

BIOCHEMICAL JOURNAL

ACCEPTED MANUSCRIPT

Assembly of the elongated collagen prolyl 4-hydroxylase $\alpha_2\beta_2$ heterotetramer around a central α_2 dimer

M Kristian Koski, Jothi Anantharajan, Petri Kursula, Prathusha Dhavala, Abhinandan V Murthy, Ulrich Bergmann, Johanna Myllyharju & Rik K Wierenga

C-P4H-I, collagen prolyl 4-hydroxylase isoenzyme I; ER, endoplasmic reticulum; 2OG, 2-oxoglutarate; CAT, catalytic domain of the C-P4H α subunit; PSB, peptide-substrate-binding domain of the C-P4H α subunit; DD, double domain region of the C-P4H α subunit; PDI, protein disulfide isomerase; Cr-P4H-1, *Chlamydomonas reinhardtii* prolyl 4-hydroxylase isoenzyme 1; MS, mass spectrometry; SAXS, small-angle X-ray scattering; eC-P4H-I, human C-P4H-I expressed in *Escherichia coli*. iC-P4H-I, human C-P4H-I expressed in insect cells; yC-P4H-I, human C-P4H-I expressed in yeast cells; α MF, *Saccharomyces cerevisiae* α -mating factor sequence; SEC, size exclusion chromatography; (PPG)_n, synthetic peptide within Pro-Pro-Gly triplets; Pn, synthetic peptide with n proline residues; ESI, electrospray ionization; LC, liquid chromatography; SLS, static light scattering; PDC, pyridine 2,4-dicarboxylate; eC-P4H-I(®(®), complete eC-P4H-I tetramer; eC-P4H-I(®(®'), truncated variant of the eC-P4H-I tetramer; yC-P4H-I(®(®), complete yC-P4H tetramer; yC-P4H-I(®(®'), truncated

Cite as Biochemical Journal (2016) DOI: 10.1042/BCJ20161000

Copyright 2016 The Author(s).

Use of open access articles is permitted based on the terms of the specific Creative Commons Licence under which the article is published. Archiving of non-open access articles is permitted in accordance with the Archiving Policy of Portland Press (<http://www.portlandpresspublishing.com/content/open-access-policy#Archiving>).

Assembly of the elongated collagen prolyl 4-hydroxylase $\alpha_2\beta_2$ heterotetramer around a central α_2 dimer

M Kristian Koski^{*}, Jothi Anantharajan^{*}, Petri Kursula^{*†}, Prathusha Dhavala^{*}, Abhinandan V Murthy^{*}, Ulrich Bergmann^{*}, Johanna Myllyharju[‡] & Rik K Wierenga^{*1}

^{*}Biocenter Oulu and Faculty of Biochemistry and Molecular Medicine, PO Box 5400, University of Oulu, FIN-90014 Oulu, Finland.

[†]Department of Biomedicine, University of Bergen, PO Box 7804, N-5009 Bergen, Norway.

[‡]Oulu Center for Cell-Matrix Research, Biocenter Oulu and Faculty of Biochemistry and Molecular Medicine, PO Box 5400, University of Oulu, FIN-90014 Oulu, Finland.

¹To whom correspondence should be addressed (email rik.wierenga@oulu.fi).

SUMMARY STATEMENT

The SAXS model of human collagen prolyl 4-hydroxylase, a crucial enzyme in collagen biosynthesis, shows that this heterotetrameric enzyme has a bilobal shape. Enzyme kinetic studies provide insight on the topology of proline-rich peptide binding to the competent complex.

KEYWORDS

small-angle X-ray scattering, extracellular matrix proteins, prolyl hydroxylase, multidomain assembly, collagen biosynthesis, 2-oxoglutarate dependent dioxygenase

