Journal of Membrane Biology 251 (2018) 653-666

Evolutionary analysis of the lysine-rich N-terminal cytoplasmic domains of the gastric H^+,K^+-ATPase and the Na^+,K^+-ATPase

Dil Diaz[1] and Ronald J. Clarke[1,2,*]

[1]*School of Chemistry, University of Sydney, Sydney, NSW 2006, Australia*

[2] *The University of Sydney Nano Institute, Sydney, NSW 2006, Australia*

Address correspondence to Assoc. Prof. Ronald J. Clarke, School of Chemistry, University of Sydney, Sydney, NSW 2006, Australia. Tel.: 61-2-93514406; Fax: 61-2-93513329; E-mail: ronald.clarke@sydney.edu.au; ORCID 0000-0002-0950-8017.

Abstract

The catalytic α-subunits of both the Na^+,K^+-ATPase and the gastric H^+,K^+-ATPase possess lysine-rich N-termini which project into the cytoplasm. Due to conflicting experimental results it is currently unclear whether the N-termini play a role in ion pump function or regulation, and, if they do, by what mechanism. Comparison of the lysine frequencies of the N-termini of both proteins with those of all of their extramembrane domains showed that the N-terminal lysine frequencies are far higher than one would expect simply from exposure to the aqueous solvent. The lysine frequency was found to vary significantly between different vertebrate classes, but this is due predominantly to a change in N-terminal length. As evidenced by a comparison between fish and mammals, an evolutionary trend towards an increase of the length of the N-terminus of the H^+,K^+-ATPase on going from an ancestral fish to mammals could be identified. This evolutionary trend supports the hypothesis that the N-terminus is important in ion pump function or regulation. In placental mammals, one of the lysines is replaced by serine (Ser-27), which is a target for protein kinase C. In most other animal species a lysine occupies this position and hence no protein kinase C target is present. Interaction with protein kinase C is thus not the primary role of the lysine-rich N-terminus. The disordered structure of the N-terminus may, via increased flexibility, facilitate interaction with another binding partner, e.g. the surrounding membrane, or help to stabilize particular enzyme conformations via the increased entropy it produces.

Keywords: sodium pump; gastric proton pump; protein intrinsic disorder; stomach pH; amino acid sequence analysis; protein kinase C

Introduction

P-type ATPases are a family of enzymes whose major physiological role is the pumping of either ions or phospholipids across biological membranes. One of its prominent members is the Na^+,K^+-ATPase (or sodium pump), which is responsible for the Na^+ and K^+ electrochemical gradients across the plasma membrane of all multicellular animal cells (Kaplan 2002). Another is the gastric H^+,K^+-ATPase (or proton pump), whose activity creates the low pH necessary for the activation of pepsin, the main protein-digesting enzyme of the stomach (Shin et al. 2009). The Na^+,K^+-ATPase and the gastric H^+,K^+-ATPase are closely related, both belonging to the subfamily of type IIC P-type ATPases (Axelsen and Palmgren 1998). The H^+,K^+-ATPase is thought to have evolved by gene duplication from an ancestral Na^+,K^+-ATPase before the evolution of cartilaginous fish ~400 million years ago (Okamura et al. 2002; Choe et al. 2004).

Both of these ion pumps possess lysine-rich N-terminal extensions of their catalytic α subunits which project into the cytoplasm of the cell. From the study of tryptic digestion patterns (Jørgensen 1975; Jørgensen et al. 1982; Jørgensen and Collins 1986; Jørgensen and Andersen 1988) and experiments in which the N-terminus was removed either by proteolytic cleavage (Cornelius et al. 2005; Jørgensen and Collins 1986) or by mutagenesis (Scanzano et al. 2007), it is known that the N-terminus of the Na^+,K^+-ATPase undergoes substantial movement during conformational changes crucial to its ion pumping mechanism. It is also known that serine residues of the N-terminus are capable of phosphorylation by protein kinase C (Beguin et al. 1994; Feschenko and Sweadner 1995; Logvinenko et al. 1996). Similarly, phosphorylation of the N-terminus of the H^+,K^+-ATPase by protein kinase C has been demonstrated and been shown to affect the maximal enzyme activity (Cornelius and Mahmmoud, 2003). Togawa et al. (1996) previously identified the site of protein kinase C phosphorylation of the H^+,K^+-ATPase as serine-27. Therefore, it is thought that the N-termini

are involved in pump regulation, and they are even sometimes referred to as R (i.e. regulatory) domains of the proteins (Morth et al. 2011). Recent experimental and theoretical studies on the Na^+,K^+-ATPase via a variety of techniques have shown (Jiang et al. 2017; Garcia et al. 2017; Nguyen et al. 2018) that the lysine residues of the N-terminus are capable of interacting with the negatively charged head groups of anionic phospholipids in the neighbouring membrane. Unfortunately no x-ray crystallographic data on the N-terminus of either the Na^+,K^+- or the H^+,K^+-ATPase is available, because it could either not be resolved (Morth et al. 2007; Shinoda et al. 2009; Kanai et al. 2013; Nyblom et al. 2013) or it was removed prior to crystal formation (Abe et al. 2018). Furthermore, in contrast to the experiments in native tissue described above, experimental data on the Na^+,K^+-ATPase and H^+,K^+-ATPase after exogenous expression in cell lines do not support a role of the N-terminus in determination of ion pumping activity. Thus, Daly et al. (1996) found that deletion of the Na^+,K^+-ATPase N-terminus had no effect on activity. Similarly, Asano et al. (2000) found that mutating all of the lysines of the H^+,K^+-ATPase N-terminus to alanines had no effect on its kinetic properties. Therefore, the detailed function of the N-terminus remains unclear.

Analysis of the amino acid sequence of the N-terminus of the α_1 subunit of vertebrate Na^+,K^+-ATPases up to its conserved LKKE motif has shown (Nguyen et al. 2018) that the frequency of lysine residues is at least 4 times greater than the overall lysine frequency in over 1021 unrelated proteins (McCaldon and Argos 1988). However, this alone is not proof of a functional role of the N-terminal lysine residues in the Na^+,K^+-ATPase. Because the N-terminus extends into the cytoplasm, one would naturally expect a higher frequency of charged amino acids simply because of the surrounding aqueous environment. Therefore, one open question is whether there is a preferential enrichment of lysine in the N-terminus relative to other extramembrane domains of the Na^+,K^+- and H^+,K^+-ATPases. To answer this

question, here we have compared the lysine frequencies of the solvent-accessible extramembrane domains, the N-terminus alone and the transmembrane domains of the Na^+,K^+- and H^+,K^+-ATPases. The analysis clearly shows that lysine is preferentially concentrated in the N-termini, thus supporting the hypothesis of a mechanistic or regulatory role for the lysine residues rather than simply charging the protein surface and improving its hydration.

To further investigate the role of the ATPase N-terminus, we have also analysed variations in the lysine frequency of the N-terminus of the H^+,K^+-ATPase across different animal species. We have chosen the H^+,K^+-ATPase for this analysis rather than the Na^+,K^+-ATPase, because, in contrast to the Na^+,K^+-ATPase, the H^+,K^+-ATPase is a tissue-specific enzyme. It is only present in the gastric mucosa of the stomach, whereas the Na^+,K^+-ATPase is present in the plasma membrane of every multicellular animal cell, where, amongst other functions, it is crucial to the maintenance of the osmotic conditions across the membrane necessary to avoid any cell volume changes. If there is no significant variation in osmotic conditions between the cells of different animal species, then for the purposes of its ion pumping function alone there is no reason to expect any significant variation in Na^+,K^+-ATPase molecular activity. Differences in Na^+,K^+-ATPase activity between warm- and cold-blooded animals have been reported (Else and Wu 1999), however, which are probably related to the contribution of the Na^+,K^+-ATPase to heat production and maintenance of body temperature. The proton gradient produced by the H^+,K^+-ATPase across the membrane of the stomach parietal cells on the other hand is known to be very species-dependent. The pH inside an animal's stomach depends strongly on its diet (Beasely et al. 2015). Carnivores, which have a high protein diet, require a low stomach pH, not only in order to activate the digestive enzyme pepsin, but also as a protection against pathogenic foreign microbes in their food. In contrast, herbivores rely heavily on gut bacteria for the digestion of their food which

can only function effectively at pH values closer to neutral. Thus, carnivores must have a higher net pumping rate of H^+ ions into their stomachs than herbivores, either via a higher forward rate of H^+ pumping by the H^+,K^+-ATPase, by a reduced backward movement of H^+ ions out of the stomach into the cytoplasm of the surrounding cells or by a higher expression level of the H^+,K^+-ATPase. If the lysines of the N-terminus play a functional role in ion pumping, it is possible, therefore, that their frequencies could vary between the H^+,K^+-ATPases of carnivores and herbivores. Significant differences in the lysine frequency of the H^+,K^+-ATPase N-terminus were indeed found between different vertebrate classes, but the differences did not correlate with the animal's diet.

