Journal, 297*, 249-260.

Wiltzius, J. J. W., Sievers, S. A., Sawaya, M. R., Cascio, D., Popov, D., Riekel, C., & Eisenberg, D. (2008). Atomic structure of the cross-β spine of islet amyloid polypeptide (amylin). *Protein Science, 17*, 1467-1474.

Zarrinpar, A., Park, S. H., & Lim, W. A. (2003). Optimization of specificity in a cellular protein interaction network by negative selection. *Nature, 426*, 676-680.

Zeng, Y., & Gu, X. (2010). Genome factor and gene pleiotropy hypotheses in protein evolution. *Biology Direct, 5*, Article Number: 37

Supplementary Data

Casein micelle structures, functions and diversity C. Holt (2016) Int Dairy J.

Data S1. Accession codes of sequence data

P02662	CASA1_BOVIN	Alpha-S1-casein	CSN1S1	Bos taurus (Bovine)
O62823	CASA1_BUBBU	Alpha-S1-casein	CSN1S1	Bubalus bubalis (Domestic water buffalo)
P18626	CASA1_CAPHI	Alpha-S1-casein	CSN1S1	Capra hircus (Goat)
P04653	CASA1_SHEEP	Alpha-S1-casein	CSN1S1	Ovis aries (Sheep)
CX987842	CSN1S1_DOG	Alpha-S1-casein	CSN1S1	Canis lupus familiaris
P39035	CASA1_PIG	Alpha-S1-casein	CSN1S1	Sus scrofa (Pig)
H2QPK8	H2QPK8_PANTR	Alpha-S1-casein	Uncharacterized protein	Pan troglodytes (Chimpanzee)
O97943	CASA1_CAMDR	Alpha-S1-casein	CSN1S1	Camelus dromedarius (Dromedary) (Arabian camel)
C3W972	C3W972_EQUAS	Alpha s1 casein	csn1S1	Equus asinus africanus (donkey)
P47710	CASA1_HUMAN	Alpha-S1-casein	CSN1S1 CASA CSN1	Homo sapiens (Human)
Q9XSE3	Q9XSE3_TRIVU	Alpha-casein		Trichosurus vulpecula (Brush-tailed possum)
D0QJ96	D0QJ96_ORNAN	Alpha casein	CSN1	Ornithorhynchus anatinus (Duckbill platypus)
D0QJA2	D0QJA2_9MAMM	Alpha casein	CSN1	Tachyglossus aculeatus (Australian echidna)
P09115	CASA1_RABIT	Alpha-S1-casein	CSN1S1	Oryctolagus cuniculus (Rabbit)
P19228	CASA1_MOUSE	Alpha-S1-casein	Csn1s1 Csn1 Csna	Mus musculus (Mouse)
P04656	CASA1_CAVPO	Alpha-S1-casein	CSN1S1	Cavia porcellus (Guinea pig)
P02663	CASA2_BOVIN	Alpha-S2-casein	CSN1S2	Bos taurus (Bovine)
B6VPY2	B6VPY2_BUBBU	Alpha s2 casein	csn1s2	Bubalus bubalis (Domestic water buffalo)
P33049	CASA2_CAPHI	Alpha-S2-casein	CSN1S2	Capra hircus (Goat)
P04654	CASA2_SHEEP	Alpha-S2-casein	CSN1S2	Ovis aries (Sheep)
P39036	CASA2_PIG	Alpha-S2-casein	CSN1S2	Sus scrofa (Pig)
O97944	CASA2_CAMDR	Alpha-S2-casein	CSN1S2	Camelus dromedarius (Dromedary) (Arabian camel)
B7VGF9	CASA2_EQUAS	Alpha-S2-casein	CSN1S2	Equus asinus (Donkey)
P04655	CASA2_CAVPO	Alpha-S2-casein	CSN1S2	Cavia porcellus (Guinea pig)
P50419	CASA2_RABIT	Alpha-S2-casein	CSN1S2	Oryctolagus cuniculus (Rabbit)
Q02862	CS2LA_MOUSE	Alpha-S2-casein-like A	Csn1s2a Csng	Mus musculus (Mouse
P02667	CS2LA_RAT	Alpha-S2-casein-like A	Csn1s2a Csng	Rattus norvegicus (Rat)
P50418	CS2LA_RABIT	Alpha-S2-casein-like A	CSN1S2A	Oryctolagus cuniculus (Rabbit)
P02664	CS2LB_MOUSE	Alpha-S2-casein-like B	Csn1s2b Csnd Csne	Mus musculus (Mouse)

