Harnessing natural diversity to identify key residues in Prolidase

Hanna Marie Schilbert[1,2*],Vanessa Pellegrinelli[1], Sergio Rodriguez-Cuenca[1], Antonio Vidal-Puig[1], Boas Pucker[3,4,5*]

1 Metabolic Research Laboratories, Wellcome Trust MRC Institute of Metabolic Science, Addenbrooke's Hospital, University of Cambridge, Cambridge, UK.

2 Proteome and Metabolome Research, Center for Biotechnology (CeBiTec), Bielefeld University, Universitätsstraße 27, Bielefeld, Germany

3 Genetics and Genomics of Plants, Faculty of Biology, Bielefeld University, Germany

4 Center for Biotechnology (CeBiTec), Bielefeld University, Germany

5 Evolution and Diversity, Department of Plant Sciences, University of Cambridge, UK

* corresponding authors:

HS: hschilbe@cebitec.uni-bielefeld.de

BP: bpucker@cebitec.uni-bielefeld.de

Key words: PEPD, Peptidase D, Xaa-Pro dipeptidase, cancer, natural variation, polymorphism, prolidase deficiency, conservation, phylogeny

Abstract

Prolidase (PEPD) catalyses the cleavage of dipeptides with high affinity for proline at the C-terminal end. This function is required in almost all living organisms and orthologues of PEPD were thus detected across a broad taxonomic range. In order to detect strongly conserved residues in PEPD, we analysed PEPD orthologous sequences identified in data sets of animals, plants, fungi, archaea, and bacteria. Due to conservation over very long evolutionary time, conserved residues are likely to be of functional relevance. Single amino acid mutations in *PEPD* cause an autosomal disorder called prolidase deficiency and were associated with various cancer types. We provide new insights into 15 additional residues with putative roles in prolidase deficiency and cancer. Moreover, our results confirm previous reports identifying five residues involved in the binding of metal cofactors as highly conserved and enable the classification of several non-synonymous single nucleotide polymorphisms as likely pathogenic and seven as putative polymorphisms. Moreover, more than 50 conserved residues across species, which were not previously described, were identified. Conservation degree per residue across the animal kingdom were mapped to the human PEPD 3D structure revealing the strongest conservation close to the active site accompanied with a higher functional implication and pathogenic potential, validating the importance of a characteristic active site fold for prolidase identity.

52 ## Introduction

53 Human peptidase D (PEPD) or prolidase (EC 3.4.13.9) is a multifunctional manganese-requiring
54 homodimeric iminodipeptidase. Its enzymatic activity was reported in 1937 for the first time with the
55 observation of Glycyl-Proline dipeptides degradation [1]. PEPD belongs to the metalloproteinase M24
56 family. Its major function is the hydrolysis of peptide bonds of imidodipeptides with a C-terminal
57 proline or hydroxyproline, thus liberating proline [2].

58 The biological significance of *PEPD* is indicated by the presence in the genomes of most animal species
59 and its expression in several tissues [3–7]. Moreover, *PEPD* has been identified in fungi [8,9], plants
60 [10], archaea [11], and even bacteria [12–15]. Especially the presence of PEPD in several mycoplasma
61 species stresses its essential role in their metabolism and maintaining cellular functions, as these
62 intracellular parasites display an otherwise extremely reduced gene set [16].

63

64 **Physiological role of PEPD**

65 PEPD is the only known metalloenzyme in eukaryotes catalysing the hydrolysis of X-P [17]. Therefore,
66 deleterious mutations in *PEPD* in human lead to a rare autosomal disease called prolidase deficiency
67 (PD), which is characterized by skin ulcerations -due to defective wound healing-, immunodeficiency,
68 mental retardation, splenomegaly, recurrent respiratory infections and imidodipeptiduria [18–20]. To
69 date, 29 different pathogenic variants have been reported and associated with PD, resulting in a partial
70 or complete enzyme inactivation [21]. In addition to this autosomal disease, perturbations in PEPD
71 expression, (serum) activity or serum levels have been associated with several (patho)physiological
72 processes, including remodelling of the extracellular matrix, inflammation, carcinogenesis,
73 angiogenesis, cell migration, and cell differentiation [22–27]. Moreover, alterations of PEPD serum
74 activity are associated with a spectrum of mental diseases, like post-traumatic stress disorder [28] and
75 depression [29].
76 In bacteria and archaea, PEPD is assumed to be involved in the degradation of intracellular proteins
77 and proline recycling [30]. In animals, PEPD is involved in the degradation proline-rich dietary proteins
78 and seems to play an important role in proline recycling [2]. Since collagen (a major components of
79 extracellular matrix) consists of 25% proline and hydroxyproline, PEPD is thought to be the rate limiting
80 step in collagen turnover [2,31]. Interestingly, there is a growing body of evidence showing that PEPD
81 may also have additional pleiotropic effects, independently from its enzymatic activity. Thus, PEPD has

been reported to influence the p53 pathway by direct protein-protein interaction [32] and acts as ligand for EGFR and ErbB2 when released by injured cells [33,34].

Characterization of the enzymatic and structural properties of PEPD

The crystal structure of PEPD has been extensively investigated in several species, including bacteria [16,35], archaea [36], and eukaryotes [17]. PEPD belongs together with methionine aminopeptidase (MetAP; EC 3.4.11.18) and aminopeptidase P (APP; EC 3.4.11.9) to the “pita-bread” family, which is able to hydrolyse amido-, imido-, and amidino-containing bonds [37,38]. Characteristic for this family is the highly conserved characteristic pita-bread fold in the catalytic C-terminal domain including the metal centre and a well-defined substrate binding pocket [37,39]. The catalytic C-terminal domain comprises five highly conserved residues for the binding of the metal cofactors: D276, D287, H370, E412, and E452 (positions refer to human sequence) [17].

The preferable substrate, optimal pH and temperature, and required metal ions (e.g. Mn^{2+}, Zn^{2+} or Co^{2+}) are species-dependent [2]. Although PEPD appears to be a (homo)dimer in most species including humans, it can be also active as a monomer or even as a tetramer in certain species [2]. The homodimeric human PEPD preferably hydrolyses G-P, is adapted to a pH value of 7.8 with a temperature optimum of 50°C, and shows long-term activity at 37°C [17,40]. *In vitro* studies based on recombinant PEPD produced in CHO cell lines and *E. coli* as well as endogenous PEPD of human fibroblasts, revealed G-P as preferred substrate followed by a lower substrate specificity for A-P, M-P, F-P, V-P, and L-P dipeptides [40]. Moreover, in human PEPD the substrate specificity for dipeptides is determined through the presence of specific residues, like R398 and T241, which prevent the binding of longer substrates [17].

Regulation of PEPD

PEPD is a phosphotyrosine and phosphothreonine/serine enzyme [41,42]. Phosphorylation results in an increase of PEPD activity and is mediated by the MAPK pathway and NO/cGMP signalling for tyrosine and threonine/serine residues, respectively [41,42]. Phosphorylation mediated up-regulation of PEPD activity was reported without an increased gene expression, indicating the importance of post-translational modification in its regulation [41,42]. *In silico* analysis of human PEPD indicated post-translational modifications like glycosylations. N-glycosylation was predicted for N13 and N172, while O-glycosylation was thought to effect T458 [22].

113 We anticipate the detailed profiling of conserved residues in PEPD during evolution may help to
114 identify and understand essential components for mentioned PEPD functions and structure. This
115 increased knowledge could help explain the role of PEPD in diseases, especially prolidase deficiency.
116 Taxon-specific conservation of residues provides additional insights e.g. into post-translational
117 modification in eukaryotes. This study identified orthologous sequences of PEPD in peptide sequence
118 sets of several hundred organisms including bacteria, archaea, animal, fungi, and plant species to
119 investigate the conservation of residues in PEPD across the tree of life. We further identified highly
120 conserved residues, which are likely to play key functional roles.

121

122 Results and Discussion

123 Sequence lengths differentiate between high-level taxonomic groups

124 In total, 769 putative PEPD orthologues were identified in animals (440), plants (122), fungi (72),
125 archaea (42), and bacteria (93) (Supplementary File 1). PEPD orthologues in animals revealed an
126 average sequence length of 493 amino acids (aa), while plants and fungi orthologues had an average
127 sequence length of 499 aa and 507 aa, respectively (Supplementary File 2). Compared to these three
128 kingdoms, PEPD sequences of bacteria were slightly smaller with an average sequence length of 455
129 aa. However, PEPD orthologues identified in archaea showed the smallest average sequence length of
130 a kingdom with 360 aa. These findings matched previous reports of 349 aa (*P. furiosus*) and 493 aa
131 (*H. sapiens*) [11,17]. In general, our observations indicate that PEPD sequence length has changed
132 during evolution. This length difference could be due to an increase of complexity and functionality of
133 PEPD in eukaryotes, where it is known as a multifunctional enzyme [2], or due to a loss of domains in
134 prokaryotes. Observing longer version in eukaryotes is not surprising, because eukaryotes are probably
135 more likely to tolerate larger proteins than bacteria due to differences in the relative metabolic burden
136 [43].