ABBREVIATION LIST

C-P4H-I, collagen prolyl 4-hydroxylase isoenzyme I; ER, endoplasmic reticulum; 2OG, 2-oxoglutarate; CAT, catalytic domain of the C-P4H α subunit; PSB, peptide-substrate-binding domain of the C-P4H α subunit; DD, double domain region of the C-P4H α subunit; PDI,

protein disulfide isomerase; Cr-P4H-1, *Chlamydomonas reinhardtii* prolyl 4-hydroxylase isoenzyme 1; MS, mass spectrometry; SAXS, small-angle X-ray scattering; eC-P4H-I, human C-P4H-I expressed in *Escherichia coli*. iC-P4H-I, human C-P4H-I expressed in insect cells; yC-P4H-I, human C-P4H-I expressed in yeast cells; α MF, *Saccharomyces cerevisiae* α mating factor sequence; SEC, size exclusion chromatography; (PPG)_n, synthetic peptide with n Pro-Pro-Gly triplets; Pn, synthetic peptide with n proline residues; ESI, electrospray ionization; LC, liquid chromatography; SLS, static light scattering; PDC, pyridine 2,4-dicarboxylate; eC-P4H-I($\beta\alpha\alpha\beta$), complete eC-P4H-I tetramer; eC-P4H-I($\beta\alpha\alpha'$), truncated variant of the eC-P4H-I tetramer; yC-P4H-I($\beta\alpha\alpha\beta$), complete yC-P4H tetramer; yC-P4H-I($\beta\alpha'$), truncated variant of the yC-P4H-I tetramer; R_g, radius of gyration.

ABSTRACT

Collagen prolyl 4-hydroxylase (C-P4H), an $\alpha_2\beta_2$ heterotetramer, is a crucial enzyme for collagen synthesis. The α subunit consists of an N terminal dimerization domain, a central peptide-substrate-binding (PSB) domain, and a C terminal catalytic (CAT)

domain. The β subunit (also known as protein disulfide isomerase (PDI)) acts as a chaperone, stabilizing the functional conformation of C-P4H. C-P4H has been studied for decades, but its structure has remained elusive. Here, we present a three-dimensional small-angle X-ray scattering (SAXS) model of the entire human C-P4H-I heterotetramer. C-P4H is an elongated, bilobal, symmetric molecule with a length of 290 Å. The dimerization domains from the two α subunits form a protein-protein dimer interface, assembled around the central antiparallel coiled-coil interface of their N terminal α helices. This region forms a thin waist in the bilobal tetramer. The two PSB/CAT units, each complexed with a PDI/ β subunit, form two bulky lobes pointing outward from this waist region, such that the PDI/ β subunits locate at the far ends of the $\beta\alpha\alpha\beta$ complex. The PDI/ β subunit interacts extensively with the CAT domain. The asymmetric shape of two truncated C-P4H-I variants, also characterized in this study, agrees with this assembly. Furthermore, data from these truncated variants show that dimerization between the α subunits has an important role in achieving the correct PSB-CAT assembly competent for catalytic activity. Kinetic assays with various proline-rich peptide substrates and inhibitors suggest that in the competent assembly, the PSB domain binds to the procollagen substrate downstream from the CAT domain.

INTRODUCTION

Collagen is the most abundant protein in animals and the major structural protein of the extracellular matrix [1–3]. Collagen prolyl 4-hydroxylases (C-P4Hs, EC 1.14.11.2) are

endoplasmic reticulum (ER) enzymes essential for collagen synthesis [4–6]. In the triple-helical collagen molecule, the individual chains adopt a poly-L-proline type II conformation and assemble into a parallel, right-handed supercoil. Repetitive X-Y-Gly motifs in the collagen chains are required for the formation of this triple helix. Proline is the most frequent amino acid in the X and Y positions [1,3], and the proline residue in the Y position is very often 4-hydroxylated. This modification, catalyzed by C-P4H (Figure 1a), is essential for the thermal stability of the collagen triple helix [3]. C-P4Hs belong to the 2-oxoglutarate (2OG)-dependent dioxygenase superfamily, and in addition to 2OG, their reaction mechanism requires Fe^{2+} and O_2 (Figure 1a) [4-6]. In addition, C-P4H activity requires vitamin C (ascorbic acid), essential for reactivating the inactive Fe^{3+} form of C-P4H into the active Fe^{2+} form, whereas its absence causes the scurvy disease [5,7]. C-P4H activity is also involved in excessive collagen accumulation in fibrotic diseases and severe scarring [5,6,8]. Recently, it has also been reported that C-P4H plays a role in breast cancer metastasis [9,10]. C-P4H is also involved in other cellular processes, such as hydroxylating a specific proline (Pro700) of argonaute-2, which is involved in regulating the RNA interference machinery [11]. C-P4Hs are, thus, attractive targets for pharmacological intervention, the most potent inhibitors currently being bipyridinedicarboxylic acids [12].