Q8CGR3	CS2LB_RAT	Alpha-S2-casein-like B	Csn1s2b Csnd	Rattus norvegicus (Rat)
D0QJ98	D0QJ98_ORNAN	Beta-like casein 2		Ornithorhynchus anatinus (Duckbill platypus)
D0QJA6	D0QJA6_9MAMM	Beta-like casein 2 variant 3		Tachyglossus aculeatus (Australian echidna)
P02666	CASB_BOVIN	Beta-casein	CSN2	Bos taurus (Bovine)
Q9TSI0	CASB_BUBBU	Beta-casein	CSN2	Bubalus bubalis (Domestic water buffalo)
P33048	CASB_CAPHI	Beta-casein	CSN2	Capra hircus (Goat)
P11839	CASB_SHEEP	Beta-casein	CSN2	Ovis aries (Sheep)
P39037	CASB_PIG	Beta-casein	CSN2	Sus scrofa (Pig)
Q9TVD0	CASB_CAMDR	Beta-casein	CSN2	Camelus dromedarius (Dromedary) (Arabian camel)
Q9GKK3	CASB_HORSE	Beta-casein	CSN2	Equus caballus (Horse)
Q9N2G8	Q9N2G8_CANFA	Beta-casein		Canis familiaris (Dog) (Canis lupus familiaris)
P05814	CASB_HUMAN	Beta-casein	CSN2 CASB	Homo sapiens (Human)
P09116	CASB_RABIT	Beta-casein	CSN2	Oryctolagus cuniculus (Rabbit)
P02665	CASB_RAT	Beta-casein	Csn2 Csnb	Rattus norvegicus (Rat)
P10598	CASB_MOUSE	Beta-casein	Csn2 Csnb	Mus musculus (Mouse)
G3U197	G3U197_LOXAF	Uncharacterized protein	CSN2	Loxodonta africana (African elephant)
Q9XSE4	CASB_TRIVU	Beta-casein	CSN2 BCAS	Trichosurus vulpecula (Brush-tailed possum)
D0QJ95	D0QJ95_ORNAN	Beta casein	CSN2	Ornithorhynchus anatinus (Duckbill platypus)
D0QJ99	D0QJ99_9MAMM	Beta casein	CSN2	Tachyglossus aculeatus (Australian echidna)
D0QJA4	D0QJA4_9MAMM	Beta-like casein 2 variant 1		Tachyglossus aculeatus (Australian echidna)
P02668	CASK_BOVIN	Kappa-casein	CSN3 CSN10 CSNK	Bos taurus (Bovine)
P11840	CASK_BUBBU	Kappa-casein	CSN3 CSN10 CSNK	Bubalus bubalis (Domestic water buffalo)
P02669	CASK_SHEEP	Kappa-casein	CSN3 CSN10 CSNK	Ovis aries (Sheep)
P02670	CASK_CAPHI	Kappa-casein	CSN3 CSN10 CSNK	Capra hircus (Goat)
P79139	CASK_CAMDR	Kappa-casein	CSN3 CSN10 CSNK	Camelus dromedarius (Dromedary) (Arabian camel)
P11841	CASK_PIG	Kappa-casein	CSN3 CSN10 CSNK	Sus scrofa (Pig)
P33618	CASK_RABIT	Kappa-casein	CSN3 CSN10 CSNK	Oryctolagus cuniculus (Rabbit)
P19442	CASK_CAVPO	Kappa-casein	CSN3 CSN10 CSNK	Cavia porcellus (Guinea pig)
P04468	CASK_RAT	Kappa-casein	Csn3 Csn10 Csnk	Rattus norvegicus (Rat)
P06796	CASK_MOUSE	Kappa-casein	Csn3 Csn10 Csnk	Mus musculus (Mouse)
P07498	CASK_HUMAN	Kappa-casein	CSN3 CASK CSN10	Homo sapiens (Human)
P82187	CASK_HORSE	Kappa-casein	CSN3 CSN10	Equus caballus (Horse)
E2QXF8	E2QXF8_CANFA	Uncharacterized protein	CSN3	Canis familiaris (Dog) (Canis lupus familiaris)
G3UDT9	G3UDT9_LOXAF	Uncharacterized protein	CSN3	Loxodonta africana (African elephant)

D0QJA9	D0QJA9_ORNAN	Kappa casein	CSN3	Ornithorhynchus anatinus (Duckbill platypus)
D0QJA7	D0QJA7_9MAMM	Kappa casein	CSN3	Tachyglossus aculeatus (Australian echidna)
Q9XSD6	Q9XSD6_TRIVU	Kappa casein	CASK	Trichosurus vulpecula (Brush-tailed possum)
F7E1V6			CSN3	Monodelphis domestica (Gray-tailed opossum)

Multiple Sequence alignments

Multiple amino acid sequence alignments α_{S1}-casein (CSN1S1; Data S2), α_{S2}-casein (CSN1S2; Data S3), α_{S2}-casein B-type (CSN1S2B; Data S4), β-casein (CSN2; Data S5) and κ-casein (CSN3; Data S6) were based on the previous work of Kawasaki *et al.*, (Kawasaki, Lafont, & Sire, 2011) but with an increased number of sequences, alignments of sequences into 5 rather than 4 groups and a refined method of alignment. Monotreme CSN2B amino acid sequences coded by exons 2-6 were aligned with CSN2, whereas those coded by exon 7 and the following exons were aligned with CSN1S2B. Polar tract sequences are coloured green. Canonical sites of phosphorylation by the Golgi kinase are coloured red. Identities are denoted by # in the last row and indels by -. Boxed sequences are encoded by candidate orthologous exons.

Data S2. Alignment of amino acid sequences of CSN1S1 genes

cow	MKLLILTCLVAVALARP	----------	--------	KHPIKHQGL--PQ-	---------	----------	EVLNE-NLLRFFVA---------	PFPEVFGK	EKVNELSK
buffalo	MKLLILTCLVAVALARP	----------	--------	KQPIKHQGL--PQ-	---------	----------	GVLNE-NLLRFFVA---------	PFPEVFGK	EKVNELST
goat	MKLLILTCLVAVALARP	----------	--------	KHPINHRGL--SP-	---------	----------	EVPNE-NLLRFVVA---------	PFPEVFRK	ENINELSK
sheep	MKLLILTCLVAVALARP	----------	--------	KHPIKHQGL--SS-	---------	----------	EVLNE-NLLRFVVA---------	PFPEVFRK	ENINELSK
pig	MKLLIFICLAAVALARP	----------	--------	KPPLRHQEH--LQ-	---------	NE------PDSRE	ELFKERKFLRFPEV---------	PLLSQFRQ	EIINELNR
camel	MKLLILTCLVAVALARP	----------	--------	KYPLRYPEV--FQ-	---------	NE------PDSIE	EVLNK-RKILELAV---------	VSPIQFRQ	ENIDEL-K
donkey	MKLLILTCLVAVALARP	----------	--------	KLPHRHPEI--IQ-	---------	NE------QDSRE	KVLKE-RKFPSFAL---------	HT---SRE	EYINELNR
dog	MKFLILTCLVAVALARP	----------	--------	KLPLRHPEL--TQ-	---------	NE------LDSRE	EVLKERQFLRF-AL---------	PTPRELRE	--------
human	MRLLILTCLVAVALARP	----------	--------	KLPLRYPER--LQ-	---------	NP------SESSE	----------	PIPLESRE	EYMNGMNR
chimpanzee	MRLLILTCLVAVALARP	----------	--------	KLPLRYPER--LQ-	---------	NP------SESSE	----------	PIPLESRQ	EYMNGMNR
guinea pig	MKLLILTCLVASAVAMP	----------	--------	KFPFRHTEL--FQ-	---------	TQRGGSSSSSSSE	ERLKE-ENIFKFDQ---------	QKELQ-RK	--------
rabbit	MKLLILTCLVATALARH	----------	--------	KFHLGHLKL--TQ-	---------	EQP-----ESSEQ	EILKE-RKLLRFVQ---------	TVPLELRE	EYVNELNR
mouse	MKLLILTCLVAAAFAMP	----------	--------	RLHSRNAVS--SQ-	---------	TQQQ----HSSSE	EIFKQ-PKYLNLNQ---------	----DLRQ	EFVNNMNR
possum	MKLLIFSCLMALALARP	----------	DVLHLSID	R-HIKHREVENRS-	NEDLIPLNE	--------VSSSE	ESLHQLNRDRRSPEKYELNKYRE	--------	--------
platypus	MKVLILACLVAVAVAMP	ESPSSSSSSE	EAPRLLTK	KRILRNQEYYLPHL	---------	EE---SRSSSSSE	----------	--------	ESTRPTLK
echidna	MKVLILACLVAFVVAMP	ESPSSSSSSE	EASKILTK	KRVQRDQEYYLPHQ	---------	EE---SVSSSSSE	----------	--------	ESTDRLKR
	# ## ## # #								