137

138 Analysis of previously described residues

139 Our broad taxonomic sampling captured vast natural diversity, which was harnessed to identify highly
140 conserved residues. From conservation of amino acid residues over billions of years during evolution,
141 we infer functional relevance. A huge diversity of different species and thus sequences is key to
142 distinguish relevant residues from the phylogenetic background. To ensure an accurate alignment of
143 all analysed sequences, the alignment was performed with permutations of the input sequences and

144 repeated with different alignment tools. The average difference per position in the resulting
145 alignments is low (Supplementary File 3 and 4).

146

147 **Conservation of functional and structural relevant residues**

148 Highly conserved residues are likely to have a high functional, and/or structural relevance. Aiming to
149 extend the knowledge about the already existing crystallization models of especially human PEPD, we
150 analysed the conservation degree of known residues relevant for the structure and function of PEPD
151 [17]. Despite the high diversity of metal ions accepted by different species [2], the amino acids
152 responsible for the binding of the metal ions (D276, D287, H370, E412, and E452) are highly conserved
153 across species (Supplementary File 5). All residues reported for the interaction with metal ions were
154 detected in over 90% of all sequences. Sequences without these particular residues are likely to be
155 partial and thus not covering this position leading to a lower observed conservation value. When
156 excluding sequence gaps, almost 100% match is reached for all five positions. Based on these results,
157 we conclude that all selected sequences are *bona fide* prolidases. This finding marks the conservation
158 of these five residues as one important structural and functional characteristic of PEPD (Figure 1).

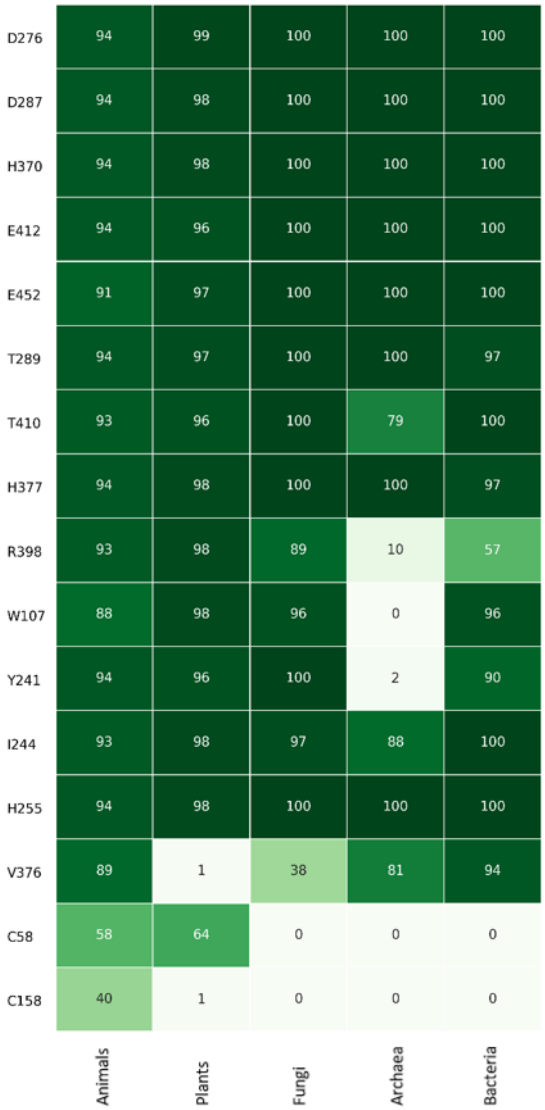


Figure 1: Heatmap of reported functionally important residues of PEPD. The conservation degree of reported residues important for PEPD functionality and structure is displayed in percentage across species. Each column represents a kingdom, while the rows display the analysed residue and its corresponding position in the human PEPD amino acid sequence. A dark green background indicates high conservation, while white means no conservation.

Additionally, strong conservation of T289 and T410 in proximity to the manganese ions supports previous reports and hypotheses of their functional relevance in PEPD [22].

Nevertheless, one plant- and three animal PEPD orthologues showed an amino acid substitution of one metal binding residue: *Ancylostoma ceylanicum* (H370V), *Arachis duranensis* (D287N), *Oncorhynchus kisutch* (E452K) and *Tetraodon nigroviridis* (E452R). Crystal structures and enzyme assays could illuminate the consequences of these substitutions thus providing natural sequences to assess the contribution of each residue. Since D287N was reported before as a probably deleterious substitution [44], these prolidases may have lost their ability to cleave X-P dipeptides.

173 Another essential step for the enzymatic catalysis of prolidases is the binding of their dipeptide
174 substrate (e.g. G-P)[17]. For example, H255 binds to the carboxylate group of the C-terminal proline
175 residue of the substrate and its side chain moves upon substrate binding by about 6 A° narrowing down
176 the size of the active site [17]. The importance of such substrate binding residues, like H255 and H377
177 [17], was validated through a high conservation degree of minimum 94% in all living organisms (Figure
178 1). Interestingly, another residue involved in G-P binding in human PEPD, R398 [17], is highly conserved
179 except in archaea (Figure 1). Besides its role in G-P binding, this residue is also important for the
180 specificity of PEPD for dipeptides by determining the length of the ligand at the C-terminus through its
181 large side chain [16,17]. These results suggest that the majority of analysed archaeal prolidases might
182 not be capable of G-P degradation and may have a broader substrate spectrum due to the missing
183 R398. In line with the hypotheses, Ghosh *et al.* showed that PEPD purified from the archaeon
184 *P. furiosus* revealed no substrate specificity for G-P, but for longer substrates like K-W-A-P and
185 P-P-G-F-S-P, although this specificity was rather weak [11]. However, the preferred substrates of this
186 enzyme were the dipeptides M-P and L-P [11]. Interestingly, *P. furiosus* still has a corresponding
187 arginine residue at the position 295 [16]. This R295 was reported to have dual functionality for cleaving
188 di- and tripeptides due to the intermediate position of this arginine [16]. These reports support the
189 hypothesis that archaeal prolidases have a broader substrate spectrum compared to the prolidases of
190 the other kingdoms. In turn, the strong conservation of R398 in eukaryotes may indicate an adaptation
191 to the specific recognition of dipeptides. In in line with the hypothesis, the bulky side chain of R398
192 was reported to prevent the acceptance of tripeptides [17]. Moreover, a strong conservation of W107,
193 except in archaea, was identified (Figure 1). After G-P binding, W107 is shifted inwards to the active
194 site, sealing the active site [17]. The low conservation of W107 in archaea suggests that archaeal
195 prolidases might use a different conformational change, probably due to their putative expanded
196 substrate spectrum.