Methods

H^+,K^+-ATPase and Na^+,K^+-ATPase amino acid sequence analysis

Sequences of the main catalytic α_1 subunit of the gastric H^+,K^+-ATPase and the Na^+,K^+-ATPase were obtained from the protein database of the National Center for Biotechnology Information. (https://www.ncbi.nlm.nih.gov/protein/). All available entire vertebrate sequences were aligned using the MUSCLE program (Edgar 2004) within the MEGA7 suite of evolutionary genetics programs (Kumar et al. 2016).

Lysine frequency analysis

The lysine frequencies of different domains of the Na^+,K^+- and H^+,K^+-ATPases were determined either using the ProtParam tool (https://web.expasy.org/protparam/) within the ExPASy server of the Swiss Institute of Bioinformatics (Gasteiger et al. 2005) or within the MEGA7 suite of evolutionary genetics programs (Kumar et al. 2016). The ProtParam tool was also used to separately identify the extramembrane and transmembrane domains of both enzymes by entering the Swiss-Prot/TrEMBL accession numbers of the two enzymes

(P05023 for human Na^+,K^+-ATPase and P20648 for human H^+,K^+-ATPase). This allowed the percentages of lysines in extra- and trans-membrane domains to be separately determined.

Phylogenetic analysis

Phylogenetic relationships between either the whole α_1 subunit of the H^+,K^+-ATPase of different vertebrate species or of its N-terminus alone were calculated using the Neighbor-Joining method (Saitou and Nei 1987) within the MEGA7 suite of evolutionary genetics programs (Kumar et al. 2016). Evolutionary distances were computed using the Poisson correction method (Zuckerkandl and Pauling 1965). Each tree was rooted on the branch leading to the sequence of *Dasytis sabina* (Atlantic stingray), the only cartilaginous fish in our analysis. Each species was classified as a carnivore, omnivore or herbivore based on either common knowledge (e.g. koalas are herbivorous) or information freely available on the internet from a variety of sources.

Prediction of disorder

The normalized disorder tendency along the N-terminus of the *Homo sapiens* H^+,K^+-ATPase and Na^+,K^+-ATPase α_1 subunits were predicted using the meta Protein Disorder prediction system server, metaPrDOS (Ishida and Kinoshita 2008) (http://prdos.hgc.jp/cgi-bin/meta/top.cgi), which in our case used the following combination of predictors: PrDOS, DISOPRED2, DISPROT (VSL2P) and IUPred. This combination of predictors is estimated to have a receiver-operator (ROC) score of 0.897, which represents higher prediction accuracy than if any of the predictor servers were used alone. In order to increase the reliability of the prediction and reduce end-effects, the prediction calculation was carried out not just for the region of interest, i.e. the cytoplasmic N-terminal domains, but for the entire N-terminus up to the end of the first transmembrane helix.

Results

H^+,K^+-ATPase and Na^+,K^+-ATPase amino acid sequence analysis

To first determine which amino acid residues in the N-terminus of the gastric H^+,K^+-ATPase α_1 subunit are potentially physiologically relevant we carried out a sequence alignment of all available vertebrate sequences. Alignments of the N-termini are shown in Fig. 1. The numbering is based on the human sequence, with the initiation methionine defined as amino acid number 1 (Maeda et al. 1988). All conserved amino acid residues are highlighted in yellow. The sequence alignment shown in Fig. 1 starts at *Homo sapiens* amino acid residue number 5 simply because the H^+,K^+-ATPase sequences of several species (e.g. *Myotis brandtii* and *Chlorocebus sabaeus* in particular) have long species-specific segments prior to this position which don't match at all well with those of most of the others. Thus, by starting at *Homo sapiens* residue 5 these species-specific segments are excluded. The reason for the unusual sequences of these species is unclear. There is no obvious evolutionary connection between the animals which possess them. An alignment of the N-terminus of the α_1 subunit of the Na^+,K^+-ATPase is shown in Fig. 2.

From the H^+,K^+-ATPase alignment shown in Fig. 1, one region of note is the sequence $^{46}M^{47}K^{48}K^{49}E$, which is conserved in every sequence except that of *Stegastes partitus* (bicolour damselfish), where the second lysine (K) in the motif is substituted by the other basic amino acid residue, arginine (R). The MKKE (or MKRE motif) of the H^+,K^+-ATPase parallels the conserved $^{30}L^{31}K^{32}K^{33}E$ motif of the Na^+,K^+-ATPase, which is found in this enzyme's N-terminus (see Fig. 2). Another similarity between the two enzymes is the high frequency of lysines in the N-terminus. For the Na^+,K^+-ATPase the frequency of lysines was found to be in the range 22.9 – 36.4% up to the LKKE motif. Carrying out the same analysis for the H^+,K^+-ATPase up to and including the MKKE motif (but excluding the

initiation methionine) yields values ranging between 7.8% (*Myotis brandtii*, Brandt's bat) and 40.0% (*Boleophthalmus pectinirostris*, Great blue-spotted mudskipper and *Monopterus albus*, Asian swamp eel). The relatively low value for *Myotis brandtii* is mainly due to its very long N-terminus, i.e., 218 residues up to the MKKE, in comparison to only 35 residues for *Boleophthalmus pectinirostris* and *Monopterus albus*. But even the *Myotis brandtii* enzyme has an accumulation of lysine residues shortly before the MKKE motif.

As in the case of the Na^+,K^+-ATPase, the lysine frequencies of the H^+,K^+-ATPase N-terminus of all species are higher than the 5.7% determined for 1021 unrelated proteins (McCaldon and Argos 1988). This would seem to suggest a functional role of the lysine residues. However, it is important to bear in mind that the value of 5.7% reported by McCaldon and Argos (1988) is based on entire protein sequences, and, therefore, includes many protein domains buried deep in the nonpolar interior of proteins where charged residues such as lysines are unlikely to be present purely on electrostatic grounds. Therefore, to determine whether or not lysine is enriched in the N-termini of the Na^+,K^+- and H^+,K^+-ATPase above what one would expect for protein segments exposed to an aqueous medium it is necessary to consider the transmembrane and extramembrane domains of the proteins separately. Such an analysis is described in the following section.

Apart from the lysine frequency, another point worth noting from the alignment is the large number of glycine residues present in the H^+,K^+-ATPase N-terminal sequence, particularly the cluster between residues 33 – 37, which is specific to mammals. Another feature which is specific to most mammals is the pair of prolines in positions 16 and 18. Serine-27, which, as mentioned in the Introduction, has been identified as a site of phosphorylation by protein kinase C (Togawa et al. 1996), is specific to placental mammals. The possible significance of these residues will be discussed later.