cow	----------	DIGSESTE	DQAMEDIK	QMEAESISSSE	E---IVPNSVE	------	QKHIQK-EDVPSERYLGYL	EQLLRLKKYKVPQL	EIVPNSAE	------
buffalo	----------	DIGSESTE	DQAMEDIK	QMEAESISSSE	E---IVPISVE	------	QKHIQK-EDVPSERYLGYL	EQLLRLKKYNVPQL	EIVPNLAE	------
goat	----------	DIGSESTE	DQAMEDAK	QMKAGSSSSSE	E---IVPNSAE	------	QKYIQK-EDVPSERYLGYL	EQLLRLKKYNVPQL	EIVPKSAE	------
sheep	----------	DIGSESIE	DQAMEDAK	QMKAGSSSSSE	E---IVPNSAE	------	QKYIQK-EDVPSERYLGYL	EQLLRLKKYNVPQL	EIVPKSAE	------
pig	----------	--------	NHGMEGHE	Q-RGSSSSSSE	E---VVGNSAE	------	QKHVQKEEDVPSQSYL---	GHLQGLNKYKLRQL	--------	EAIHDQ
camel	----------	DTRNEPTE	DHIMEDTE	RKESG-SSSSE	E---VVSSTTE	------	QKDILK-EDMPSQRYL---	EELHRLNKYKLLQL	--------	EAIRDQ
donkey	QRELLKEKQKDEHK	--------	EYLIEDPE	QQESSSTSSSE	E---VVPINTE	------	QKRIPR-EDMLYQHTL---	EALRRLSKYNQLQL	--------	QAIYAQ
dog	-RELLREKQNEGIK	--------	--------	QRQSSSTSSSE	E---VVPNNTE	------	QRQIPR-EDILYQRYL---	EQLRRLSQHNQLQ-	--------	GTIHDQ
human	QRNILREKQTDEIK	DTRNESTQ	NCVVAEPE	KMESSISSSSE	E---MSLSKCA	------	----------	EQFCRLNEYNQLQL	--------	QAAHAQ
chimpanzee	QRNILREKQTDEIK	DTRNESTQ	NCVMAEPE	KMESSISSSSE	E---ISLSKCA	------	----------	EQFCRLNEYNQLQL	--------	QAVHAQ
guinea pig	QS--------EKIK	EIISESTE	--------	QREASSISSSE	E---VVPKNTE	------	QKHIPQ-EDALYQQAL---	EQLSRLIKYHQLQM	--------	EVVHAQ
rabbit	QRELLREKENEEIK	GTRNEVTE	EHVLADRE	-TEASISSSSE	E---IVPSSTK	------	QKYVPR-EDLAYQPYV---	----------	--------	------
mouse	QRALLTE-QNDEIK	VTMDAASE	EQAMASAQ	E-DSSISSSSE	ESEEAIPNITE	------	QKNIAN-EDMLNQCTL---	EQLQRQFKYNQLLQ	--------	KASLAK
possum	----------	--------	--------	DLKT---SSSE	ES--VAP-STE	ESVRRQ	VEYNFN-EQEDASASRE--	RKIEDVSEQYRQYL	--------	------
platypus	-RLLLKEKPILHIL	--------	--------	KAPE--SSSSE	ES----DSAAE	------	KRLLR--EREFYQQQL---	----------	--------	------
echidna	-R-LLKDKPIFRLL	--------	--------	KATE--SSSSE	ES----DSAIE	------	KRILR--ERQYYQQKL---	----------	--------	------
				####	#					

cow	------	------	------	------	------	------	------	------	------	------	------	------	------	------	--------------
buffalo	------	------	------	------	------	------	------	------	------	------	------	------	------	------	--------------
goat	------	------	------	------	------	------	------	------	------	------	------	------	------	------	--------------
sheep	------	------	------	------	------	------	------	------	------	------	------	------	------	------	--------------
pig	------	------	------	------	------	------	------	------	------	------	------	------	------	------	--------------
camel	------	------	------	------	------	------	------	------	------	------	------	------	------	------	KLIPRVKLSSHPYL
donkey	------	------	------	------	------	------	------	------	------	------	------	------	------	------	--------------
dog	------	------	------	------	------	------	------	------	------	------	------	------	------	------	--------------
human	------	------	------	------	------	------	------	------	------	------	------	------	------	------	--------------
chimpanzee	------	------	------	------	------	------	------	------	------	------	------	------	------	------	--------------
guinea pig	------	------	------	------	------	------	------	------	------	------	------	------	------	------	--------------
rabbit	------	------	------	------	------	------	------	------	------	------	------	------	------	------	--------------
mouse	QASLFQ	QPSLVQ	QASLFQ	QPSLLQ	QASLFQ	QPSMAQ	QASLLQ	QLLLAQ	QPSLAL	QVSPAQ	QSSLVQ	QAFLAQ	QASLAQ	KHHPRLSQSYYPHM	
possum	------	------	------	------	------	------	------	------	------	------	------	------	------	RRRPEERALNLRYL	
platypus	------	------	------	------	------	------	------	------	------	------	------	------	------	--------------	
echidna	------	------	------	------	------	------	------	------	------	------	------	------	------	--------------	