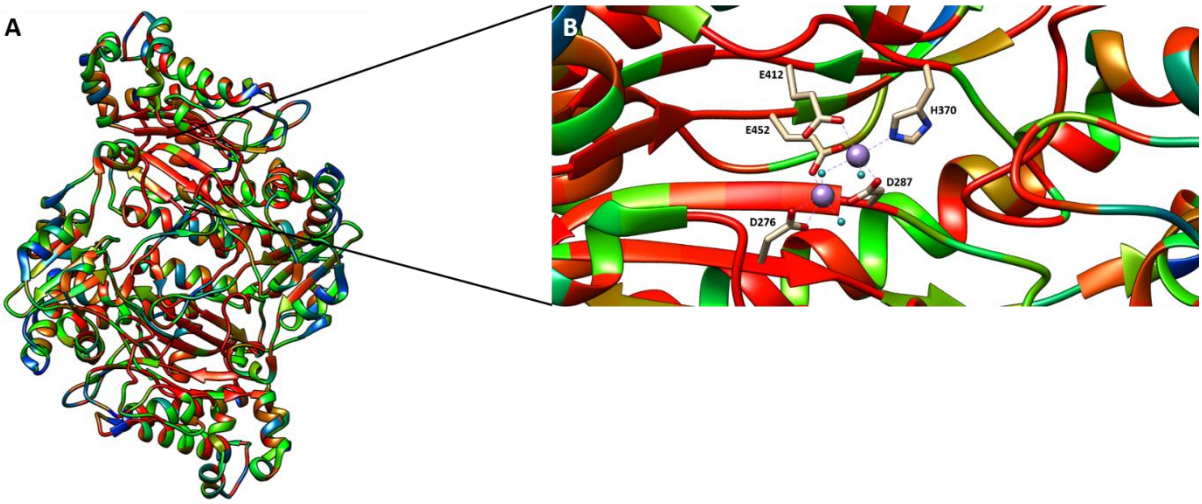
197

198 Furthermore, some residues were reported to be involved in the interaction of L-P, another potential
199 prolidase substrate: Y241, I244, H255, and V376 [17]. H255 and I244 are highly conserved across
200 species (Figure 1). V376 is less conserved in fungi and not conserved in plants. Y241 is not conserved
201 in archaea. Since *P. furiosus* PEPD is capable of binding and degrading L-P, Y241 is probably not
202 essential for this binding process in archaea. Another reason for the flexibility in archaea might be the
203 putatively expanded substrate spectrum due to the absence of Y241, which is reported to close the
204 active site on the side where the N-terminus of the substrate is placed [21]. To the best of our
205 knowledge, the effect of the absence of V376 in plants was not investigated yet.

206 In order to identify a common disulfide bond responsible for the common dimer formation of
207 prolidases previously reported cysteine residues [17] were analysed. In human PEPD an intramolecular
208 disulfide bridge was observed between C58 from chain A and C158 from chain B [17]. However, this
209 bond was only present in the inactive (Mn^{2+} free) enzyme complex, while the substrate was bound in
210 the active site [17]. These amino acids are weakly conserved in the animal kingdom (58% and 40%
211 respectively), but showed an almost complete conservation among vertebrata likely due to their
212 relevance in the dimer formation in this group. However, these cysteines might not be responsible for
213 the dimer formation in the active form of the enzyme, which occurs in most of the prolidases [8,17,45].
214 Therefore, we aimed to identify a better candidate for this common PEPD conformation. However, we
215 could not identify a highly conserved cysteine across species, suggesting (I) the presence of different
216 interactions for stabilization of e.g. PEPD dimers or (II) frequent occurrence of PEPD as a monomer.
217

218 **Analysis of residues known to be mutated in prolidase deficiency**

219 The majority of amino acids that are hot spots causing PD (6/11: D276, G278, L368, E412, G448, G452)
220 are localised near or in the active side of PEPD [22,46]. These amino acids are conserved across species,
221 thus suggesting a negative correlation between the distance of a residue to the active site and its
222 conservation in animals. As expected, highly conserved (>85%) residues are more likely to be located
223 close to the active site (p-value= 3.76e-06, Mann-Whitney U test)(Figure 2, Supplementary File 6).



224

225 **Figure 2: The catalytic cavity is highly conserved in the animal kingdom.** (A) Three dimensional heat map of
226 residue conservation degree in the animal kingdom, displayed represented by the PEPD structure of human
227 prolidase (5M4G). The colour scale ranges from red (highly conserved residues) over orange and green to blue
228 (weakly conserved residues). (B) Conservation degree of the catalytic site of human PEPD. The metal binding
229 residues (D276, D287, H370, E412, and E452) are shown together with the bound Mn^{2+} ions (violet) and water
230 molecules (cyan).

231 As mentioned previously the metal binding residue E452 is highly conserved across species and its
232 deletion results surprisingly in a preservation of the active site [21], likely because it can be replaced
233 by neighbouring residues. However, the mutated protein shows less than 5% of the WT activity [47]
234 supporting our findings. Additionally, our results are in line with findings of Bhatnager and Dang, 2018,
235 who identified the mutation of D276N, G278D, E412K, and G448R as damaging substitutions [44],
236 because we observed a strong conservation of all four residues. Recently the structural basis of these
237 and other PD mutations have been analysed in detail [21]. Once again in accordance with our results,
238 Wilk *et al.* claimed that the D276N mutation results in an excessive reduction of the PEPD activity due
239 to the loss of one of the catalytic metal ions derived from the charge change caused by the substitution
240 [21]. Similarly, in the G278D mutant the loss of one metal ion and additional enhanced disorder were
241 observed [21]. Interestingly, the previously as highly conserved identified Y241 seems to have high
242 functional relevance since its displacement in this mutant results in a destabilization of two metal
243 binding residues (D276 and D287)[21]. In addition, the highly conserved substrate coordinating residue
244 H255 is completely absent from the active site of the G278D mutant [21] stressing its importance in
245 maintaining PEPD functionality. H255 is also absent in the G448R mutant contributing to a
246 dysfunctional protein core [21]. The substitution of the metal binding E412 to K results once again in
247 the loss of one metal ion by an amino acid side chain leading to PEPD inactivation [21].

248 R184 is defined by the shortest atom-to-atom distance to G-P in human PEPD and marks the end of
249 the N-terminal chain of human PEPD [21]. The deletion or mutation of R184 to G in PD patients results
250 in an inactive PEPD or one with highly reduced enzyme activity, respectively [21]. Therefore, R184
251 might be essential for the functionality and structure of PEPD, which is supported by its high
252 conservation across many species [22].In this study, this finding was validated with a minimum
253 conservation degree of 92% of all sequences analysed. Moreover, D375 and D378 were identified as
254 highly conserved across species. Interestingly, these residues were both recently reported to directly
255 interact with R184 [21]. In the PD mutation variant R184G, the interaction between R184 with D375
256 and D378 is lost, due to the replacement of the positive charged guanidinum group of R184 to the
257 neutral amide group of G [21]. The resulting protein shows only residual activity, supporting the
258 hypothesis that D375 and D378 are highly important for PEPD functionality.

259 Additional relevant residues in PD are not particular conserved across different phyla. Among them
260 are S202 (90%) and Y231 (89%) highly conserved in animals. While the deletion of Y231 results in
261 alterations in the dimer interface with remaining PEPD activity, the S202F substitution increases PEPD
262 disorder resulting in the inability to hydrolyse G-P [21]. Y241 is affected by S202F contributing to loss
263 of PEPD activity, since Y becomes disordered even though all other metal binding residue are not

affected [21]. Since Y241 interacts in the WT human PEPD structure with the metal binding aspartates [21], its disorder might result in the loss of this interaction, thus destabilizing PEPD. However, A212 (45%) and R265 (35%) show a substantially smaller conservation degree compared to S202 and Y231. Strong conservation of A212 and R265 is limited to vertebrates thus suggesting a pathogenic role limited to this branch. The phenotype of S202P, A212P, and L368R are not distinguishable from each other, posing an example for relevant residues in PD without strong conservation [46].

Identification of polymorphisms in damage-associated SNPs in human prolidase gene

Recently, Bhatnager and Dang (2018), identified damage associated single-nucleotide polymorphisms (SNPs) in human prolidase gene based on a comprehensive *in silico* analysis [44]. We observed that some of their non-synonymous SNPs are leading to substitutions at variable positions thus qualifying as polymorphisms instead of pathogenic variants. Such a SNP is causing the substitution of V to I at position 305, while our analysis revealed V in 78% and I in 16% of all animal PEPD sequences. Six out of seven tools predicted this SNP as neutral, supporting our assumption [44]. Similar ratios and even dominance of a different amino acid were observed for I45V, E227L, and L435F indicating three additional polymorphisms. Additionally, we hypothesize that nsSNPs leading to T137M, V456M, and D125N are likely to be polymorphisms as the conservation of the canonical amino acid is low.

However, the remaining nsSNPs showing a higher conservation degree in the animal kingdom indicate that they may be important for structure or function of PEPD in the animal kingdom and that substitutions of these residues have a pathogenic potential [44]. This is especially the case for the overlaps of the identified consensus nsSNPs, which were predicted from all tools as damage associated, with our results stressing that these residues are highly conserved not only in the animal kingdom, but also across species [44](Table 1).

Table 1: Conservation degree across species for positions, which were reported to be derived from damage-associated nsSNPs. The conservation degree of positions, which were reported to be derived from damage-associated nsSNPs are stated for animals (An), plants (Pl), fungi (Fu), bacteria (Ba) and archaea (Ar). The first column contains the position of each amino acid based on the human PEPD sequence (Reference sequence position, RSP; UniProt ID: P12955). The amino acid frequency (AAF) ranging from 0 to 1 (1=100% conserved) of the most abundant (1) and second abundant (2) amino acid at a certain position is listed. Gaps in the alignment are indicated through a "-" followed by the conservation degree in the kingdom. Only a "-" is given, when the first amino acid is 100% conserved.