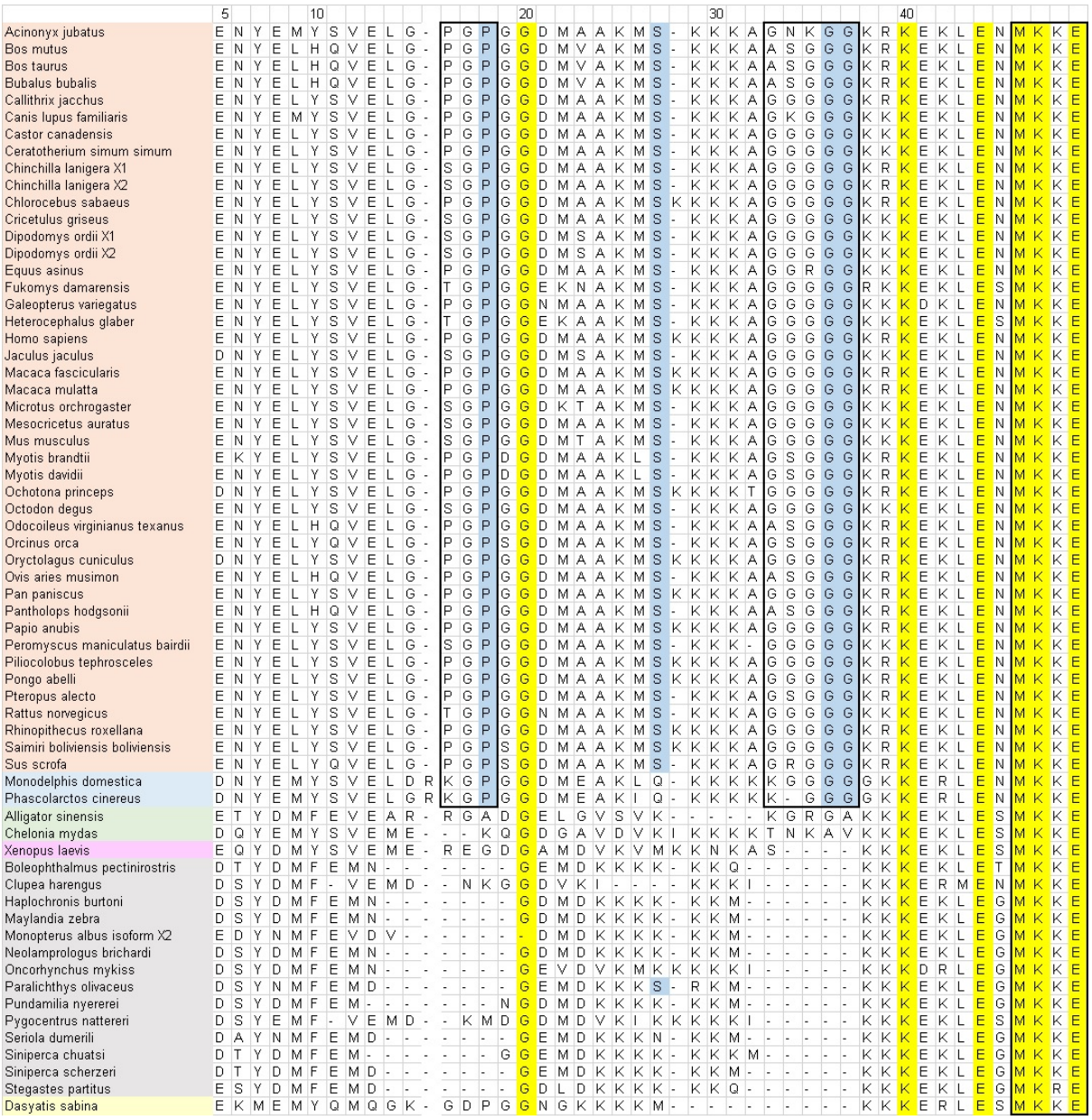

Figure 1: Sequence alignment of the N-terminus of the α_1 isoform of the catalytic α-subunit of the gastric H^+,K^+-ATPase from vertebrates. Residues which are conserved across all species are highlighted in yellow. Serine 27, which is conserved in all placental mammals, is highlighted in light blue, as are glycine 36, glycine 37 and proline 18, which are conserved in all mammals (placental and marsupial). The $^{46}M^{47}K^{48}K^{49}E$ (or $^{46}M^{47}K^{48}R^{49}E$) motif and the glycine cluster and proline locations of all mammals are boxed. The numbering of the

residues is based on the *Homo sapiens* sequence. The species have been grouped according to classes, i.e. placental mammals (orange, top), marsupial mammals (blue), reptiles (green), amphibian (pink), bony fish (grey) and cartilaginous fish (pale yellow, bottom).

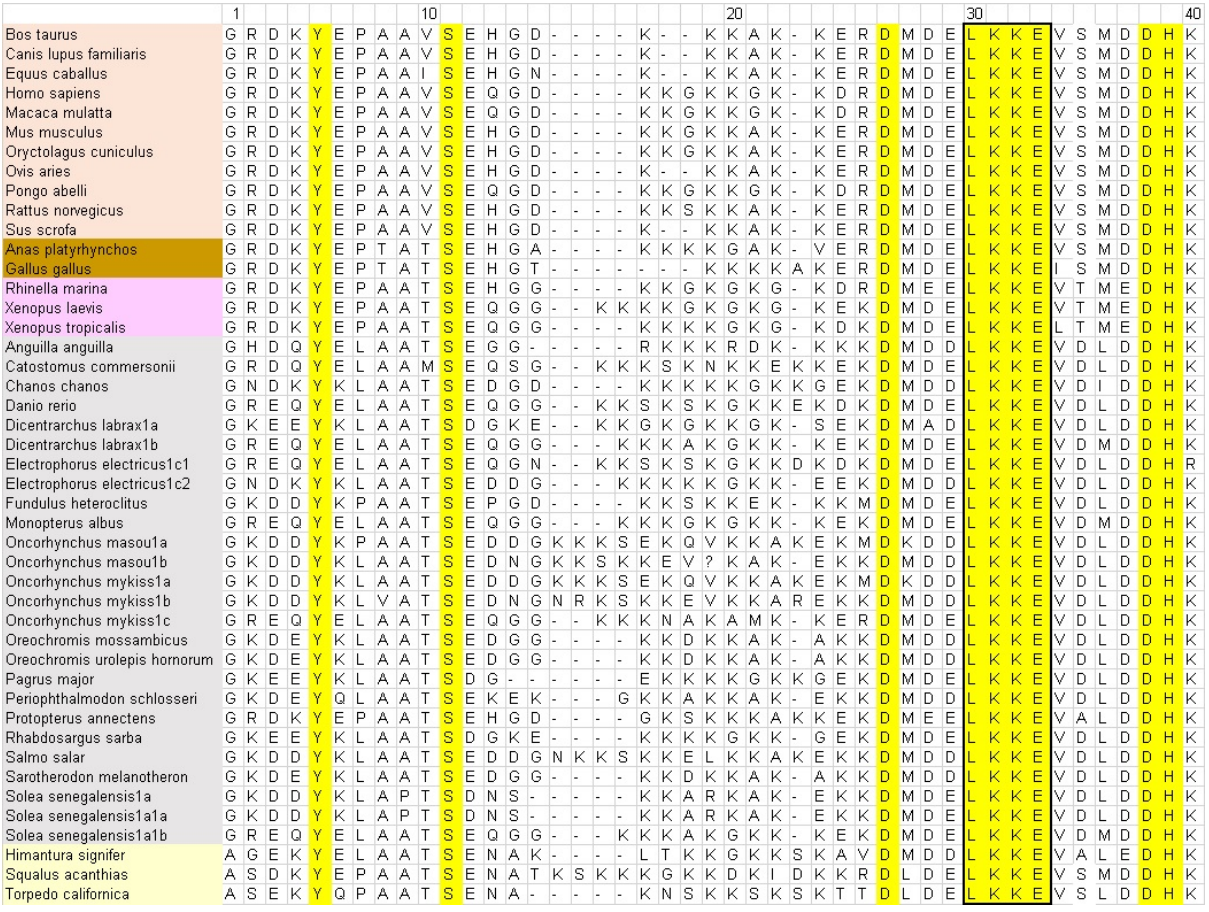

Figure 2: Sequence alignment of the N-terminus of the α_1 isoform of the catalytic α-subunit of the Na^+,K^+-ATPase from vertebrates. Residues which are conserved across all species are highlighted in yellow. The $^{30}L^{31}K^{32}K^{33}E$ motif is boxed. The numbering of the residues is based on the *Homo sapiens* sequence. The species have been grouped according to classes, i.e. placental mammals (orange, top), birds (brown), amphibian (pink), bony fish (grey) and cartilaginous fish (pale yellow, bottom)