Vertebrate C-P4Hs are heterotetramers consisting of two glycosylated α subunits (~65 kDa) and two non-glycosylated β subunits (55 kDa), with a total molecular mass of around 240 kDa [4-6]. Three vertebrate C-P4H isoenzymes, C-P4H I-III, differing only in their respective α subunits, have been characterized [13,14]. The four critical catalytic residues, two histidines, one aspartate and one lysine, are located in the C terminal catalytic (CAT) domain of the α subunit [15-17] (Figure 1b). The -His-X-Asp-.....-His- catalytic triad binds the Fe^{2+} in the active site, whereas the conserved lysine makes a salt bridge with the buried 2OG. These catalytic residues are conserved in all members of the large 2OG-dependent dioxygenase superfamily, in which the active site framework is provided by a characteristic

eight-stranded jelly-roll motif, also known as the double stranded β helix fold or the cupin fold [18,19]. Two other domains have been identified in the C-P4H α subunit, (i) an N terminal dimerization domain (residues 1-143 in the human α (I) subunit) important for the dimerization between the two α subunits [20], and (ii) a middle region containing the peptide-substrate-binding (PSB) domain (residues 144-244 in the human α (I) subunit) [21,22] (Figure 1b). Protein disulfide isomerase (PDI) [23,24] serves as the β subunit of the C-P4Hs. PDI contains four domains, referred to as **a**, **b**, **b'**, and **a'** (Figure 1b). Each domain has the typical thioredoxin fold, but only the **a** and **a'** domains catalyze disulfide bond formation and isomerization *via* their conserved -Cys-Gly-His-Cys- motifs (Figure 1b). The first cysteine is rather solvent exposed, whereas the second cysteine is more buried. In the oxidized state, these two cysteine residues form a disulfide bridge. It appears that the main function of PDI in C-P4H is to facilitate the folding of the α chain, as well as to solubilize the otherwise insoluble α subunit in a catalytically active, non-aggregated state [25].

Although C-P4Hs have been studied since the 1960s, structural information on the complete enzyme tetramer is still limited. The crystal structure of human PDI with all four thioredoxin domains was solved recently [26]. Crystal structures of the dimerization and PSB domains of the α (I) subunit of human C-P4H-I have also been determined [20,22]. The structure of the CAT domain of the α subunit is not available because of its insolubility, but several crystal structures are currently known from other P4H enzymes having a jelly-roll fold [27-30]. One of these is the monomeric algal P4H isoenzyme 1 from *Chlamydomonas reinhardtii* (Cr-P4H-1) [28], a homologue with 26% amino acid sequence identity to the human α (I) CAT domain (Figure 2). Cr-P4H-1 has also been crystallized in the presence of a substrate peptide with Ser-Pro repeats [31]. Crystallographic binding studies with the Cr-P4H-1 [31] and the human C-P4H α (I) PSB domain [20] have revealed that both of these domains bind a 9-residue proline-rich peptide that is folded into a poly-L-proline type II helix.

Here, we present a detailed bioinformatics analysis of the human C-P4H-I α subunit and report a low-resolution small-angle X-ray scattering (SAXS) structure of the complete C-P4H-I heterotetramer in solution. The SAXS analysis shows that the overall shape of C-P4H-I is elongated with bulky lobes at both ends, connected by a thinner waist. The N terminal ends of the dimerized α subunits form the central axis of this enzyme while the two PDI/ β subunits cover the highly insoluble C terminal CAT domains of its two bulky lobes. Enzyme kinetics provide insight concerning the spatial organization of the binding pockets of the PSB and CAT domains in each of the two lobes.