RSP	An AAF1	An AAF2	Pl AAF1	Pl AAF2	Fu AAF1	Fu AAF2	Ba AAF1	Ba AAF2	Ar AAF1	Ar AAF2
19	P_0.71	S_0.18	P_0.73	-_0.1	P_0.97	D_0.01	-_0.62	P_0.26	-_1.0	-
35	R_0.51	K_0.24	R_0.76	-_0.06	L_0.31	R_0.17	P_0.37	A_0.22	E_0.25	Y_0.1

188	T_0.73	S_0.2	S_0.89	T_0.08	D_0.82	T_0.1	D_0.5	T_0.43	D_0.63	E_0.13
192	L_0.67	I_0.25	L_0.8	I_0.14	I_0.64	V_0.19	I_0.38	L_0.34	I_0.5	L_0.38
224	S_0.81	A_0.11	S_0.95	A_0.02	A_0.92	G_0.06	G_0.28	L_0.25	A_0.43	G_0.2
240	S_0.84	A_0.08	S_0.90	-_0.02	A_0.38	G_0.35	P_0.44	G_0.4	S_0.48	A_0.48
247	S_0.79	T_0.13	T_0.89	S_0.07	S_0.74	A_0.21	L_0.41	S_0.24	S_0.45	F_0.38
255	H_0.94	-_0.05	H_0.98	-_0.02	H_1.0	-	H_1.0	-	H_1.0	-
276	D_0.94	-_0.05	D_0.99	-_0.01	D_1.0	-	D_1.0	-	D_1.0	-
278	G_0.94	-_0.06	G_0.99	-_0.01	G_0.97	A_0.03	G_0.99	T_0.01	G_0.95	T_0.05
287	D_0.94	-_0.05	D_0.98	-_0.02	D_1.0	-	D_1.0	-	D_1.0	-
296	G_0.94	-_0.05	G_0.98	-_0.02	G_0.97	T_0.01	G_0.62	S_0.19	G_0.45	-_0.18
373	G_0.94	-_0.06	G_0.98	-_0.02	G_1.0	-	G_1.0	-	G_1.0	-
378	D_0.93	-_0.05	D_0.98	-_0.02	D_1.0	-	D_0.94	E_0.06	E_0.85	D_0.15
403	L_0.80	V_0.12	L_0.96	-_0.02	L_0.94	V_0.04	L_0.78	I_0.15	L_0.85	I_0.13
410	T_0.93	-_0.06	T_0.96	-_0.02	T_1.0	-	T_1.0	-	T_0.78	S_0.23
412	E_0.94	-_0.06	E_0.96	-_0.02	E_1.0	-	E_1.0	-	E_1.0	-
447	G_0.93	-_0.07	G_0.97	-_0.03	G_1.0	-	G_0.53	-_0.4	F_0.6	G_0.25
448	G_0.93	-_0.06	G_0.97	-_0.03	G_1.0	-	G_1.0	-	G_1.0	-

PEPD in cancer

Altered PEPD activity and serum level have been frequently described in different cancer types suggesting an involvement of PEPD in cancer [2,23,24,48]. The investigation of curated SNPs in *PEPD*, which are associated with specific cancer types (BioMuta database [49]), revealed missense mutations in various cancer types to be distributed across the whole PEPD sequence (Supplementary File 7). As many SNPs were associated with a low frequency, we focused on a small set of more frequent ones. Surprisingly, the amino acid affected by the most frequent SNPs in various cancer types is A74, a residue located in the non-catalytic N-terminal domain. While the general frequency in animals is low (38%), it displays a strong conservation in mammals thus suggesting a functional role. Other frequently effected residues are A122, H155, G257, R311, M329, and D378. All of them are conserved to different extents in the animal kingdom, while three (G257, M329, and D378) are also conserved in plants. However, D378 is the only amino acid conserved across all species. Being in proximity to the metal binding residue H370, the high conservation degree of D378 might be due to its role in forming a functional catalytic site. However, we could not identify a “cancer specific hot spot residue” in the animal kingdom and thus the appearance of SNPs in *PEPD* in various cancer types is likely not to be the driving force of a specific cancer type and the identified SNPs might be polymorphisms.

312

313 **Post-translational regulation of PEPD**

314 Since there is experimental evidence of PEPD activity being regulated at the post-translational level
315 through phosphorylation [41,42], we aimed to validate previously predicted post-translational
316 modifications (PTMs) [50] in human PEPD. None of the examined sites were highly conserved across
317 species (Supplementary File 5), which could be explained by differences in the PTM mechanisms
318 between prokaryotes and eukaryotes [51,52]. Nevertheless, some residues were conserved in the
319 animal kingdom e.g. R196 (88%). The low conservation values could be due to differences in PTMs
320 between different groups of eukaryotes [51]. The lack of conservation for some of these residues (S8,
321 K36, S113, T487, A490, K493) could be explained in three ways: (I) no strong functional relevance for
322 PEPD, (II) false positive prediction, or (III) a human specific regulation system. *Vice versa*, three residues
323 are highly conserved at least in the animal kingdom (T15:80%, Y128:78%, R196:88%) posing good
324 candidates for a PTM site. Two of the three amino acids are predicted to be phosphorylated (T15 and
325 Y128), while R196 is thought to be monomethylated [50].

326 Lupi *et al.* predicted putative PTMs at N13, N172 (NetNGly), and T458 (NetOGlyc) [22]. These residues
327 were found to be highly conserved among vertebrates. This situation could be explained by a more
328 recently evolved function or a relaxed ancestral function in species without strong conservation. *In*
329 *silico* prediction of new phosphorylation sites resulted in T90, S113, Y121, Y128, S202, S224, S138,
330 S240, S247 and S460 as best candidates. Conservation degrees generally support these predictions
331 (Supplementary File 5) and distribution across species suggests a more recently increased relevance of
332 S113 and S138.

333

334 **Identification of novel conserved residues**

335 All structure related observation and hypothesis are based on human prolidase crystallization structure
336 (PDB: 5M4G). As we already validated through the correlation in the animal kingdom, highly conserved
337 residues are located nearby or in the substrate binding site. Therefore, it was not surprising that
338 residues near the metal binding residue E452 are highly conserved across species especially R450:92%
339 along with the previously reported G448:93%. The side chain of R450 is near the metal binding site,
340 indicating that it might be essential for the formation of a functional metal ion binding site
341 (Supplementary File 8 (A)). Another two conserved residues, T458 and G461, are located in the curve
342 of a C-terminal loop near the binding site (Supplementary File 8 (B)). The small size of these amino

acids might be necessary to form this structural feature. However, T458 could be a putative phosphorylation site. Since it is located on the outer surface of the enzyme, it is accessible for modifications. Additionally, we observed a cluster of highly conserved residues (G406-V408), which are part of the pita-bread structure, stressing the importance of this fold for the function of PEPD as metalloproteinase.

Again, highly conserved residues across species were identified near another known metal binding residue E412: Y416:94%, P413:94%, and G414:93% are located near the active site and are therefore good candidates for generating a functional binding site. The glycine and proline seem to be important to allow the proper arrangement of the metal binding residues by providing space between them. The side chain of Y416 is pointing into the active side, indicating it might have an additional functional role (Figure 3 (A)).