Lysine frequencies of the Na^+,K^+- and H^+,K^+-ATPases

The lysine frequencies of the entire α_1 chain of the human Na^+,K^+-ATPase (UniProtKB P05023, Gene ATP1A1) and of the human gastric H^+,K^+-ATPase (UniProtKB P20648, Gene ATP4A) are 5.4% (55 of 1018 residues) and 4.6% (48 of 1034 residues), respectively. The propetide sequence of the Na^+,K^+-ATPase (MGKGV) and the initiation methionine of the H^+,K^+-ATPase were excluded in this calculation. Both of these percentages are below the 5.7% lysine frequency of proteins in general determined by McCaldon and Argos (1988). This is to be expected, because lysines would be energetically unfavoured in the nonpolar transmembrane domains of membrane proteins. More interesting is a separate analysis of the transmembrane and extramembrane domains.

For the human Na^+,K^+-ATPase and H^+,K^+-ATPase the frequency of lysines in transmembrane domains are 2.94% (6 of 204 residues) and 3.43% (7 of 204 residues), respectively. In the case of extramembrane domains, the corresponding values are 6.02% (49 of 814 residues) and 4.94% (41 of 830 residues). Thus, again, as one would expect based on electrostatic considerations alone, the charged lysine residue is more prevalent in extramembrane domains. However, if one considers the N-terminus alone, for the entire N-terminus up to the start of the first transmembrane helix the lysine frequencies of the Na^+,K^+-ATPase and the H^+,K^+-ATPase are 13.41% (11 of 82 residues) and 13.40% (13 of 97 residues). Comparing these values to those of the extramembrane domains as a whole for each of the two enzymes, i.e., 6.02% and 4.94%, it is clear that the lysines are not evenly distributed across the extramembrane domains. They are much more prevalent within the N-terminus. As shown in the last section, if one just considers the first part of the N-terminus up to the MKKE motif of the H^+,K^+-ATPase or the LKKE motif of the Na^+,K^+-ATPase, for

some species the lysine frequency reaches values of 40% and 36% for the H^+,K^+- and the Na^+,K^+-ATPase, respectively.

If the high lysine content of the N-terminus were simply due to the energy stabilisation caused by the interaction of the charged lysine residues with surrounding water dipoles, one would expect a relatively even lysine distribution across the extramembrane domains. The fact that this is not the case, i.e., that lysine is concentrated in the N-termini of both enzymes, supports the hypothesis that it is performing an important functional or regulatory role. This conclusion is further supported by results of Dunker et al. (2002), who found from an analysis of the sequences of 157 intrinsically disordered proteins a lysine frequency in the range 7-10%, still significantly lower than the frequencies of the Na^+,K^+- and H^+,K^+-ATPase N-termini.

Disorder profiles and secondary structure

Interestingly, based on their sequences, much of the N-termini of both the H^+,K^+-ATPase and the Na^+,K^+-ATPase are themselves predicted to be intrinsically disordered domains (see Fig. 3). The metaPrDOS server predicts that 79% of the residues up to and including the $M^{46}K^{47}K^{48}E^{49}$ motif of the H^+,K^+-ATPase have a normalized disorder tendency of 0.5 or greater, with a maximum in disorder tendency at residue 35 in the middle of the cluster of five glycine residues (see Fig. 1). Qualitatively similar disorder tendency profiles were obtained for all other mammalian H^+,K^+-ATPases. In the case of the Na^+,K^+-ATPase, the server predicts a normalized disorder tendency of greater than 0.5 for every residue up to the conserved $^{30}L^{31}K^{32}K^{33}E$ motif. However, in contrast to the H^+,K^+-ATPase, there is no maximum in disorder tendency in the middle of the N-terminus, but rather a more gradual drop as one proceeds along the polypeptide chain. The absence of any significant disorder tendency maximum in the Na^+,K^+-ATPase N-terminus is most likely because it lacks both the

glycine cluster of the H^+,K^+-ATPase as well as the bulky prolines in positions 16 and 18 of the H^+,K^+-ATPase sequence. The dramatic drop in disorder tendency of the Na^+,K^+-ATPase which occurs at around residue 75 coincides with the start of the first transmembrane helix. In the case of the H^+,K^+-ATPase, because of its longer cytoplasmic N-terminus, the first transmembrane helix doesn't start until residue 99. In an analogous fashion to the Na^+,K^+-ATPase, there is a steady drop in disorder tendency just prior to this position.

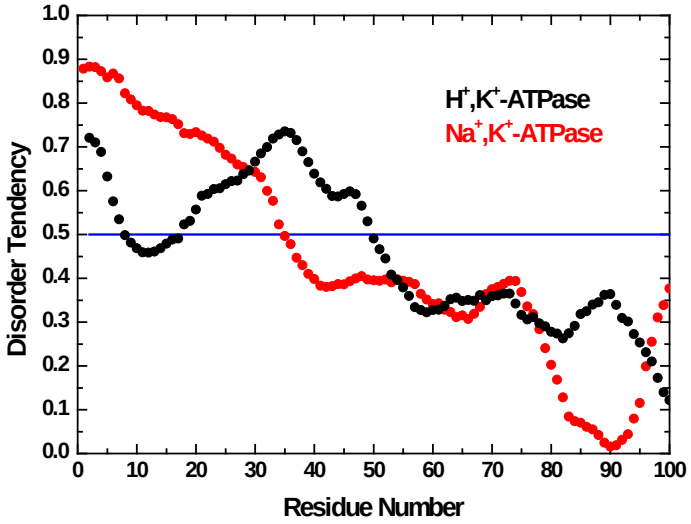

Figure 3: Disorder tendency of the N-terminus of the *Homo sapiens* H^+,K^+-ATPase (black) and Na^+,K^+-ATPase (red) α_1 subunits, as predicted by the metaPrDOS server.

To confirm the origin of the disorder tendency maximum in the mammalian H^+,K^+-ATPase N-terminus sequence, predictions have also been carried out on the sequences of bony fish, which lack both the glycine cluster and the proline residues (see Fig. 1). Qualitatively similar results were obtained for all fish sequences investigated. As an example, Fig. 4 shows a comparison of the results obtained for the human sequence and that of *Siniperca scherzeri* (golden mandarin fish). In contrast to *Homo sapiens*, the *Siniperca scherzeri* sequence shows a disorder tendency profile with no pronounced maximum in the

30 – 40 residue range. In fact, the *Siniperca scherzeri* profile is not dissimilar to that of the *Homo sapiens* Na^+,K^+-ATPase sequence, showing a more gradual drop in disorder tendency along the peptide chain. Therefore, loss of both the cluster of disorder-promoting flexible glycine residues and the disorder-inhibiting bulky proline residues largely abolishes the peak in disorder tendency of the mammalian H^+,K^+-ATPase N-terminus.