cow	-ERLHSMKEGIHAQQ	KEPMIGVNQ	ELAYFYPE	-----LFRQFYQLDAYPSGAWYYVPLGTQYTDAPSFSDIPNPIGSENSEKTT-MPLW----------	--
buffalo	-EQLHSMKEGIHAQQ	KEPMIGVNQ	ELAYFYPQ	-----LFRQFYQLDAYPSGAWYYVPLGTQYPDAPSFSDIPNPIGSENSGKTT-MPLW----------	--
goat	-EQLHSMKEGNPAHQ	KQPMIAVNQ	ELAYFYPQ	-----LFRQFYQLDAYPSGAWYYLPLGTQYTDAPSFSDIPNPIGSENSGKTT-MPLW----------	--
sheep	-EQLHSMKEGNPAHQ	KQPMIAVNQ	ELAYFYPQ	-----LFRQFYQLDAYPSGAWYYLPLGTQYTDAPSFSDIPNPIGSENSGKIT-MPLW----------	--
pig	-E-LHRTNEDKHTQQ	GEPMKGVNQ	EQAYFYFE	-----PLHQFYQLDAYPYATWYYPP---QYIAHPLFTNIPQPTAPEKGGKTEIMPQW----------	--
camel	-EQLYRINEDNHPQL	GEPVKVVTQ	EQAYFHLE	-----PFPQFFQLGASPYVAWYYPPQVMQYIAHPSSYDTPEGIASEDGGKTDVMPQWW----------	--
donkey	-EQLLRMKENS--QR	K-PMRVVNQ	--------	-----PFQPSYQLDVYPYAAWFHPAQIMQHVAYSPFHDTAKLIASENSEKTDIIPEW----------	--
dog	-QQLRRVNENNLLQL	---------	--------	-----PFQQFYQLDAYPYAAWYFPAQIMQYIAYPPSLDITKPIASENIENADVVPQW----------	--
human	-EQIRRMNENSHVQV	---------	--------	-----PFQQLNQLAAYPYAVWYY-PQIMQYVPFPPFSDISNPTAHENYEKNNVMLQW----------	--
chimpanzee	-EQIRRMNENSHVQV	---------	--------	-----PFQQLNQLAAYPCAVWYY-PQIMQYVPFPPFSDISNPTAHENYEKNNVMLQW----------	--
guinea pig	-EQFHRINEHNQAQV	KEPMRVFNQ	--------	------------LDAYPFAAWYYGPE-VQYMSFLPFSSIPQPIFPEDAQNTEVMPEWVM----------	--
rabbit	QQQLLRMKERYQIQE	REPMRVVNQ	ELAQLYLQ	-----PFEQPYQLDAYLPAPWYYTPEVMQYVLSPLFYDLVTPSAFESAEKTDVIPEWLKN----------	--
mouse	-EQPYRMNAYSQVQM	RHPMSVVDQ	ALAQFSVQ	-----PFPQIFQYDAFPL--WAYFPQDMQYLTPKAVLNTFKPIVSKDTEKTNV---W----------	--
possum	----------------	---------	EPL-YYAT	----------EP-DFYYTYVPISMPRFFPYPAEAPVFSTRKAPVPSINRATEAVYTYSEEK----------	KN
platypus	----------------	---------	--------	D---EYYRQFEP-DFYPRAYPKK--EVMPYPLEYFIPQAAVYSIPQLVYRVPQEVTFPSPLRFRYAFPQPTLPVE	RK
echidna	----------------	---------	--------	DELKEYFRQFEP-YFYPVAYQKK--EVMPYQLEYFVPQPEVYSIPQPVYRVPQEVTFPSLLHFRYAFPQSTLPIE	RK

Data S3. Alignment of amino acid sequences from CSN1S2 genes

cow	MKFFIFTCLLAVALAKN	TMEHVSSSE	ESI--ISQE-----	TYKQEKNMAINPSK	ENLCSTFCK	EVVRNA--NEE	--------	EYSIGSSSE	ESAEVATE
buffalo	MKFFIFTCLLAVALAKH	TMEHVSSSE	ESI--ISQE-----	TYKQEKNMAIHPSK	ENLCSTFCK	EVIRNA--NEE	--------	EYSIGSSSE	ESAEVATE
goat	MKFFIFTCLLAVALAKH	KMEHVSSSE	EPI-NIFQE-----	IYKQEKNMAIHPRK	EKLCTTSCE	EVVRNA--NEE	--------	EYSIRSSSE	ESAEVAPE
sheep	MKFFIFTCLLAVALAKH	KMEHVSSSE	EPI-NISQE-----	IYKQEKNMAIHPRK	EKLCTTSCE	EVVRNA--DEE	--------	EYSIRSSSE	ESAEVAPE
pig	MKFFIFTCLLAVAFAKH	EMEHVSSSE	ESI-NISQE-----	KYKQEKNVINHPSK	EDICATSCE	EAVRNI--KEV	--------	GYASSSSSE	ESVDIPAE
camel	MKFFIFTCLLAVVLAKH	EMDQGSSSE	ESI-NVSQQ-----	KFKQVKKVAIHPSK	EDICSTFCE	EAVRNI--KEV	--------	---------	ESAEVPTE
donkey	MKFFIFTCLLAVALAKH	NMEHRSSSE	DSV-NISQE-----	KFKQEKYVVIPTSK	ESICSTSCE	EATRNI--NEM	ESAKFPTE	VYSSSSSSE	ESAKFPTE
guinea pig	MKLFIFTCLLAVALAKH	KSEQQSSSE	ESV-SISQE-----	KFK-DKNMDTISSE	ETICASLCK	EATKNT--PKM	--------	AFFSRSSSE	EFADIHRE
rabbit a	MRFFVFTCLLAVALAKN	GIEQRSASE	EIV-SFYQE-----	KYKQDSNAAIYPTN	---------	----------	--------	---SVSSSE	ESVEVQTE
mouse a	MKFFIFACLVVVALAKH	EIKDKSSSE	ESSASIYPG-----	KSKLDNSVFFQTTK	---------	----------	--------	DSASSSSSE	ESSEEVSE
rat a	MKFFIFTCLVAAALAKH	AVKDKPSSE	ESA-SVYLG-----	KYKQGNGVFFQTPQ	---------	----------	--------	DSASSSSSE	ESSEEISE
	# # # ## ###	##		#					# #