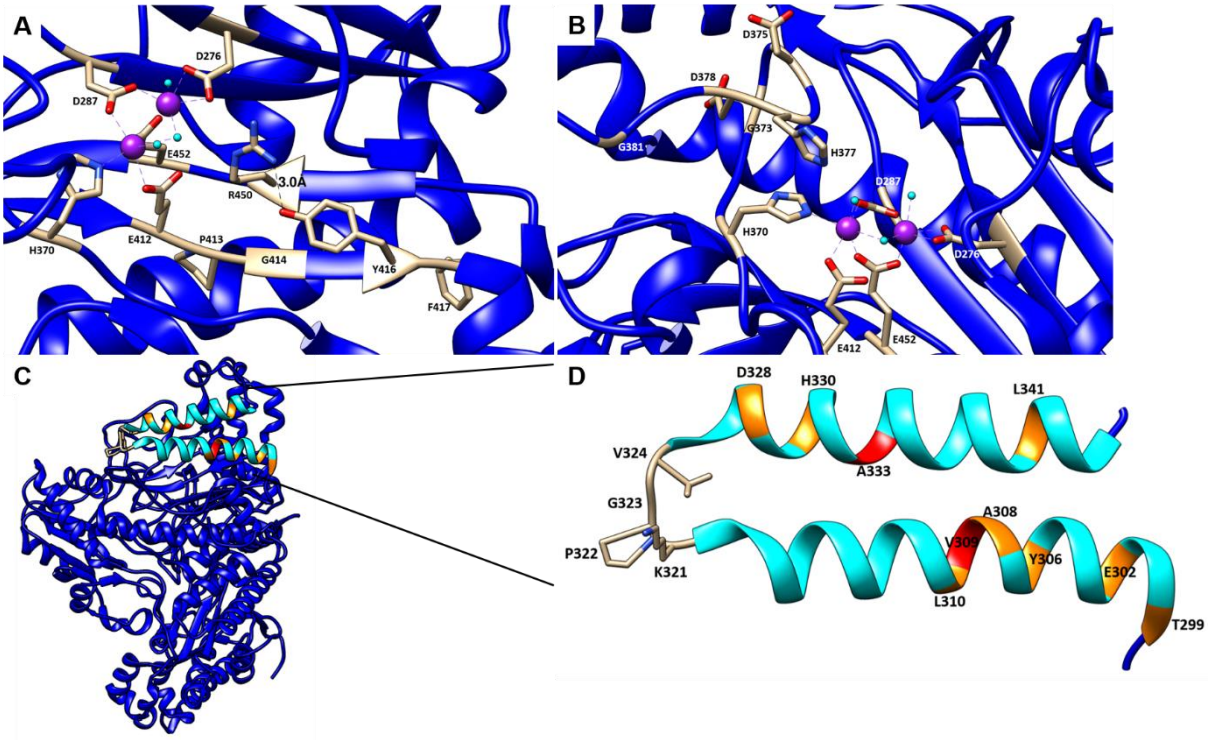


Figure 3: Novel highly conserved residues with functional and/or structural importance in PEPD. The ribbon of the human PEPD 3D model is shown in blue, while residues of interest are lettered. The metal ions are shown in violet and water molecules are shown in cyan. (A) Highly conserved residues P413, G414, and Y416 are located near the metal binding residue E412 and are likely to be involved in generating a functional binding cavity. Y416 might stabilize the anti-parallel β-strand through interaction with R450. (B) G373, D375, D378, and G381 are involved in the stabilization of the loop, which results in an optimal position of the substrate-binding residue H377. (C) Peripheral localisation of the helix with highly conserved residues. (D) The peripheral helix contains two highly conserved residues (A333 and V309), which are marked in red and other conserved residues, which are marked in orange. Moreover, residues building the loop (V324, G323, P322, and K321) are conserved, too.

364 However, it is more likely that it has a stabilizing effect building a hydrogen bond with the NH group of
365 R450:92% (Figure 3 (A)) thus stabilizing the anti-parallel β-strand. This anti-parallel β-strand seems to
366 be highly important for PEPD functionality, since substitutions in the parallel β-strand e.g. G447R or
367 G448R were reported to null PEPD activity [44]. The insertion of a bulky arginine side chain, which
368 prevents the correct assembly of the β-sheet, could be the explanation [44]. Furthermore, F417:82%
369 is highly conserved in every kingdom except archaea, expanding the number of conserved residues in
370 this conserved region (Figure 3 (A)).

371 The conserved G373 is located in a tied turn of the peptide chain, suggesting its interplay with the
372 conserved residues D375, D378, and G381 to form a loop. As a result, the important dipeptide-binding
373 residue H377 is placed near the catalytic site (Figure 3 (B)). Weak conservation of these residues in
374 archaea vindicates the previously mentioned hypothesis that archaea PEPD might be able to hydrolyze
375 a broader substrate spectrum. Additionally, we identified the two conserved residues G369 and H366
376 near the metal binding residue H370 (Supplementary File 8 (C)). The side chain of H366 is pointing into
377 the active site, indicating that it will narrow down the active site, therefore contributing to substrate
378 specificity. Interestingly, residues near H366 e.g. P365, G367, and L368 are highly conserved with
379 exception of the archaea kingdom. This could explain the ability of archaeal prolidases to process
380 tripeptides in addition to dipeptides.

381 The highly conserved residues T299, E302, Y306, A308, V309, L310, K321, P322, G323, V324, D328,
382 H330, and L341 form two parallel helices located in the periphery of PEPD, thus exposed to the solvent
383 (Figure 3 (C)). Based on their extremely high conservation, V309 and A333 are probably most important
384 for this structure (Figure 3 (D)). Whether this region could be the cause for some of extracellular
385 functions of PEPD, e.g. EGFR or ErbB2 binding [33,34] or might be a target for a regulatory protein,
386 needs to be investigated in the future.

387 Moreover, T299, F298, G296 and P293 are highly conserved across species except archaea. These
388 residues might stabilize the pita-bread fold by strengthening a loop near the catalytic site
389 (Supplementary File 8 (D)).

390 Interestingly, Y284 is highly conserved across species, especially in the archaea, bacteria, fungi, and
391 plant kingdom with a minimal conservation of 93%. The conservation degree in animals is only 68%,
392 but the human prolidase contains a F at this position. Across all animals, the conservation of F at this
393 position is 25% ranking it second to Y. Most mammal sequences displayed F at this position, thus
394 indicating (I) a specific function of F in this group or (II) a polymorphism at a permissive site.

395 Additionally, near the metal binding residue D276, some amino acids display strong conservation
396 including G278, G270, E280, and L274.

397 Interestingly, investigation of residues near the highly conserved H255 revealed an exclusive
398 conservation of the region between L257 and A259 in animals and plants. It is located in a loop
399 structure at the periphery of PEPD. This region and other similar observations e.g. G385, V386, M236,
400 G149, N151, T152, Q49, and G50 indicating that plant and animal prolidases might have distinct
401 structural features compared to archaea, bacteria, and fungi. However, the flanking amino acids of
402 H255 are highly conserved at a minimum of 94% in animals, plants and fungi, stressing its importance
403 in eukaryotes.

404 The highly conserved K187:93% separates two helices from each other in human prolidase and might
405 therefore be of structural relevance.

406 Another highly conserved residue is E219:90%, which is likely to stabilize a β-strand from the pita-bread
407 fold, possibly through the interaction with the side chain of another conserved residue N250 or S247
408 (Supplementary File 8 (E)). Moreover, R401 is highly conserved in animal and plant sequences facing
409 the side chain of another conserved residue in the N-terminal region: E182 (Supplementary File 8 (F)).
410 The atom distance between both side chain atoms matches the range of hydrogen bonds with ~2.7 Å
411 to ~3 Å [53]. Thus, both residues could be involved in stabilizing the structure of PEPD.

412 Overall, we observe more conserved residues in the C-terminal catalytic region compared to the N-
413 terminal region. Nevertheless, P98, L95, P80, G76, and F65 are examples for conserved residues in the
414 N-terminal part. Their functions are yet to be determined.

415

416 **Limitations and perspectives**

417 Numerous PEPD orthologues were identified across all living organisms to pinpoint key residues in this
418 protein. The selection of sequences from different groups is not balanced and we do not attempt to
419 assign evolution events to certain groups, which would be possible based on an even more
420 comprehensive sample. A high natural diversity allowed us to distinguish between variable positions
421 with low if any functional relevance and highly conserved residues, which are likely to play key
422 catalytic, structural, or regulatory roles in PEPD. The results match previously reported residues and
423 enabled us to identify additional residues, which should be subjected to in-depth investigation and will
424 eventually shed light on function and structure of PEPD. However, 264 (27%) of the screened data sets
425 did not reveal a PEPD candidate based on our bait sequences. A majority of species without PEPD

426 candidates (175) were bacteria (Supplementary File 9). Since PEPD is a relevant enzyme at least in
427 eukaryotes, it is unlikely to be missing in many species. Technical limitations like incomplete assemblies
428 or annotations could be the reasons for the absence of PEPD from some data sets. Therefore, we
429 checked the completeness of all analysed data sets through the identification of suitable benchmarking
430 genes that are assumed to be present in the respective species (Supplementary File 9) and discussed
431 it in detail (Supplementary File 10). The identification of additional PEPD orthologues would facilitate
432 further analyses e.g. improve the differentiation between pathogenic substitutions and harmless
433 polymorphisms. We used our observations to predict the functional impact of nsSNPs and expect that
434 this approach will be useful in the future for similar applications. We anticipate that the use of *in silico*
435 tools integrating evolutionary genetics and structural data available will help to gain knowledge e.g.
436 regarding the molecular characterization of PEPD, the identification of new regulatory residues, the
437 extracellular role of PEPD, and new therapeutic strategies against prolidase deficiency and other PEPD
438 associated disorders.