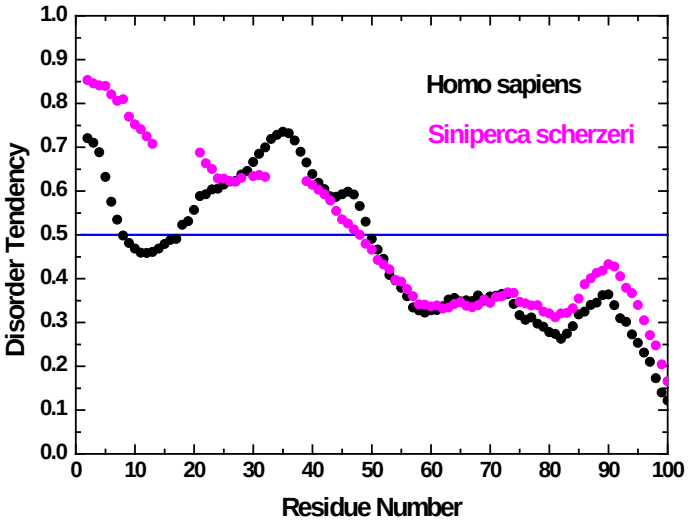

Figure 4: Disorder tendency of the N-terminus of H^+,K^+-ATPase α_1 subunits of *Homo sapiens* (black) and *Siniperca scherzeri* (magenta), as predicted by the metaPrDOS server. The amino acid residue numbering is that of the *Homo sapiens* sequence, with which the *Siniperca scherzeri* sequence has been aligned.

For a comparison with the disorder profiles of the N-terminus of the *Homo sapiens* and *Siniperca scherzeri* H^+,K^+-ATPases we have also carried out an *ab initio* predictions of their secondary structure using the QUARK server (http://zhanglab.ccmb.med.umich.edu/QUARK) (Xu and Zhang 2012). The predicted secondary structures of the entire N-termini up to the start of the first transmembrane domain are shown in Fig. 5. The N-terminus of the *Homo sapiens* enzyme is predicted to have 3 α-

helical segments between residues 39-49, 58-65 and 75-84. The first helix, 39-49, contains the conserved MKKE (or MKRE) motif. However, in agreement with the disorder calculations, the majority of the N-terminus (64% of the residues) is predicted to be in a random coil state. This includes the glycine cluster of the *Homo sapiens* sequence. Previously published secondary structure predictions for the N-terminus of the Na^+,K^+-ATPase α_1 subunit from rat have yielded similar results, i.e., 3 α-helical segments, with the first containing the conserved LKKE motif (Segall et al. 2002).

In comparison, the N-terminus of vertebrates other than mammals is predicted to contain a greater proportion of α-helix than that of mammals. As an example, the predicted secondary structures of the N-termini of the *Homo sapiens* and *Siniperca scherzeri* sequences are shown in Fig. 5. In particular it can be seen that the N-terminus of *Siniperca scherzeri* is predicted to have an extended region of α-helix between residues 22 and 40, which is missing in the *Homo sapiens* sequence. The reason for this is most likely the presence of the glycine cluster in the mammalian sequence in positions 33-37 which may act as a "secondary-structure-breaker". If the length of the *Homo sapiens* N-terminus input into the QUARK server is varied, slightly different lengths and positions of the α-helical segments are predicted, but the helix-breaking effect of the glycine cluster is always apparent.

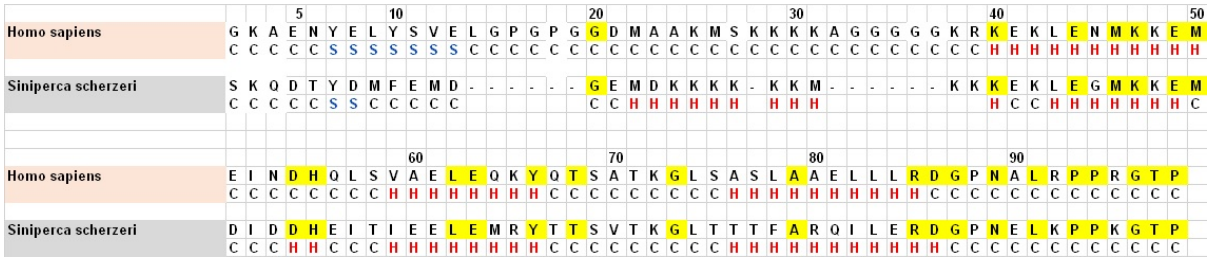

Figure 5: Predicted secondary structures of the N-terminus of the H^+,K^+-ATPase α_1 subunit of *Homo sapiens* and *Siniperca scherzeri*. The top line gives the *Homo sapiens* sequence

numbering. The second line shows the primary sequence, with conserved amino acid residues highlighted in yellow. The third line gives the predicted secondary structure (C = random coil, S = β-strand, H = α-helix).

Dependence of H^+,K^+-ATPase N-terminal lysine frequency on diet and animal class

The animals for which H^+,K^+-ATPase sequence data is available (see Fig. 1) were first grouped according to their diet as carnivores, omnivores and herbivores, and the lysine frequency of each group was averaged. The results obtained were: carnivores (N = 15) 29 (± 9)%, omnivores (N = 24) 25 (± 6)% and herbivores (N = 25) 23 (± 5)%. Although there appears to be a slightly higher preference for lysine in the order carnivores > omnivores > herbivores, the standard deviations are all overlapping. It is also important to note that the carnivore group contains a large number of fish (9 of 15 species). If the fish are removed from the carnivore group, the lysine frequency drops to 19 (± 6) %, i.e., even below that of the herbivore group. Thus, it appears doubtful that there is a link between animal diet and N-terminal lysine frequency.

To minimise the impact of species sampling on the analysis, we have also considered sister pairs, i.e., species which have a close phylogenetic relationship, but different diets. We selected the sister pairs based on the phylogenetic relationships of the species, as determined by a phylogenetic tree constructed using the entire sequence of the H^+,K^+-ATPase (see Fig. 5). Only pairs consisting of one carnivore and one herbivore were chosen in order to maximise the difference in the type of diet. In addition, all pairs were phylogenetically independent of one another, i.e., there was no overlap between the pairs based on the positions of the species on the tree (see Fig. 1). The pairs and their N-terminal lysine frequencies are given in Table 1. Unfortunately the number of available pairs precludes a statistical analysis. However, the available data show that for two of the sister pairs there is

absolutely no difference in lysine frequency between the carnivore and herbivore of the sister pair. For two pairs the carnivore's lysine frequency is higher than that of the herbivore, but there is also a pair in which the herbivore's lysine frequency is higher than that of the carnivore's. Thus, there is no convincing evidence for the suggestion that a higher lysine frequency of the N-terminus is advantageous for the digestion of meat-eating animals.

Sister pairs	% Lys	Sign (C – H)
Orcinus orca (C)	21.28	0
Pantholops hodgsonii (H)	21.28	
Canis lupus familiaris (C)	25.53	+
Pteropus alecto (H)	21.28	
Alligator sinensis (C)	20.93	–
Phascolarctos cinereus (H)	25.53	
Siniperca chuatsi (C)	38.89	+
Stegastes partitus (H)	37.14	
Monopterus albus (C)	40.00	0
Boleophthalmus pectinirostris (H)	40.00	

Table 1: Analysis of the lysine frequency (% Lys) in the N-terminus of the α_1 subunit of the H^+,K^+-ATPase of sister pair species consisting of a carnivore (C) and a herbivore (H). In the third column of the table the sign of the difference between the carnivore's lysine frequency and that of the herbivore is given.

To further investigate any possibility of a diet-related link we carried out a phylogenetic analysis of the sequences of the entire α_1 subunit of the H^+,K^+-ATPase and its N-terminus (see Figs. 6 and 7, respectively).

Figure 6: Phylogenetic tree of the α_1 isoform of the catalytic α-subunit of the gastric H^+,K^+-ATPase from vertebrates. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) is shown next to the branches (Felsenstein 1985).