cow	EVKITVDDKHYQKAL	NEINQFYQK--FPQYLQYLYQG-PIVLNPWDQVKRNAVPI-TPTL	NR--EQLSTSE	---------	ENSKKTVDM	---------
buffalo	EVKITVDDKHYQKAL	NEINQFYQK--FPQYLQYLYQG-PIVLNPWDQVKRNAVPI-TPTL	NR--EQLSTSE	---------	ENSKKTVDM	---------
goat	EIKITVDDKHYQKAL	NEINQFYQK--FPQYLQYPYQG-PIVLNPWDQVKRNAGPF-TPTV	NR--EQLSTSE	---------	ENSKKTIDM	---------
sheep	EVKITVDDKHYQKAL	NEINQFYQK--FPQYLQYLYQG-PIVLNPWDQVKRNAGPF-TPTV	NR--EQLSTSE	---------	ENSKKTIDM	---------
pig	NVKVTVEDKHYLKQL	EKISQFYQK--FPQYLQALYQA-QIVMNPWDQTKTSAYPF-IPTV	IQSGEELSTSE	-EPVSSSQE	E-NTKTVDM	---------
camel	----------------	NKISQFYQKWKFLQYLQALHQG-QIVMNPWDQGKTRAYPF-IPTV	NT--EQLSISE	---------	E-STEVPTE	---------
donkey	REEKEVEEKHHLKQL	NKINQFYEKLNFLQYLQALRQP-RIVLTPWDQTKTGASPF-IPIV	NT--EQLFTSE	---------	EIPKKTVDM	---------
guinea pig	----------------	NKKDQLYQKWMVPQYNPDFYQR-PVVMSPWNQIYTRPYPIVLPTL	GK--EQISTIE	---------	DILKKTTAV	ESSSSSSTE
rabbit a	KDEQIEEENVYLKQL	KRIKQIFQKFYIPQYPE-VYQQ-QIVMNPWKHVKTTTYPV--PI-	----------	PETTRIPLE	EIVKKIVEM	---------
mouse a	KIVQSEEQKVNLNQQ	KKFKQFSQESSFSQCCTPLHQQQQSSVNQWPQP--NAIHN-TPTQ	----------	-ESISTSVE	EILKKIIDM	---------
rat a	KIEQSEEQKVNLNQQ	KKSKQFSQDSSFPQICT-PYQQ-QSSVNQWPQP--NAIYD-VPSQ	----------	-ESTSTSVE	EILKKIIDI	---------
		# # # # #				

cow	ESTEVFTK	KTKLTEEEKNRLNFL	KKISQRYQKFALPQYLKTVYQHQKAMKPWIQPKTKV-----IPYV	RYL
buffalo	ESTEVITK	KTKLTEEDKNRLNFL	KKISQHYQKFTWPQYLKTVYQYQKAMKPWTQPKTNV-----IPYV	RYL
goat	ESTKVFTK	KTKLTEEEKNRLNFL	KIISQYYQKFAWPQYLKTVDQHQKAMKRWTQPKTNA-----IPYV	RYL
sheep	ESTEVFTK	KTKLTEEEKNRLNFL	KKISQYYQKFAWPQYLKTVDQHQKAMKPWTQPKTNA-----IPYV	RYL
pig	ESMEEFTK	KTELTEEEKNRIKFL	NKIKQYYQKFTWPQYIKTVHQKQKAMKPWNHIKTNSYQI--IPNL	RYF
camel	ESTEVFTK	KTELTEEEKDHQKFL	NKIYQYYQTFLWPEYLKTVYQYQKTMTPWNHIK----------	RYF
donkey	ESTEVVTE	KTELTEEEKNYLKLL	NKINQYYEKFTLPQYFKIVHQHQTTMDPQSHSKTNSYQI--IPVL	RYF
guinea pig	KSTDVFIK	KTKMDEVQKLIQSLL	NIIHEYSQKAFWSQTLEDVDQYLKFVMPWNHYNTNADQVD-ASQE	RQA
rabbit a	--------	----------	IKFNQ-LHQFVIPQYVQALQQ-RIAMNPWHHVTPFRS----FPV-	LNF
mouse a	--------	----------	IKYIQ-YQQVTIPQLPQALHP-QIPVSYWYPSKDYTFPN--AHYT	RFY
rat a	--------	----------	VKYFQ-YQQLTNPHFPQAVHP-QIPVSSWAPSKDYTFPT--ARYM	--A

Data S4. Alignment of CSN1S2B-like sequences

rabbit b	MKFFIFTCLLAVALAKP	KIEQ-SSSE	ETI-AVSQEVSPNL	--------------	ENICSTACE	EPIKNI--NEV	EYVEVPTE
mouse b	MKFIILTCLLAVALAKQ	RMEQYISSE	ESM-DNSQE-----	NFKQNMDVAFFPSQ	---------	ETVENIYIPQM	ESVEAPMK
rat b	MKFIILTCLLAVALAKQ	---------	ESK-DNSQE-----	DFKQTVDVVIFPGQ	---------	ETVKNIPIPQM	ESVEAPIK
	### # ##########					# ##	# ## #

platypus 2b			KYQQR-LRLFKPTYL	VPVNKFVE-	RHPFRNILFPEELPEAYQPIE
echidna 2b			IYQQG-LRPFKPTHL	---------	RRPLKYIFFSEEPPKVYQPIQ
rabbit b	IKDQEFYQKVNLLQYLQALYQY-PTVMDPWTRAETKAIPF-IRTM	---------	QYKQEK-DATKHTSQ	---------	----------
mouse b	----------	-VSDIISQQ	QYNQKMMDMSVSARE	---------	----------
rat b	NKCYQSIQTFKPPQALKGLYQY-HMAKNPWGYTVNRAFPS-TRTL	---------	QYNQKTMDLSMRARE	---------	----------
			# #		

platypus 2b	KEDSSSSSE	ETVQVPVE	K-HLLRLRK-LHVPQ	-KLRP--LRFYPNHQVPFQRHPLPYAG----TQVHQPVEVPFPLP	VQY
echidna 2b	NEDSSSSSE	EPVEVPAE	QNHVLRLKK-LQVLQ	-NLQP--LRRLPNYQVPLQRHPLPFVRLPNVFQAPHPVELPFPLP	QVV
rabbit b	---------	--------	KTELTEEEKAFLKYL	DEMKQYYQKFVFPQYLKNAHHFQKTMNPWNHVKTIIYQS--VPTL	RYL
mouse b	---------	--------	KTVMTEESKNIQDYM	NKMKR-YSKITWPQFVKLLHQYQKTMTPWSYYPST-------PSQ	--V
rat b	---------	--------	KIVMSEIKKNIQDYV	TKMKQ-YSKITWPRFVKSLQQYQKTMNPWSCYPYT-------LLQ	--V
			#		