439

440 ## Material and methods

441 **Data set collection**

442 The peptide sequence sets of 475 animals, 122 plants, 72 fungi, 49 archaea, and 236 bacteria were
443 retrieved from the NCBI. All sequences were pre-processed with a dedicated Python script to generate
444 customized data files mainly with adjusted sequence names as long sequence names can pose a
445 problem to some alignment tools (https://github.com/bpucker/PEPD). Next, peptide sequence sets
446 were subjected to BUSCO v3 [54] to assess their completeness based on the reference sequence sets
447 'metazoa odb9' (animals), 'embryophyta odb9' and 'eukaryota odb9' (plants), 'eukaryote odb9' (fungi),
448 and 'bacteria odb9' (bacteria). Since there is no dedicated reference sequence set available for
449 archaea, we used the eukaryota and bacteria sets. PEPD bait sequences (Supplementary File 11 and
450 12) were selected manually based on the literature and/or curated UniProt entries [8,36]. Initial
451 selection of related sequences was based on a pipeline combining previously published scripts and
452 using their default parameters [55]. Candidate sequences were identified in a sensitive similarity
453 search by SWIPE v2.0.12 [56] and filtered through iterative steps of phylogenetic analyses involving
454 MAFFT v7.299b [57], phyx [58], and FastTree v2.1.10 [59]. Results were manually inspected and
455 polished to identify *bona fide* orthologous genes with a high confidence. As the average length of PEPD
456 in animals and plants is around 500 amino acids, sequences outside the range 200-700 amino acids

457 were filtered out to avoid bias in downstream analyses through partial sequences or likely annotation
458 artefacts.

459

460 **Identification and investigation of conserved residues**

461 MAFFT v.7.299b [57] was applied for the generation of multiple sequence alignments. Resulting
462 alignments were cleaned by removal of all alignment columns with less than 30% occupancy.
463 Conserved residues were identified and listed based on positions in the human PEPD sequence
464 (UniProt ID: P12955) using the Python script 'conservation_per_pos.py' (Supplementary File 1). This
465 analysis was repeated 50 times with randomly reshuffled sequences as the order of sequences can
466 heavily impact the alignment process [60]. In addition, we compared the alignments generated by
467 MAFFT v.7.299b to ClustalO v.1.2.4 [61] and MUSCLE v.3.8.31 [62] alignments of the same data sets.
468 The alignment bias through the order of input sequences was quantified for all positions of the aligned
469 *Homo sapiens* sequence. For the *in silico* prediction of phosphorylation sites the *H. sapiens* PEPD
470 sequence (UniProt ID: P12955) was submitted to NetPhos 3.1 [63,64]. Only the best prediction for each
471 residue with a high confident score of >0.8 was considered for further analyses.

472

473 **Sources of previously reported data**

474 Previously reported residues with functional implications (Supplementary File 7) were checked for
475 conservation. Additionally, the alignment was screened for highly conserved residues to the best of
476 our knowledge not previously reported in respect to functionality or structure of PEPD. The results of
477 the residue conservation analysis for the animal kingdom were mapped to a 3D structure of human
478 PEPD (PDB: 5G4M). Putative post-translational modification sites were obtained from PhosphoSitePlus
479 and literature [22,50]. Residues associated with PD were retrieved from literature [22,46].
480 Non-synonymous single-nucleotide polymorphisms (nsSNPs) [44] and details about observations were
481 retrieved from the curated BioMuta database [49].

482

483 **Correlation analysis of conservation degree and distance to the active site of PEPD**

484 To determine the conservation degree in correlation to the distance to the active site, the average
485 localisation of the five metal binding residues was identified and used to calculate the distance of each
486 residue to this focus of the catalytic site (Supplementary File 13). Information about the position of

487 each residue was taken from the PDB file 5M4G of human PEPD [17]. The Python modules matplotlib
488 [65] and seaborn (https://github.com/mwaskom/seaborn) were applied to construct a conservation
489 heatmap. In addition, the conservation of all residues in animals was mapped to the 3D model of the
490 human PEPD by assigning colours within a colour gradient to each amino acid representing its
491 conservation among animal sequences.

492

493 **Phylogenetic analysis**

494 A phylogenetic tree was constructed via FastTree v.2.1.10 [59] based on alignments generated via
495 MAFFT v.7.299b [57] and trimmed via pxclsq [58] to a minimal occupancy of 60%. The conservation of
496 different key residues was mapped to this tree for visualization. A Python script
497 (https://github.com/bpucker/PEPD) was deployed to colour all leaves representing sequences with the
498 conserved residue in red.

499

500 **Authors' contributions**

501 HMS and BP designed the experiments, performed bioinformatics analyses, interpreted the results,
502 and wrote the manuscript.

503

504 **Acknowledgements**

505 We thank Samuel F. Brockington, Nathanael Walker-Hale, and Kali Swichtenberg for critical reading of
506 the manuscript and very helpful comments.

507

508 ## References

509 [1] Bergmann M, Fruton J. On proteolytic enzymes. XII. Regarding the specificity of aminopeptidases
510 and carboxypeptidases. A new type of enzyme in the intestinal tract. J Biol Chem 1937;177:189–
511 202.
512 [2] Kitchener R l., Grunden A m. Prolidase function in proline metabolism and its medical and
513 biotechnological applications. J Appl Microbiol 2012;113:233–47. doi:10.1111/j.1365-
514 2672.2012.05310.x.

515 [3] Davis NC, Smith EL. Purification and some properties of prolidase of swine kidney. J Biol Chem
516 1957;224:261–75.
517 [4] Baksi K, Radhakrishnan AN. Purification and properties of prolidase (imidodipeptidase) from
518 monkey small intestine. Indian J Biochem Biophys 1974;11:7–11.
519 [5] Browne P, O'Cuinn G. The purification and characterization of a proline dipeptidase from guinea
520 pig brain. J Biol Chem 1983;258:6147–54.
521 [6] Endo F, Hata A, Indo Y, Motohara K, Matsuda I. Immunochemical analysis of prolidase deficiency
522 and molecular cloning of cDNA for prolidase of human liver. J Inherit Metab Dis 1987;10:305–7.
523 doi:10.1007/BF01800088.
524 [7] Myara I, Cosson C, Moatti N, Lemonnier A. Human kidney prolidase—purification, preincubation
525 properties and immunological reactivity. Int J Biochem 1994;26:207–14. doi:10.1016/0020-
526 711X(94)90147-3.
527 [8] Jalving R, Bron P, Kester HCM, Visser J, Schaap PJ. Cloning of a prolidase gene from Aspergillus
528 nidulans and characterisation of its product. Mol Genet Genomics MGG 2002;267:218–22.
529 doi:10.1007/s00438-002-0655-8.
530 [9] Johnson GL, Brown JL. Partial purification and characterization of two peptidases from
531 Neurospora crassa. Biochim Biophys Acta 1974;370:530–40.
532 [10] Kubota Y, Shoji S, Motohara K. Purification and properties of prolidase for germinating soybeans.
533 Yakugaku Zasshi 1977;97:111–5.
534 [11] Ghosh M, Grunden AM, Dunn DM, Weiss R, Adams MW. Characterization of native and
535 recombinant forms of an unusual cobalt-dependent proline dipeptidase (prolidase) from the
536 hyperthermophilic archaeon Pyrococcus furiosus. J Bacteriol 1998;180:4781–9.
537 [12] Booth M, Jennings PV, Nífhaolain I, O'cuinn G. Endopeptidase activities of Streptococcus
538 cremoris. Biochem Soc Trans 1990;18:339–40. doi:10.1042/bst0180339.
539 [13] Suga K, Kabashima T, Ito K, Tsuru D, Okamura H, Kataoka J, et al. Prolidase from Xanthomonas
540 maltophilia: Purification and Characterization of the Enzyme. Biosci Biotechnol Biochem
541 1995;59:2087–90. doi:10.1271/bbb.59.2087.
542 [14] Mikio F, Yuko N, Shigeyuki I, Toshio S. Purification and Characterization of a Prolidase from
543 Aureobacterium esteraromaticum. Biosci Biotechnol Biochem 1996;60:1118–22.
544 doi:10.1271/bbb.60.1118.
545 [15] Fernández-Esplá MD, Martín-Hernández MC, Fox PF. Purification and characterization of a
546 prolidase from Lactobacillus casei subsp. casei IFPL 731. Appl Environ Microbiol 1997;63:314–6.
547 [16] Weaver J, Watts T, Li P, Rye HS. Structural Basis of Substrate Selectivity of E. coli Prolidase. PLOS
548 ONE 2014;9:e111531. doi:10.1371/journal.pone.0111531.