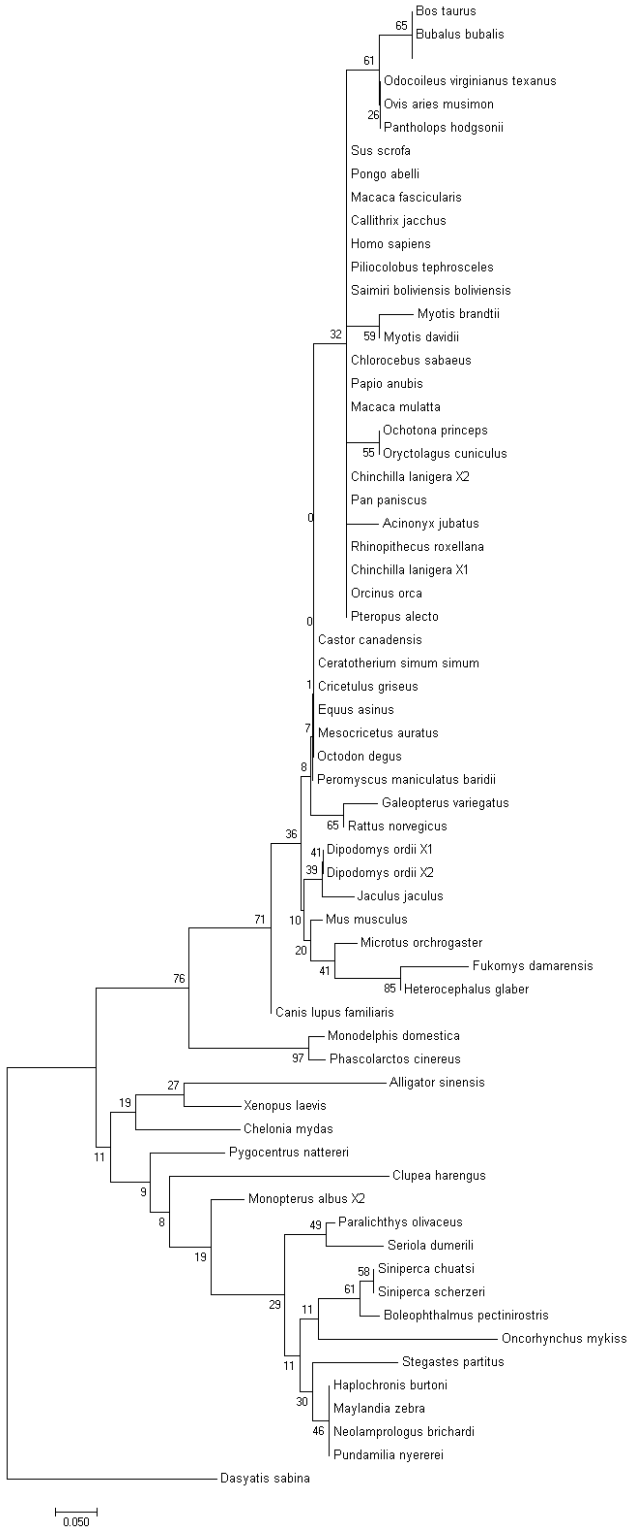

Figure 7: Phylogenetic tree of the N-terminus of the α_1 isoform of the catalytic α-subunit of the gastric H^+,K^+-ATPase from vertebrates. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) is shown next to the branches (Felsenstein 1985).

Although the trees shown in Figs. 6 and 7 are not identical, the important point is that in both trees animals cluster according to their accepted evolutionary positions, i.e., placental mammals cluster with other placental mammals, marsupials with marsupials, fish with other fish, regardless of their diet. In comparison with the whole protein, there is no significant reorganisation of the N-terminus tree to suggest that the N-terminus has a major role in facilitating an animal's digestion, i.e., there is no evidence for convergent evolution of a high N-terminal lysine frequency of species with different origins for any diet-related purpose. This doesn't mean, however, that the N-terminus has no role in determining the proton pumping activity of the H^+,K^+-ATPase. It simply means that factors other than the amino acid sequence of the H^+,K^+-ATPase N-terminus must be playing a more dominant role in controlling stomach pH. The most obvious factor would be expression level. If carnivores expressed a greater number of copies of the H^+,K^+-ATPase in their parietal cells than herbivores and omnivores, this could lead to a lower stomach pH even in the absence of any differences in amino acid sequence of the protein.

After establishing that there is no link between diet and the N-terminus sequence, we turned our attention to an investigation of whether the lysine content of the N-terminus varies with animal class. The results obtained are shown in Table 2. Because the number of reptiles (2) and amphibians (1) in the data set are so low, standard deviations of the values tabulated were only calculated for fish (both bony and cartilaginous) and mammals (placental and marsupial).

Animal Class	Av. % Lys	No. Lys	No. aa
Fish (*N* = 15)	36 (±4)	13 (±1)	36 (±3)
Amphibian (*N* = 1)	27	12	44
Reptiles (*N* = 2)	26	11.5	45
Mammals (*N* = 46)	22 (±4)	11 (±2)	54 (±26)
Mammals (*N* = 45) (exc. *Myotis brandtii*)	22 (±3)	10 (±2)	50 (±8)

Table 2: Values of the average % of lysines (Av. % Lys), the number of lysine residues (No. Lys) and the total number of amino acids (No. aa) in the N-terminus of the α_1 subunit of the H^+,K^+-ATPase for fish and mammals. The error values given in brackets are the standard deviations from the means. The fifth row shows the mammalian results obtained if the sequence of *Myotis brandtii* is excluded from the analysis.

The results obtained clearly show that the lysine frequency in the N-terminus of fish H^+,K^+-ATPases is significantly greater than that of mammals. Because the sequence of Brandt's bat, *Myotis brandtii*, is so unusual, with an N-terminus over a 100 residues longer than all other mammalian species, we have also included averaged values for an analysis excluding this species. However, even excluding *Myotis brandtii*, the lysine frequency is much greater for fish than mammals. As can be seen from Table 2, there are two reasons for the higher lysine frequency of fish. One is that the number of lysine residues is slightly higher for fish than for mammals. However, the more dominant cause is that the length of the N-terminus is shorter for fish than for mammals. In the course of evolution it appears, therefore,

that there has been a tendency towards a slight reduction in the number of lysine residues in the N-terminus and an increase in its length.

As already briefly mentioned earlier, other clear differences between the H^+,K^+-ATPase N-terminus of fish and mammals can be clearly seen by inspection of the alignment shown in Fig. 1. Togawa et al. (1996) showed that serine-27 can be phosphorylated by protein kinase C in the pig. However, Fig. 1 shows that this residue is only conserved as a serine among placental animals. In almost every bony fish this position is occupied instead by a lysine. Only in *Paralichthys olivaeus* (Olive flounder) is there a serine in this position, like the placental mammals. In the bony fish *Seriola dumerili* (Greater amberjack) the position is taken by asparagine. In the only cartilaginous fish in the data set (*Dasytis sabina*, Atlantic stingray) there is a gap in this position. In the only amphibian in the data set (*Xenopus laevis*) its position is occupied by a methionine. In both reptiles in the data set, like the bony fish, the position is occupied by a lysine, and in both marsupial mammals the position is occupied by glutamine. Therefore, it is only the N-terminus of placental mammals which provides a conserved substrate for protein kinase C. It is known that basic residues surrounding serine residues provide a consensus sequence for phosphorylation by protein kinase C (Kennelly and Krebs 1991; Zhu et al. 2005). However, this cannot be the primary purpose of the clustering of lysine residues, because, as explained above, lysine clustering still occurs in fish even when phosphorylation by protein kinase C is impossible due to the absence of a serine residue in the sequence.

The N-terminus of the Na^+,K^+-ATPase also possesses at least one serine residue, i.e., Ser-11 (or Ser-16 if one includes the propetide sequence), which is a potential target for protein kinase C. Even though it is not within a typical consensus sequence for protein kinase C, Beguin et al. (1994) showed, both *in vitro* and in intact cells, that this residue can be phosphorylated by protein kinase C. In contrast to Ser-27 of the H^+,K^+-ATPase, however,

Ser-11 of the Na^+,K^+-ATPase is not present almost exclusively in placental mammals; it is conserved across all vertebrate species (see Fig. 2). Another possible site of phosphorylation by protein kinase C in the Na^+,K^+-ATPase is in position 18 (23 if the propeptide sequence is included). This site is occupied by a serine in the rat (*Rattus norvegicus*), but in most other species it is occupied by a lysine or a glycine residue. Therefore, although phosphorylation of the serine in this position by protein kinase C has been demonstrated in the rat (Feschenko and Sweader, 1995; Logvinenko et al. 1996), this cannot represent part of a general regulatory mechanism of the Na^+,K^+-ATPase across all vertebrate species.