Data S5. Alignment of amino acid sequences of CSN2 genes

cow	MKVLILACLVALALARE---	LEELNVPGE	IVESL------SSSE	--------	----------	----------	ESITRINK	-KIEKFQSEEQQQTE
buffalo	MKVLILACLVALALARE---	LEELNVPGE	IVESL------SSSE	--------	----------	----------	ESITHINK	-KIEKFQSEEQQQME
goat	MKVLILACLVALAIARE---	QEELNVVGE	TVESL------SSSE	--------	----------	----------	ESITHINK	-KIEKFQSEEQQQTE
sheep	MKVLILACLVALALARE---	QEELNVVGE	TVESL------SSSE	--------	----------	----------	ESITHINK	-KIEKFQSEEQQQTE
pig	MKLLILACFVALALARA---	KEELNASGE	TVESL------SSSE	--------	----------	----------	ESITHISK	EKIEKLKREEQQQTE
camel	MKVLILACRVALALARE---	KEEFKTAGE	ALESI------SSSE	--------	----------	----------	ESITHINK	QKIEKFKIEEQQQTE
horse	MKILILACLVALALARE---	KEELNVSSE	TVESLSSNEPDSSSE	--------	----------	----------	--------	EKLQKFKHEGQQQRE
dog	MKVFILACLVALALARE---	KEELTLSNE	TVESL------SSSE	--------	----------	----------	ESITHINK	QKLENFKHEEQQQRE
human	MKVLILACLVALALARE---	----------	TIESL------SSSE	--------	----------	----------	ESITEY-K	QKVEKVKHEDQQQGE
rabbit	MKVLILACLVALALARE---	KEQLSVPTE	AVGSV------SSSE	--------	----------	----------	E-ITHINK	QKLETIKHVEQLLRE
rat	MKVFILACLVALALARE---	KDAFTVSSE	T-GSI-------SSE	--------	----------	----------	ESVEHINE	-KLQKVKLMGQVQSE
mouse	MKVFILACLVALALARE---	-TTFTVSSE	T-DSI-------SSE	--------	----------	----------	ESVEHINE	QKLQKVNLMGQLQAE
elephant	MKVFILACLVAFALGRE---	KEEIIVSTE	TVENL------SSSE	--------	----------	----------	ESVTQVNK	QKPEGVKHEEQQ-RE
opossum	MKLLILSCLVALAVARP---	----------	MVEKI------SETE	--------	----------	----------	EFVTVIPE	QQI----RREDVPVK
possum	MKLLILTCLVVLAVARP---	----------	MVEKI------SESE	--------	----------	----------	EHVTDVPE	----------------
platypus 2a	MKVFILSCLLAVAMAMP---	----------	---KL--QSSSSSSE	ETDQLLVK	EKLVKRRELM	-DLPTTLSSE	--------	-EHVMEEKEFYQPRL
echidna 2a	MKVFILACLVAVAMALP---	----------	---KQ--HSSSSSSE	ESDRLLVK	EKLMRRRKLM	-DIPTAFSSE	--------	-EHSVDPKELYEPRQ
platypus 2b	MKVFILACLVAAAVAVPVST	----------	----------	EFDKLLVK	EKLLKHRDLV	KDLPTIFSSE	--------	WEQFLRHPEVYVPLE
echidna 2b	MKVFIFACLVAVAMAVP---	----------	---KQ--QSSSSSSE	ETDKQLVM	ENLLKHRALV	KDIPTTFSSE	ENINYEKQ	WEQLLRQPMVYEPFE
	## # # #							

cow	DELQDKIHPFAQTQSLVYPFP--GPIPN-SLPQNIPPLTQTPVVVP---PFLQPEVMGVSKVKEAMAPKQKEMPFPKYP-VEPFTESQSLTL---TDVENLHLPLP
buffalo	DELQDKIHPFAQTQSLVYPFP--GPIPK-SLPQNIPPLTQTPVVVP---PFLQPEIMGVSKVKEAMAPKHKEMPFPKYP-VEPFTESQSLTL---TDVENLHLPLP
goat	DELQDKIHPFAQAQSLVYPFT--GPIPN-SLPQNILPLTQTPVVVP---PFLQPEIMGVPKVKETMVPKHKEMPFPKYP-VEPFTESQSLTL---TDVEKLHLPLP
sheep	DELQDKIHPFAQAQSLVYPFT--GPIPN-SLPQNILPLTQTPVVVP---PFLQPEIMGVPKVKETMVPKHKEMPFPKYP-VEPFTESQSLTL---TDVEKLHLPLP
pig	NERQNKIHQFPQPQPLAHPYT--EPIPYPILPQNILPLAQVPVVV----PLLHPEVMKDSKAKETIVPKRKGMPFPKSP-AEPFVEGQSLTL---TDFEVLSLPL-
camel	DEQQDKIYTFPQPQSLVYSHT--EPIPYPILPQNFLPPLQPAVMV----PFLQPKVMDVPKTKETIIPKRKEMPLLQSP-VVPFTESQSLTL---TDLENLHLPLP
horse	VERQDKISRFVQPQPVVYPYA--EPVPYAVVPQSILPLAQPPIL-----PFLQPEIMEVSQAKETILPKRKVMPFLKSP-IVPFSERQILNP---TNGENLRLPVH
dog	DERQNKIHPLFQQQPLVSPYA--DPIHYAILPQNILPLAQPAVVV----PFLQPEIMEVPKVKENIFPRHKVMPFLKSP-VTPFLDSQILNV---ADLENVHFPLP
human	DEHQDKIYPSFQPQPLIYPFV--EPIPYGFLPQNILPLAQPAVVL----PVPQPEIMEVPKAKDTVYTKGRVMPVLKSP-TIPFFDPQIPKL---TDLENLHLPLP
rabbit	EKLQDKILPFIQS---LFPFA--ERIPYPTLPQNILNLAQLDMLL----PLLQPEIMEDPKAKETIIPKHKLMPFLKSPKTVPFVDSQILNL---REMKNQHLLLP
rat	DVLQNKFHSGIQSEPKAIPYA--QTISCSPIPQNIQPIAQPPVV-PTDGPIISPELESFLKAKATVLPKHKQMPFLNSETVLRLFNSQIPSL----DLANLHLPQS
mouse	DVLQAKVHSSIQSQPQAFPYAQAQTISCNPVPQNIQPIAQPPVV-PSLGPVISPELESFLKAKATILPKHKQMPLLNSETVLRLINSQIPSL---ASLANLHLPQS
elephant	DEHQNKIQPLFQPQPLVYPFA-EPIPYTVFPPNAIPLAQPIVVL-----PFPQPEVKQLPEAKEITFPRQKLMSFLKSP---VMPFFDPQIPNLGTDLENLHLPLP
opossum	NERHPEINRFIPLEAETMSF-YVPVYWPEEMRDAKMTSPLKEKRMTLANPIAPEEELPHLQHKSLSLAKQRFLASLRP-KAAQPFYAPRMAPLPHKLFTMPKEQA-
possum	NEHRLEINRYLRPEYEMMNLYYQPFYWSEEMRNLKMTSLPKDRRMAVLKSVVSDDMLPPLQHKSLSLPKPKVLPLSH--RQILPPHTLRMVPLSHKLFTIPKREM-
platypus 2a	----------KYPYPFFPPIKTYVNPHIYQKPAVLPVTHPETLTYLQPQQNPEDMPLP--KKEVLPYLKAVVVPYPQVQVMPYPETEVMPYFPPMTMSLVQPDIV
echidna 2a	---------SYSYPWQSVRPINTYTYPRAYQIPAVLPMTHPQTLTYLQPQFKPEDMSISQ--KQIPPYVQAVVMPYPQVEAIPFPGAEFMPYAQPITTPLLQPEVF
	#