METHODS

Expression and purification of human C-P4H-I

Human C-P4H-I (Uniprot code P13674-1) was expressed in bacteria (*Escherichia coli*, Origami strain) [32], insect cells (*Spodoptera frugiperda*) [33], and yeast cells (*Pichia pastoris*) [34]. C-P4H-I expression, cell lysis, and protein purification followed the protocols described earlier [32-34] with slight modifications. The purified proteins are referred to as eC-P4H-I, iC-P4H-I, and yC-P4H-I, respectively, according to the expression host. In the yeast expression system, the PDI/ β subunit is preceded by the 85-residue N terminal *Saccharomyces cerevisiae* α mating factor sequence (α MF) to facilitate protein translocation in the yeast ER [34,35]. The cells from all three expression systems were lysed in a buffer containing 10 mM Tris-HCl (pH 7.8), 100 mM NaCl, 100 mM glycine, 10 μ M DTT (C-P4H affinity buffer). The soluble fraction of the cell lysates was loaded onto a poly-L-proline (>40000 Da, Sigma) Sepharose matrix immediately after cell lysis. The bound protein was eluted using the C-P4H affinity buffer, supplemented with 3 mg/ml poly-L-proline (1000–10000 Da, Sigma). The fractions containing C-P4H-I were pooled, and the NaCl concentration of the pooled fractions was reduced to 50 mM. The C-P4H-I sample was loaded onto a 5 ml HiTrap Q Sepharose column (GE Healthcare), and the protein was eluted using an

increasing NaCl gradient (200 mM to 500 mM, $20 \times$ column volume). The C-P4H-I-containing fractions were pooled and concentrated using Amicon Ultra centrifugal filters (Millipore) with a molecular weight cutoff of 50 kDa. Size exclusion chromatography (SEC) was carried out in 50 mM Tris-HCl (pH 7.8), 30 mM NaCl, 100 mM glycine, 10 μ M DTT, using a 24 ml 10/300 SuperdexTM200 GL column (GE Healthcare) connected to an ÄKTApurifier (GE Healthcare) at a flow rate of 0.25 ml/min. In SEC experiments, the pooled C-P4H-I samples eluted in two or three peaks, the first peak (major peak) being the complete C-P4H-I tetramer (Supplementary Figures S1-S3). The purified C-P4H-I samples in SEC buffer were divided into 10-50 μ l aliquots and stored at -70°C after quick freezing in liquid nitrogen.

During and after purification, C-P4H-I samples were analyzed by 8% SDS-PAGE under reducing conditions followed by Coomassie Blue staining. Western blotting was done using a polyclonal antibody generated against the purified recombinant PSB domain of the human C-P4H α (I) subunit [36] (Innovagen) and enhanced chemiluminescence detection (Amersham Biosciences). For N terminal sequencing, the protein samples were electroblotted onto a ProBlott membrane (Applied Biosystems Inc.), and protein bands (visualized using Ponceau S) were loaded onto a 492A Protein Sequencer (Applied Biosystems Inc.).

Activity assays and peptides

The catalytic properties of purified C-P4H-I were characterized using a method based on the hydroxylation-coupled decarboxylation of 2-oxo[1-¹⁴C]glutarate [37]. Synthetic peptides with three, five, or ten Pro-Pro-Gly triplets ((PPG)₃, (PPG)₅, and (PPG)₁₀, respectively) were used as substrates, and the liberation of radioactive CO₂ was measured using a Tri-Carb 2900 TR liquid scintillation counter (Perkin Elmer). eC-P4H-I was used for further activity assays with various synthetic peptides, with variations in the number and order of the -PPG- and -PPP- triplets. These hybrid peptides included (PPG)₂(PPP)₂P, P(PPP)₂(PPG)₂, (PPG)₃(PPP)₃, (PPP)₃(PPG)₃, (PPG)₄(PPP)₄, and (PPP)₄(PPG)₄. The inhibitory and/or binding properties of

poly-L-proline peptides consisting of nine (P9) and 24 (P24) proline residues were also studied. The peptides used in the SAXS measurements were P24 (inhibitor) and (PPG)₈ (substrate). All peptides were from TAG Copenhagen A/S.