Discussion

The evolutionary analysis carried out here on the H^+,K^+-ATPase has revealed a number of important points. Firstly, the lysine frequency of the H^+,K^+-ATPase N-terminus is not linked to animal diet and hence not linked to stomach pH. This could be explained in two possible ways. It could mean that the lysines of the N-terminus do not play any role in determining the molecular activity of the H^+,K^+-ATPase under physiological conditions. This is possible, but it would appear to be inconsistent with direct experimental data on both the H^+,K^+-ATPase and the Na^+,K^+-ATPase, which support a functional role for the N-terminus (Cornelius and Mahmmoud 2003; Beguin et al. 1994; Feschenko and Sweadner 1995; Logvinenko et al. 1996). A more likely explanation, therefore, is that stomach pH is determined primarily by the expression level of H^+,K^+-ATPase in an animal's stomach parietal cells, rather than by the molecular activity of the individual H^+,K^+-ATPase molecules.

A second finding is that the length of the H^+,K^+-ATPase N-terminus (i.e., the length of the sequence up to the MKKE motif) appears to have increased in the course of evolution on going from fish to higher vertebrates, i.e., mammals. It is possible that this change may have occurred by chance and simply have been inherited by mammals because they have a

common ancestor. However, it is also possible that that the increased N-terminal length may have given mammals some selective advantage, which would be an argument in support of an important role of the N-terminus in H^+,K^+-ATPase function. The lysine frequencies, numbers of lysines and N-terminal length of the one amphibian in the dataset and the two reptiles are intermediate between those of fish and mammals, which would seem to support an adaptive evolutionary process, but with the small number of species this is difficult to conclude with any certainty.

From a comparison of the lysine frequency of the N-terminus of both ATPases with other extramembrane domains and with other disordered proteins, it is clear that the N-terminus lysine frequency is far above what one would expect simply through exposure to the aqueous solvent. This supports the hypothesis that the N-terminus does have a functional role for both enzymes. What this role is, however, is still unclear.

Based on the prediction that the N-terminus is largely disordered (see Figs. 3 and 4), to gain some insight into the role of the N-terminus it is worth considering the roles played by other disordered proteins. If protein domains are highly ordered via hydrogen bonding to themselves within α-helical or β-pleated sheet secondary structure it makes it harder for them to interact with other molecules, because this would involve the breaking of cooperative hydrogen bonding networks. In contrast, disordered sequences have much greater flexibility and are able to more easily interact with other molecules, undergoing transitions between disorder and order when they do so. Thus, intrinsically disordered proteins have been implicated in many regulatory and signalling processes (Wright and Dyson 2002; Dyson and Wright 2002; Burger et al. 2014). Instrinsic disorder has also been suggested as an important facilitator of regulatory protein phosphorylation by protein kinases, because the flexibility of disordered sequences allows them to interact more easily with their respective partner kinase enzyme (Iakoucheva et al. 2004). As mentioned in the Introduction, phosphorylation of the

N-terminus of both the H^+,K^+-ATPase and Na^+,K^+-ATPase by protein kinase C has been demonstrated experimentally (Beguin et al. 1994; Feschenko and Sweadner 1995; Logvinenko et al. 1996; Togawa et al., 1996; Cornelius and Mahmmoud, 2003). In the case of the H^+,K^+-ATPase the phosphorylation site has been identified as serine-27 (Togawa et al. 1996). However, this residue is only conserved in placental mammals and in almost all fish there is a lysine in this position (see Fig. 1). Thus, regulation of H^+,K^+-ATPase activity by serine phosphorylation by protein kinase C is only a feasible general regulatory mechanism for placental mammals. Although lysines do promote protein kinase C activity, this cannot be their primary role across all the H^+,K^+-ATPases of all species because of the absence of serine in position 27, not only in most fish, but also in the reptiles, marsupial mammals and amphibian in the data set studied. Since it is known that synthetic peptides with the same sequence as the Na^+,K^+-ATPase N-terminus bind to membranes (Nguyen et al 2018), rather than the lysine-rich N-terminus facilitating interaction with protein kinase C, a more likely binding partner is the surrounding membrane, with the interaction being modulated (either its strength or specificity) by the increase in N-terminal length which occurred in the course of evolution from ancestral fish to mammals.

Apart from promoting the interaction with binding partners, it has recently become recognized that intrinsically disordered regions of proteins can themselves play an important role in the thermodynamics of protein conformational transitions (Wand 2013; Drake and Pettitt 2018). If a protein conformational change occurs which involves a transition of a protein segment, such as the N-terminus, from an ordered to a disordered state, the conformational entropy thus gained must cause a decrease in free energy and hence a stabilisation of the resultant protein conformation. Because the number of accessible configurational states of a protein backbone increases with the backbone length, the degree of conformational stabilisation would also be expected to increase with the length of the protein

segment. Thus, the evolutionary increase in H^+,K^+-ATPase N-terminal length identified here in mammals could have enhanced the entropic contribution of the N-terminus to the protein's conformational energetics. As discussed in the Introduction, there is clear evidence for the Na^+,K^+-ATPase that the N-terminus moves significantly during this protein's E2-E1 conformational transition. If this movement changes the degree of motional freedom of the N-terminus, the resultant change in entropy could significantly influence the thermodynamics of the transition and hence the protein's selectivity for Na^+ and K^+ ions, which are dependent on the enzyme's conformational state. This could be equally true for the H^+/K^+ selectivity of the H^+,K^+-ATPase. By this mechanism protein segments, such as the N-terminus, which are quite distant from the actual ion binding sites themselves could have a major impact on ion pump function.

Acknowledgements

The authors acknowledge helpful discussions with Prof. Simon Ho and Prof. Keith Dunker. R.J.C acknowledges, with gratitude, financial support from the Australian Research Council (Discovery Grants DP-121003548, DP-150101112 and DP170101732).

Compliance with Ethical Standards

Conflict of interest

The authors declare that they have no conflict of interest.

Ethical approval

This article does not contain any studies with human participants or animals performed by any of the authors

References

Abe K, Irie K, Nakanishi H, Suzuki H, Fujiyoshi Y (2018) Crystal structures of the gastric proton pump. Nature 556:214-218

Asano S, Miwa K, Yashiro H, Tabuchi Y, Takeguchi N (2000) Significance of lysine/glycine cluster structure in gastric H^+,K^+-ATPase. Jap J Physiol 50:419-428

Axelsen KB, Palmgren MG (1998) Evolution of substrate specificities in the P-type ATPase superfamily. J Mol Evol 46:84-101

Beasley DE, Koltz AM, Lambert JE, Fierer N, Dunn RR (2015) The evolution of stomach acidity and its relevance to the human microbiome. PLoS ONE 10:e0134116

Beguin P, Beggah AT, Chibalin AV, Burgener-Kairuz P, Jaisser F, Mathews PM, Rossier BC, Cotecchia S, Geering K (1994) Phosphorylation of the Na,K-ATPase α-subunit by protein kinase A and C *in Vitro* and in intact cells. Identification of a novel motif for PKC-mediated phosphorylation. J Biol Chem 269:24437-24445

Burger VM, Gurry T, Stultz CM (2014) Intrinsically disordered proteins: Where computation meets experiment. Polymers 6:2684-2719

Choe KP, Verlander JW, Wingo CS, Evans DH (2004) A putative H^+,K^+-ATPase in the Atlantic stingray, *Dasyatis sabina*: primary sequence and expression in gills. Am J Physiol Regul Integr Comp Physiol 287:R981-R991

Cornelius F, Mahmmoud YA (2003) Direct activation of gastric H,K-ATPase by N-terminal protein kinase C phosphorylation. Comparison of the acute regulation mechanisms of H,K-ATPase and Na,K-ATPase. Biochemistry 84:1690-1700