cow	----------LLQSWMHQPHQPLPPTV-MFPPQSVLSLSQSKVLPVPQKAVPY--PQRDMPIQAFLLYQEPVLGPVRGPFPI----------I	V-----
buffalo	----------LLQSWMHQPPQPLPPTV-MFPPQSVLSLSQSKVLPVPQKAVPY--PQRDMPIQAFLLYQEPVLGPVRGPFPI----------I	V-----
goat	----------LVQSWMHQPPQPLSPTV-MFPPQSVLSLSQPKVLPVPQKVVP----QRDMPIQAFLLYQEPVLGPVRGPFPI----------LV	V-----
sheep	----------LVQSWMHQPPQPLPPTV-MFPPQSVLSLSQPKVLPVPQKAVP----QRDMPIQAFLLYQEPVLGPVRGPFPI----------L	V-----
pig	----------L-QSLMHQIPQPVPQTP-MFAPQPLLSLPQAKVLPVPQQVVPF--PQRDMPFQALLLYQDPLLGPLQGFYPVPQPVAPVYNP------	V-----
camel	----------LLQSLMYQIPQPVPQTP-MIPPQSLLSLSQFKVLPVPQQMVPY--PQRAMPVQAVLPFQEPVPDPVRGLHPVPQPLVPV--------I	A-----
horse	----------LIQPFMHQVPQSLLQTL-MLPSQPVLSPPQSKVAPFPQPVVPY--PQRDTPVQAFLLYQDPRLGPTGELDPATQPIVAVHNP----VI	V-----
dog	LSLPLLQPLMHQIPQPLPLLQPLMHQIPQPLPQTP-MLTPQSVLSIPQPKVLPFPQQVVPYL--QRDMPLQAFLPYQE----STHQAQPVTQPLAPLVNS----AL	V-----
human	----------LLQPLMQQVPQPIPQTL-ALPPQPLWSVPQPKVLPIPQQVVPY--PQRAVPVQALLLNQELLLNPTHQIYPVTQPLAPVHNP----IS	V-----
rabbit	----------QLLPFMHQVFQPFPQTP-IPYPQALLSLPQSKFMPIVPQVVPY--PQRDMPIQALQLFQE-LLFPTHQGYPVVQPIAPVN--------	V-----
rat	----------PAQLQAQIVQAFP-QTPAVVSSQPQLSHPQSKSQYLVQQLAPL--FQQGMPVQDLLQYLDLLLNPTLQFLATQQLHSTS---------	V-----
mouse	----------LVQLLAQVVQAFP-QTH-LVSSQTQLSLPQSKVLYFLQQVAPF--LPQDMSVQDLLQYLE-LLNPTVQFPATPQ-HSVS---------	V-----
elephant	----------LLQPLRHQLHQPLAQTP-VLP----LPLSLPKVLPVPQQVIPY--PQRGRPIQNLQLYEEPLLDPTRKIYPVAQPLAPVYNPVA----	V-----
opossum	----------LPIAKRDMLSAAELVIPAVH----------ERVIPAIDKREPLPLLAREMPA---LPDKE----------	K----Y
possum	----------LPISERERLPA-HERENLLAHEREILLAP-QREMSLIPEREILLAAERVVLP---EQEREIRPDNEREVLAVHKREILPASEKEKVLP	K----N
platypus 2a	----------PPSFYREAVIQQLAVPFVR-RESAL---PHQRAIVPVATAAAAV---RESLPLVQQ---EVVPPIMPLDVYLVRHPEVSFYNPTE---	KIPETN
echidna 2a	----------SAPFYREAV---------LFQERV--LPLHRREIVPPYQRDTIA--RREILPVDQ---RELMPEVVAVDLYPFFQPVANFYYPAELNE	KIPETD