Mass spectrometry

Purified C-P4H-I samples were analyzed by electrospray ionization(ESI)-liquid chromatography(LC)-mass spectrometry (MS) on a Q-ToF instrument (Waters Synapt G1), using reversed phase chromatography (column: Waters gradient acidified with 0.2% (v/v) formic acid). SDS-PAGE bands were analyzed by MS peptide mapping after tryptic digestion of the excised gel pieces. Coomassie-stained gels were destained with 40% (v/v) acetonitrile in 50 mM ammonium bicarbonate, reduced with 20 mM DTT in destaining solution, and alkylated with 40 mM iodoacetamide. Washed gel pieces were incubated with 20 µl trypsin solution (Sigma proteomics grade, 5 ng/µl in 10 mM ammonium bicarbonate, 2% (v/v) acetonitrile), and incubated overnight at +37°C. The supernatant was subjected to MALDI-ToF MS using an UltrafleXtreme (Bruker) instrument or acidified with trifluoroacetic acid and analyzed by nano LC-MS on a Synapt G2 (Waters), using standard protocols. Spectra were analyzed and assigned to protein sequences using the software of the instrument supplier (Bruker Biotools and Waters PLGS).

Static light scattering

Absolute molecular weight information for C-P4H-I and its truncated variants was obtained using static light scattering (SLS). The samples were analyzed with the miniDAWNTM TREOS multi-angle light scattering detector (Wyatt Technology), in on-line mode after the sample eluted from the SEC column. Before entering the SLS device, the sample flowed through a RI-101 refractometer (Shodex) for measuring the protein concentration. The molecular weights were calculated using the ASTRA software (Wyatt Technology).

Small-angle X-ray scattering

SAXS samples were concentrated to 2–21 mg/ml and stored at -70°C . After thawing, the samples were filtered using 0.22- μm centrifugal filter devices (Millipore), and the protein concentration was re-determined using absorbance at 280 nm. A sample of the identical buffer was always used as a blank measurement.

SAXS data for eC-P4H-I were also collected in the presence of 100 and 300 μM of the peptides (PPG)₈ and P24. All peptide measurements (including one apo eC-P4H-I control without peptide) were done in the presence of the C-P4H inhibitors 50 μM Zn^{2+} and 60 μM pyridine 2,4-dicarboxylate (PDC, a 2OG analogue). Zinc and PDC were added to the sample by dialyzing the concentrated protein-peptide samples in the C-P4H SEC buffer supplemented with 50 μM ZnSO_4 and 60 μM PDC.

Initial SAXS data of the eC-P4H-I and its truncated variant were collected using synchrotron radiation at MAX-Lab (Lund, Sweden) and ESRF (Grenoble, France). All the final SAXS data were collected on the EMBL/DESY (Hamburg, Germany) beamline P12 (PETRA III) [38]. SAXS data processing, analysis, and molecular modeling were carried out using ATSAS [39]. As control samples, well-behaving monomeric proteins, such as bovine serum albumin, calmodulin, and myelin protein P2 [40] were always included in the assay series, to enable reliable molecular weight determination. Molecular masses of the samples were estimated by comparing the observed forward scattering, $I(0)$, to that obtained using standard proteins. Data processing was done using PRIMUS [41], distance distribution functions were calculated using GNOM [42], and *ab initio* chain-like models were built using GASBOR [43]. Rigid body refinement, using models derived from crystal structures, coupled to building of missing loops of the C-P4H-I tetramer was done using CORAL [44] and BUNCH [45] and the scattering data of eC-P4H-I. Model building was based on the crystal structure of human PDI (residues 18-476, PDB entry 4ELI) [26], the crystal structure of the N terminal region of human C-P4H-I $\alpha(\text{I})$ (residues 3-238, PDB entry 4BT9) [20], and a model of the human C-P4H-I $\alpha(\text{I})$ CAT domain (residues 309-504, see the details below).

Theoretical scattering curves were calculated using CRY SOL [46].

Bioinformatics methods

The atomic model coordinates for the C-P4H-I α (I) CAT domain were generated with MODELLER [47] by using the crystal structure of (Ser-Pro)₅ complexed Cr-P4H-1 (PDB entry 3GZE) [31]. PredictProtein (<https://www.predictprotein.org/>), Jpred4 [48], and Phyre2 [49] were used for various sequence analyses including secondary structure predictions for the complete α subunit of C-P4H-I.