549 [17] Wilk P, Uehlein M, Kalms J, Dobbek H, Mueller U, Weiss MS. Substrate specificity and reaction
550 mechanism of human prolidase. FEBS J 2017;284:2870–85. doi:10.1111/febs.14158.
551 [18] Lupi A, Rossi A, Campari E, Pecora F, Lund AM, Elcioglu NH, et al. Molecular characterisation of
552 six patients with prolidase deficiency: identification of the first small duplication in the prolidase
553 gene and of a mutation generating symptomatic and asymptomatic outcomes within the same
554 family. J Med Genet 2006;43:e58. doi:10.1136/jmg.2006.043315.
555 [19] Viglio S, Annovazzi L, Conti B, Genta I, Perugini P, Zanone C, et al. The role of emerging techniques
556 in the investigation of prolidase deficiency: from diagnosis to the development of a possible
557 therapeutical approach. J Chromatogr B Analyt Technol Biomed Life Sci 2006;832:1–8.
558 doi:10.1016/j.jchromb.2005.12.049.
559 [20] Phang JM, Liu W, Zabirnyk O. Proline metabolism and microenvironmental stress. Annu Rev Nutr
560 2010;30:441–63. doi:10.1146/annurev.nutr.012809.104638.
561 [21] Wilk P, Uehlein M, Piwowarczyk R, Dobbek H, Mueller U, Weiss MS. Structural Basis for Prolidase
562 Deficiency Disease Mechanisms. FEBS J 2018. doi:10.1111/febs.14620.
563 [22] Lupi A, Tenni R, Rossi A, Cetta G, Forlino A. Human prolidase and prolidase deficiency: an
564 overview on the characterization of the enzyme involved in proline recycling and on the effects
565 of its mutations. Amino Acids 2008;35:739–52. doi:10.1007/s00726-008-0055-4.
566 [23] Gecit İ, Eryılmaz R, Kavak S, Meral İ, Demir H, Pirinççi N, et al. The Prolidase Activity, Oxidative
567 Stress, and Nitric Oxide Levels of Bladder Tissues with or Without Tumor in Patients with Bladder
568 Cancer. J Membr Biol 2017;250:455–9. doi:10.1007/s00232-017-9971-0.
569 [24] Kucukdurmaz F, Efe E, Çelik A, Dagli H, Kılınc M, Resim S. Evaluation of serum prolidase activity
570 and oxidative stress markers in men with BPH and prostate cancer. BMC Urol 2017;17.
571 doi:10.1186/s12894-017-0303-6.
572 [25] Surazynski A, Miltyk W, Palka J, Phang JM. Prolidase-dependent regulation of collagen
573 biosynthesis. Amino Acids 2008;35:731–8. doi:10.1007/s00726-008-0051-8.
574 [26] Uygun Ilikhan S, Bilici M, Sahin H, Demir Akca AS, Can M, Oz II, et al. Assessment of the correlation
575 between serum prolidase and alpha-fetoprotein levels in patients with hepatocellular carcinoma.
576 World J Gastroenterol WJG 2015;21:6999–7007. doi:10.3748/wjg.v21.i22.6999.
577 [27] Pirinççi N, Kaba M, Geçit İ, Güneş M, Yüksel MB, Tanık S, et al. Serum prolidase activity, oxidative
578 stress, and antioxidant enzyme levels in patients with renal cell carcinoma. Toxicol Ind Health
579 2016;32:193–9. doi:10.1177/0748233713498924.
580 [28] Demir S, Bulut M, Atli A, Kaplan İ, Kaya MC, Bez Y, et al. Decreased Prolidase Activity in Patients
581 with Posttraumatic Stress Disorder. Psychiatry Investig 2016;13:420–6.
582 doi:10.4306/pi.2016.13.4.420.

583 [29] Verma AK, Bajpai A, Keshari AK, Srivastava M, Srivastava S, Srivastava R. Association of Major
584 Depression with Serum Prolidase Activity and Oxidative Stress. Br J Med Med Res 2017;20:1–8.
585 [30] Du X, Tove S, Kast-Hutcheson K, Grunden AM. Characterization of the dinuclear metal center of
586 Pyrococcus furiosus prolidase by analysis of targeted mutants. FEBS Lett 2005;579:6140–6.
587 doi:10.1016/j.febslet.2005.09.086.
588 [31] Phang JM, Pandhare J, Liu Y. The Metabolism of Proline as Microenvironmental Stress Substrate.
589 J Nutr 2008;138:2008S-2015S.
590 [32] Yang L, Li Y, Bhattacharya A, Zhang Y. PEPD is a pivotal regulator of p53 tumor suppressor. Nat
591 Commun 2017;8:2052. doi:10.1038/s41467-017-02097-9.
592 [33] Yang L, Li Y, Ding Y, Choi K-S, Kazim AL, Zhang Y. Prolidase directly binds and activates epidermal
593 growth factor receptor and stimulates downstream signaling. J Biol Chem 2013;288:2365–75.
594 doi:10.1074/jbc.M112.429159.
595 [34] Yang L, Li Y, Zhang Y. Identification of prolidase as a high affinity ligand of the ErbB2 receptor and
596 its regulation of ErbB2 signaling and cell growth. Cell Death Dis 2014;5:e1211.
597 doi:10.1038/cddis.2014.187.
598 [35] Are VN, Jamdar SN, Ghosh B, Goyal VD, Kumar A, Neema S, et al. Crystal structure of a novel
599 prolidase from Deinococcus radiodurans identifies new subfamily of bacterial prolidases.
600 Proteins Struct Funct Bioinforma 2017;85:2239–51. doi:10.1002/prot.25389.
601 [36] Maher MJ, Ghosh M, Grunden AM, Menon AL, Adams MWW, Freeman HC, et al. Structure of the
602 Prolidase from Pyrococcus furiosus. Biochemistry 2004;43:2771–83. doi:10.1021/bi0356451.
603 [37] Bazan JF, Weaver LH, Roderick SL, Huber R, Matthews BW. Sequence and structure comparison
604 suggest that methionine aminopeptidase, prolidase, aminopeptidase P, and creatinase share a
605 common fold. Proc Natl Acad Sci U S A 1994;91:2473–7.
606 [38] Lowther WT, Matthews BW. Metalloaminopeptidases: common functional themes in disparate
607 structural surroundings. Chem Rev 2002;102:4581–608.
608 [39] Lowther WT, Matthews BW. Structure and function of the methionine aminopeptidases. Biochim
609 Biophys Acta 2000;1477:157–67.
610 [40] Lupi A, Della Torre S, Campari E, Tenni R, Cetta G, Rossi A, et al. Human recombinant prolidase
611 from eukaryotic and prokaryotic sources. Expression, purification, characterization and long-
612 term stability studies. FEBS J 2006;273:5466–78. doi:10.1111/j.1742-4658.2006.05538.x.
613 [41] Surażyński A, Pałka J, Wołczyński S. Phosphorylation of prolidase increases the enzyme activity.
614 Mol Cell Biochem 2001;220:95–101. doi:10.1023/A:1010849100540.
615 [42] Surazynski A, Liu Y, Miltyk W, Phang JM. Nitric oxide regulates prolidase activity by
616 serine/threonine phosphorylation. J Cell Biochem 2005;96:1086–94. doi:10.1002/jcb.20631.