Data S6. Alignment of amino acid sequences of CSN3 genes

platypus	-MKTLLLVGAILAMTVGFS	VAEEQKWKRLD	SSESEERWWRLRLKPSLLFRVQDKP------ERNIPRPSYPYPLLNVPHPNAINPEHQRPYVLP--RFNF-QIPN
echidna	-MKTLLLVGGILVMTVCFS	AAEDEEWKKVD	YSESEERWLRLKRQPSFPFSFQGKP------ERNIPRPYYPRPFLNIPRPYTINPEHQFAYVFPNLKF---QIPS
possum	-MKVLFLTVHILAVMVCFS	TADL-DWEKWP	CDKQNERQ-SELRQQ------PLRRSPVQYVYTPYTHQ-SYVPVIYPPRAYVRHPYFSRVAWQKPYPSYMPLLPS
opossum	-MKVLFLIGHILLAMVCFS	TAEL-DWRKWP	CEKQMERP-SELEQQ------PPGQPPVQDVYTRYTRQ-IYVPILYAPKTSIQYPYFSKLAWQRPYAAYIPLLSS
pig	MMKSSFLIVPILALTLPFL	GAEEQNQEKLT	RC-ESDKRLFNEEKVKYIPIYYMLNRFPSYGF-FYQHRSAVSPNRQFIPYPYYARPVVAGPHAQKPQWQ--DQPN
human	-MKSFLLVVNALALTLPFL	AVEVQNQKQPA	CH-ENDERPFYQKTAPYVPMYYVPNSYPYYGTNLYQRRPAIAINNPYVPRTYYANPAVV--RPHAQIPQRQYLPN
rat	MMRNFIVVMNILALTLPFL	AAEVQNPDSN-	CR-EKNEVVYDVQRVLYTPVSSVLNRN-HYEPIYYHYRTSVP--VSPYAYFPVGLKLLLL-RSPAQILKWQPMPN
mouse	MMRNFIVVVNILALTLPFL	AAEIQNPDSN-	CRGEKNDIVYDEQRVLYTPVRSVLNFN-QYEPNYYHYRPSLPATASPYMYYPLVVRLLLL-RSPAPISKWQSMPN
dog	MMKRFFLVVNIVALALPFL	GAEVQNQEQPT	CR-ENDERLFNQKTVKYIPIHYVLNSFSHYEPNYYPHRPAEPINHQ-YVPYPFYAKPAVAVRTHAQIPQWQVLPN
rabbit	MMKHFLLVVNILAVTLPFL	AADIQNQEQTT	CR-ENEERLFHQVTAPYIPVHYVMNRYPQYEPSYYLRRQAVPTLNP-FMLNPYYVKPIV-FKPNVQVPHWQILPN
horse	-MKSFFLVVNILALTLPFL	GAEVQNQEQPT	CH-KNDERFFDLKTVKYIPIYYVLNSSPRYEPIYYQHRLALLINNQ-HMPYQYYARPAA-VRPHVQIPQWQVLPN
elephant	MMKGFLLVVNILLLPLPFL	AAEVQNQEESR	CL-EKDERWFCQKAVKYIPNDYVLKSYYRYEPNYNQFRAAVPINNP-YLIYLYPAKQVA-VRPHTQIPQWQVPSN
camel	-MKSFFLVVTILALTLPFL	GAEVQNQEQPT	CF-EKVERLLNEKTVKYFPIQFVQSRYPSYGINYYQHRLAVPINNQ-FIPYPNYAKPVA-IRLHAQIPQCQALPN
cow	MMKSFFLVVTILALTLPFL	GAQEQNQEQPI	R-CEKDERFFSDKIAKYIPIQYVLSRYPSYGLNYYQQKPVALINNQ-FLPYPYYAKPAA-VRSPAQILQWQVLSN
sheep	MMKSFFLVVTILALTLPFL	GAQEQNQEQRI	C-CEKDERFFDDKIAKYIPIQYVLSRYPSYGLNYYQQRPVALINNQ-FLPYPYYAKPVA-VRSPAQTLQWQVLPN
goat	MMKSFFLVVTILALTLPFL	GAQEQNQEQPI	C-CEKDERFFDDKIAKYIPIQYVLSRYPSYGLNYYQQRPVALINNQ-FLPYPYYAKPVA-VRSPAQTLQWQVLPN
buffalo	MMKSFFLVVTILALTLPFL	GAQEQNQEQPI	R-CEKEERFFNDKIAKYIPIQYVLSRYPSYGLNYYQQKPVALINNQ-FLPYPYYAKPAA-VRSPAQILQWQVLPN
	# #		

platypus	ILP--------FLMFPE--LPPPFFPIVHPIYYDPQTPTTPR--------------NPPVTSQTPQPPVDSSANT-PEPPTTAPLTATPEAQTPL-QP
echidna	V----------FPFPLE--FLPPFYPFVHPIYYGPQTSTPPR--------------NPTVTSQTPQPPVHSSANT-PESATAAPVTATPMAQTPL-QP
possum	IYP--------WSVVSR-NLHPAFAFNPPHYAQLPVPSSPTNSPTTTIQTTNIPITNPTSTIVTPAVSSKSAATEDSAAAAMLTSPTAAQMA------
opossum	RYP--------WPVIPR-SPHPSFAFNPPQYARVPAPSGPTSSPAAPMETTTIPST--STVAATVTPDATSKFVTTEYSTTATIPTSPIPEQ----QP
pig	VYP-------PTVARRP-RPHASFIAIPPKKNQDKTA-IPAINSIATVEPTIVPAT--EPIVNAEPIVNAVVTPEASSEFLITSAPETTTVQVTSPVV
human	SHP--------PTVVRRPNLHPSFIAIPPKKIQDKII-IPTINTIATVEPTPAPAT--EPTVDSVVTPEAFSESIITSTPETTTVAVT------PPTA
rat	FPQ--------PVGVPHPIPNPSFLAIPTNEKHDNTA-IPASNTIAPIVSTPVSTT--ESVVNTVANTEAST--VPISTPETATVPVT------SPAA
mouse	FPQ--------SAGVPYAIPNPSFLAMPTNENQDNTA-IPTIDPITPIVSTPVPTM--ESIVNTVANPEAST--VSINTPETTTVPVS------STAA
dog	AYP--------PTMMHRPQLHPSFIAIPPKKIQDKTS-IPTINTIATAEATPIPL---EPKVNTAVTSDASSEFTITSTPETTTVPVT------SPVV
rabbit	IHQ----------PKVGRHSHPFFMAILPNKMQDKAV-TPTTNTIAAVEPTPIPTT--EPVVSTEVIAEASPELII--SPETTTEATAA-----SAAA
horse	IYP--------STVVRHPCPHPSFIAIPPKKLQEITV-IPKINTIATVEPTPIPTP--EPTVNNAVIPDASSEFIIASTPETTTVPVTSPV---VQKL
elephant	IYP--------SPSVPHTYLKPPFIVIPPKKTQDKPI-IPPTGTVASIEATV------EPKVNTVVNAEASSEFIATNTPEATTVPVI------SPQI
camel	IDP-------PTVERRP-RPRPSFIAIPPKKTQDKTV-NPAINTVATVEPPVIPTA--EPAVNTVVIAEASSEFITTSTPETT-VQIT------STEI
cow	TVPAKSCQAQPTTMARHPHPHLSFMAIPPKKNQDKTE-IPTINTIASGEPTSTPTT--EAVESTVATLEDSPEVIESP-PEINTVQVT------STAV
sheep	AVPAKSCQDQPTAMARHPHPHLSFMAIPPKKDQDKTE-IPAINTIASAEPTVHSTPTTEAVVNAVDNPEASSESIASAP-ETNTAQVT------STEV
goat	TVPAKSCQDQPTTLARHPHPHLSFMAIPPKKDQDKTE-VPAINTIASAEPTVHSTPTTEAIVNTVDNPEASSESIASAS-ETNTAQVT------STEV
buffalo	TVPAKSCQAQPTTMTRHPHPHLSFMAIPPKKNQDKTE-IPTINTIVSVEPTSTPTT--EAIENTVATLEASSEVIESV-PETNTAQVT------STVV
	#

Kawasaki, K., Lafont, A.-G., & Sire, J.-Y. (2011). The evolution of casein genes from tooth genes before the origin of mammals. ***Molecular Biology and Evolution, 28*****, 2053-2061.**