Cornelius F, Mahmmoud YA, Meischke L, Cramb G (2005) Functional significance of the shark Na,K-ATPase N-terminal domain. Is the structurally variable N-terminus involved in tissue-specific regulation by FXYD proteins? Biochemistry 44:13051-13062

Daly SE, Lane LK, Blostein R (1996) Structure/function analysis of the amino-terminal region of the α1 and α2 subunits of Na,K-ATPase. J Biol Chem 271:23683-23689

Drake JA, Pettitt BM (2018) Thermodynamics of conformational transitions in a disordered protein backbone model. Biophys J 114:2799-2810

Dunker AK, Brown CJ, Obradovic Z (2002) Identification and functions of usefully disordered proteins. Adv Protein Chem 62:25-49

Dyson HJ, Wright PE (2002) Coupling of folding and binding for unstructured proteins. Curr Opin Struct Biol 12:54-60

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792-1797

Else PL, Wu BJ (1999) What role for membranes in determining the higher sodium pump molecular activity of mammals compared to ectotherms? J Comp Physiol B 169:296-302

Felsenstein J (1985) Confidence limits on phylogenetics: An approach using the bootstrap. Evolution 39:783-791

Feschenko MS, Sweadner KJ (1995) Structural basis for species-specific differences in the phosphorylation of Na,K-ATPase by protein kinase C. J Biol Chem 270:14072-14077

Garcia A, Pratap PR, Lüpfert C, Cornelius F, Jacquemin D, Lev B, Allen TW, Clarke RJ (2017) The voltage-sensitive dye RH421 detects a Na^+,K^+-ATPase conformational change at the membrane surface. Biochim Biophys Acta – Biomembr 1859:813-823

Gasteiger E, Hoogland C, Gattiker A, Duvand S, Wilkins MR, Appel RD, Bairoch A (2005) Protein identification and analysis tools on the ExPASy server. In: Walker JM (ed) The proteomics protocols handbook, Humana Press, New York, pp 571-607.

Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z, Dunker AK (2004) The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Res 32:1037-1049

Ishida T, Kinoshita K (2008) Prediction of disordered regions in proteins based on the meta approach. Bioinformatics 24:1344-1348.

Jiang Q, Garcia A, Han M, Cornelius F, Apell H-J, Khandelia H, Clarke RJ (2017) Electrostatic stabilization plays a central role in autoinhibitory regulation of the Na^+,K^+-ATPase. Biophys J 112:288-299

Jørgensen PL (1975) Purification and characterization of (Na^+, K^+)-ATPase. V. Conformational changes in the enzyme transitions between the Na-form and the K-form studied with tryptic digestion as a tool. Biochim Biophys Acta 401:399-415

Jørgensen PL, Skriver E, Hebert H, Maunsbach AB (1982) Structure of the Na,K pump: crystallization of pure membrane-bound Na,K-ATPase and identification of functional domains of the α-subunit. Ann N Y Acad Sci 402:207-225

Jørgensen PL, Collins JH (1986) Tryptic and chymotryptic cleavage sites in sequence of α-subunit of (Na^+ + K^+)-ATPase from outer medulla of mammalian kidney. Biochim. Biophys. Acta 860:570-576

Jørgensen PL, Andersen JP (1988) Structural basis for E1-E2 conformational transitions in Na,K-pump and Ca-pump proteins. J Membr Biol 103:95-120

Kanai R, Ogawa H, Vilsen B, Cornelius F, Toyoshima C (2013) Crystal structure of a Na^+-bound Na^+,K^+-ATPase preceding the E1P state. Nature 502:201-206

Kaplan JH (2002) Biochemistry of Na,K-ATPase. Annu Rev Biochem 71:511-535

Kennelly PJ, Krebs EG (1991) Consensus sequences as substrate specificity determinants for protein kinases and protein phosphatases. J Biol Chem 266:15555-15558

Kumar S, Stecher G, Tamura K (2016) MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol 33:1870-1874

Logvinenko NS, Dulubova I, Fedosova N, Larsson SH, Nairn AC, Esmann M, Greegard P, Aperia A (1996) Phosphorylation by protein kinase C of serine-23 of α-1 subunit of rat Na^+,K^+-ATPase affects its conformational equilibrium. Proc Natl Acad Sci USA 93:9132-9137

Maeda M, Ishizaki J, Futai M (1988) cDNA cloning and sequence determination of pig gastric (H^+ + K^+)-ATPase. Biochem. Biophys. Res. Commun. 157:203-209

McCaldon P, Argos P (1988) Oligopeptide biases in protein sequences and their use in predicting protein coding regions in nucleotide sequences. Proteins 4:99-122

Morth JP, Pedersen BP, Toustrup-Jensen MS, Sørensen TL-M, Petersen J, Andersen JP, Vilsen B, Nissen P (2007) Crystal structure of the sodium-potassium pump. Nature 450:1043-1049

Morth JP, Pedersen BP, Buch-Pedersen MJ, Andersen JP, Vilsen B, Palmgren MG, Nissen P (2011) A structural overview of the plasma membrane Na^+,K^+-ATPase and H^+-ATPase ion pumps. Nat Rev Mol Cell Biol 12:60-70

Nguyen K, Garcia A, Sani M-A, Diaz D, Dubey V, Clayton D, Dal Poggetto G, Cornelius F, Payne RJ, Separovic F, Khandelia H, Clarke RJ (2018) Interaction of N-terminal peptide analogues of the Na^+,K^+-ATPase with membranes. Biochim Biophys Acta – Biomembr 1860:1282-1291

Nyblom M, Poulsen H, Gourdon P, Reinhard L, Andersson M, Lindahl E, Fedosova N, Nissen P (2013) Crystal Structure of Na^+,K^+-ATPase in the Na^+-bound state. Science 342:123-127

Okamura H, Yasuhara JC, Fambrough DM, Takeyasu K (2002) P-type ATPases in Caenorhabditis and Drosophila: Implications for evolution of the P-type ATPase subunit families with special reference to the Na,K-ATPase and H,K-ATPase subgroup. J Membr Biol 191:13-24

Saitou N, Nei M (1987) The neighbour-joining method: A new method for reconstructing phylogenetic trees. Mol Biol Evolution 4:406-425

Segall L, Lane LK, Blostein R (2002) New insights into the role of the N terminus in conformational transitions of the Na,K-ATPase. J Biol Chem 277: 35202-35209

Scanzano R, Segall L, Blostein R (2007) Specific sites in the cytoplasmic N terminus modulate conformational transitions of the Na,K-ATPase. J Biol Chem 282:33691-33697

Shin JM, Munson K, Vagin O, Sachs G (2009) The gastric HK-ATPase: structure, function, and inhibition. Pflugers Arch 457:609-622

Shinoda T, Ogawa H, Cornelius F, Toyoshima C (2009) Crystal structure of the sodium-potassium pump at 2.4 Å resolution. Nature 459: 446-45

Togawa K, Kaya S, Shimada A, Imagawa T, Mårdh S, Corbin J, Kikkawa U, Taniguchi K (1996) Ser-27, Tyr-10, and Tyr-7 in the α-chain as Ca^{2+}-dependent phosphorylatable sites by intrinsic and extrinsic protein kinases. Biochem Biophys Res Commun 227:810-815

Wand AJ (2013) The dark energy of proteins comes to light: conformational entropy and its role in protein function revealed by NMR relaxation. Curr Opin Struct Biol 23:75-81

Wright PE, Dyson HJ (2002) Intrinsically unstructured proteins: reassessing the protein structure-function paradigm. J Mol Biol 293:321-331

Xu D, Zhang Y (2012) *Ab initio* protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins 80: 1715-1735

Zhu G, Liu Y, Shaw S (2005) Protein kinase specificity. A strategic collaboration between peptide specificity and substrate recruitment. Cell Cycle 4:52-56

Zuckerkandl E, Pauling L (1965) Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel HJ (eds) Evolving genes and proteins, Academic Press, New York, pp 97-166.