617 [43] Lynch M, Marinov GK. The bioenergetic costs of a gene. Proc Natl Acad Sci U S A 2015;112:15690–
618 5. doi:10.1073/pnas.1514974112.
619 [44] Bhatnager R, Dang AS. Comprehensive in-silico prediction of damage associated SNPs in Human
620 Prolidase gene. Sci Rep 2018;8:9430. doi:10.1038/s41598-018-27789-0.
621 [45] Yoshimoto T, Matsubara F, Kawano E, Tsuru D. Prolidase from bovine intestine: purification and
622 characterization. J Biochem (Tokyo) 1983;94:1889–96.
623 [46] Falik-Zaccai TC, Khayat M, Luder A, Frenkel P, Magen D, Brik R, et al. A broad spectrum of
624 developmental delay in a large cohort of prolidase deficiency patients demonstrates marked
625 interfamilial and intrafamilial phenotypic variability. Am J Med Genet Part B Neuropsychiatr
626 Genet Off Publ Int Soc Psychiatr Genet 2010;153B:46–56. doi:10.1002/ajmg.b.30945.
627 [47] Ledoux P, Scriver C, Hechtman P. Four novel PEPD alleles causing prolidase deficiency. Am J Hum
628 Genet 1994;54:1014–21.
629 [48] Cechowska-Pasko M, Pałka J, Wojtukiewicz MZ. Enhanced prolidase activity and decreased
630 collagen content in breast cancer tissue. Int J Exp Pathol 2006;87:289–96. doi:10.1111/j.1365-
631 2613.2006.00486.x.
632 [49] Wu T-J, Shamsaddini A, Pan Y, Smith K, Crichton DJ, Simonyan V, et al. A framework for organizing
633 cancer-related variations from existing databases, publications and NGS data using a High-
634 performance Integrated Virtual Environment (HIVE). Database J Biol Databases Curation
635 2014;2014:bau022. doi:10.1093/database/bau022.
636 [50] Hornbeck PV, Zhang B, Murray B, Kornhauser JM, Latham V, Skrzypek E. PhosphoSitePlus, 2014:
637 mutations, PTMs and recalibrations. Nucleic Acids Res 2015;43:D512-520.
638 doi:10.1093/nar/gku1267.
639 [51] Deribe YL, Pawson T, Dikic I. Post-translational modifications in signal integration. Nat Struct Mol
640 Biol 2010;17:666–72. doi:10.1038/nsmb.1842.
641 [52] Nussinov R, Tsai C-J, Xin F, Radivojac P. Allosteric post-translational modification codes. Trends
642 Biochem Sci 2012;37:447–55. doi:10.1016/j.tibs.2012.07.001.
643 [53] Kyte J. Structure in Protein Chemistry. Garland Science; 2018.
644 [54] Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome
645 assembly and annotation completeness with single-copy orthologs. Bioinforma Oxf Engl
646 2015;31:3210–2. doi:10.1093/bioinformatics/btv351.
647 [55] Yang Y, Moore MJ, Brockington SF, Soltis DE, Wong GK-S, Carpenter EJ, et al. Dissecting Molecular
648 Evolution in the Highly Diverse Plant Clade Caryophyllales Using Transcriptome Sequencing. Mol
649 Biol Evol 2015;32:2001–14. doi:10.1093/molbev/msv081.

650 [56] Rognes T. Faster Smith-Waterman database searches with inter-sequence SIMD parallelisation.
651 BMC Bioinformatics 2011;12:221. doi:10.1186/1471-2105-12-221.
652 [57] Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements
653 in Performance and Usability. Mol Biol Evol 2013;30:772–80. doi:10.1093/molbev/mst010.
654 [58] Brown JW, Walker JF, Smith SA. Phyx: phylogenetic tools for unix. Bioinformatics 2017;33:1886–
655 8. doi:10.1093/bioinformatics/btx063.
656 [59] Price MN, Dehal PS, Arkin AP. FastTree: Computing Large Minimum Evolution Trees with Profiles
657 instead of a Distance Matrix. Mol Biol Evol 2009;26:1641–50. doi:10.1093/molbev/msp077.
658 [60] Chatzou M, Floden EW, Di Tommaso P, Gascuel O, Notredame C, Halanych K. Generalized
659 Bootstrap Supports for Phylogenetic Analyses of Protein Sequences Incorporating Alignment
660 Uncertainty. Syst Biol 2018. doi:10.1093/sysbio/syx096.
661 [61] Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-
662 quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 2014;7:539–
663 539. doi:10.1038/msb.2011.75.
664 [62] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic
665 Acids Res 2004;32:1792–7. doi:10.1093/nar/gkh340.
666 [63] Blom N, Gammeltoft S, Brunak S. Sequence and structure-based prediction of eukaryotic protein
667 phosphorylation sites. J Mol Biol 1999;294:1351–62. doi:10.1006/jmbi.1999.3310.
668 [64] Blom N, Sicheritz-Pontén T, Gupta R, Gammeltoft S, Brunak S. Prediction of post-translational
669 glycosylation and phosphorylation of proteins from the amino acid sequence. Proteomics
670 2004;4:1633–49. doi:10.1002/pmic.200300771.
671 [65] Hunter JD. Matplotlib: A 2D Graphics Environment. Comput Sci Eng 2007;9:90–5.
672 doi:10.1109/MCSE.2007.55.
673

674 ## Supplementary material

675 **Supplementary File 1: PEPD peptide sequences used for multiple sequence alignments.**

676

677 **Supplementary File 2: Length distribution of PEPD orthologues.** PEPD sequence length is displayed on
678 the x-axis, while the frequency of a sequence length in percentage is shown on the y-axis. Archaea
679 orthologues are coloured in violet, bacteria in black, fungi in blue, plants in green, and animals in red.

680

681 **Supplementary File 3: Alignment bias control.** The y-axis displays the conservation degree ratio of
682 each residue across species as well as the variation of this value between alignments (Supplementary
683 File 4). The x-axis shows the corresponding residue position in the human PEPD amino acid sequence

684 (UniProt ID: P12955). The green line shows the median of all conservation values observed across all
685 generated alignments. The red line displays the maximum conservation degree and the blue line the
686 minimum conservation degree observed for the respective position across all alignments, respectively.

687

688 **Supplementary File 4: Alignment bias control values.** The variation of the calculated conservation
689 degree based on multiple alignments by MAFFT, ClustalO, and MUSCLE is listed. The first column
690 contains the position in the reference sequence human PEPD (UniProt ID: P12955). In addition, the
691 minimal conservation degree observed over 50 alignments, the median of all these conservation
692 values, and the maximal observed value are provided.

693

694 **Supplementary File 5: Conservation degree of PEPD residues across species.** The conservation degree
695 of each residue, ranging from 0-1.0 (1.0 being perfect conservation) is listed for animals, plants, fungi,
696 bacteria, and archaea. The alignment position of each residue is given in the first column, while the
697 second column refers to the corresponding position in human PEPD (Reference sequence position,
698 UniProt ID: P12955). The amino acid frequency (AAF) of the most abundant (AAF1) and second
699 abundant amino acid (AAF2) at a certain position is given for each species. A gap is indicated by "-".

700

701 **Supplementary File 6: Distance of each residue to the active site of human PEPD.** The distance of
702 each residue to the active site of human PEPD (PDB ID: 5M4G) is stated in arbitrary units.

703

704 **Supplementary File 7: Previously reported residues for conservation analysis.** All previously reported
705 residues with relevance to structure and/or function of PEPD are listed with their associated function
706 and reference. The residue position is derived from human PEPD (UniProt ID: P12955). PTMs identified
707 in *H. sapiens* or *M. musculus* are marked through Hs and Mm in brackets, respectively.

708

709 **Supplementary File 8: Conserved residues in human PEPD 3D model with structural and/or**
710 **functional relevance.** The ribbon of the human 3D PEPD model is shown in blue, while residues of
711 interest are marked in red or alternatively in beige. The metal ions are displayed in violet and water
712 molecules are shown in cyan.

713 (A) R450 (highlighted in red) is located near the metal binding centre.

714 (B) T458 and G461 are marked in red and are located in a peripheral loop.

715 (C) G369 and H366 are located near the metal binding residue H370, where H366 might narrow down
716 the active site. Moreover, P365, G367, and L368 might be involved in substrate specificity of animal,
717 plant, fungi, and bacteria PEPD.

718 (D) T299, F298, G296, and P293 stabilize the pita-bread fold by strengthening the loop near the
719 catalytic site.

720 (E) E219 stabilizes PEPD possibly through the interaction with the side chain of another conserved
721 residue, like N250 or S247.

722 (F) Possible interaction of R401 and E182 through a hydrogen bond, thus stabilizing the structure of
723 PEPD.

Supplementary File 9: BUSCO assessment of peptide data set quality. For each analysed organism presence (+) or absence (-) of PEPD in their peptide dataset is indicated. Completeness of the data sets was assessed based on the detection of BUSCO sequences.

Supplementary File 10: Discussion of possible limitations.

Supplementary File 11: Identifier of bait sequences. Donor species and NCBI or UniProt ID of PEPD bait sequences is listed.

Supplementary File 12: Bait sequences.

Supplementary File 13: Approach for residue distance calculation. Schematic illustration of the approach used to calculate the distances of all amino acids in PEPD to the active site. Different colours indicate different amino acids with different degrees of conservation across species.