RESULTS

Bioinformatics studies of the C-P4H α subunit isoforms

The amino acid sequences of various C-P4H α subunits were studied using several bioinformatics tools. All known C-P4H α subunits contain three domains (Figure 1b), the C terminal CAT domain being the most conserved (Figure 2). Unlike for the N terminal region, consisting of the dimerization and PSB domains (also referred to as the double domain, DD) [20], a crystal structure of the C terminal CAT domain has not yet been determined, due to its poor solubility. However, according to secondary structure predictions and 26% sequence identity, the structure of the CAT domain is proposed to be very similar to that of Cr-P4H-1 (Figure 2) [28]. Thus, the modes of binding of the peptide substrate and the 2OG cofactor are predicted to be very similar in the C-P4H CAT domain. The Fe²⁺-binding catalytic triad, His412 (β II), Asp414 (β II), and His483 (β VII), and the 2OG-binding Lys493 (β VIII) (the amino acid numbers from here onwards refer to those of the human α (I) subunit), are the same as in Cr-P4H-1 [15,17]. In addition, Tyr403 (β I) and Thr449 (β IV), which interact with 2OG in Cr-P4H-1 [28], are highly conserved in the C-P4H α sequences (Figure 2). From the structure of Cr-P4H-1 complexed with a proline-rich peptide substrate [31], it is also expected that Arg362 (β 6), Trp368 (β 6), Glu396 (β I), Tyr409 (β II), Arg432 (β II), and Trp499 (β VIII)

participate in peptide substrate binding by the C-P4H CAT domain. All these residues are fully conserved in the C-P4H family (Figure 2). The major differences between algal/plant P4Hs and C-P4Hs are the absence of an extended loop (15 residues) between β IV and β V in C-P4H and the extra C terminal extension (10 residues) present in the C-P4H family but not in Cr-P4H-1 (Figure 2). The extra C terminal residues of the C-P4H α subunit are predicted to be ordered but possibly without secondary structure. It can also be noted that this extension and its immediate neighboring sequence have a number of fully conserved residues, such as Glu506, Phe507, Arg509, Pro510, and Cys511 (Figure 2). Cys511 is important for the assembly of the C-P4H tetramer and may form an intrachain disulfide bridge with Cys486 (β VII) [15]. Interestingly, the corresponding cysteine of Cr-P4H-1 (Cys230) forms a disulfide bridge with Cys195 of its extended β IV- β V loop [28].

The region between the PSB and CAT domains has unknown structure and function. This 70-residue segment, from Lys239 to Pro308, is referred to as the linker region (Figure 2). Our bioinformatics analysis proposes that the C terminal end of the linker region (Arg287-Pro308) forms part of the CAT domain, having two additional β strands (β 1 and β 2) when compared to the crystal structure of Cr-P4H-1. The N terminal part of the linker (Lys239-Arg286) is predicted to be solvent-exposed, with one α helix (α 12). The linker region contains several fully conserved residues, including Arg269, Tyr272, Glu273, Leu275, and Cys276 in the α 12 helix, and Leu291, Cys293, and Tyr295 in the β 1 strand. Cys276 and Cys293 are important for the assembly of the C-P4H-tetramer, possibly by forming an intrachain disulfide bond in the α subunit [50].

Purification and initial characterization of the complete human C-P4H-I tetramer and its truncated variants expressed in E. coli

eC-P4H-I was purified using poly-L-proline affinity, anion exchange chromatography, and SEC. The major peak of the SEC chromatogram of the eC-P4H-I sample (Supplementary

Figure S1) corresponds to a molecular mass of 250 kDa in the SLS analysis (Table 1), and Western blot and MS analyses revealed that both α and β subunits are present in the single SDS-PAGE band of 60 kDa (Supplementary Figure S1). Further MS analysis showed that the full masses of the α (including the additional N terminal Met residue) and PDI/ β subunits were 59241 Da and 55423 Da, respectively. These data indicate that the major peak of SEC represents the eC-P4H-I heterotetramer. Activity analysis of this sample showed that this tetramer is fully active when compared to earlier data [32-34].

The expression and purification of eC-P4H-I also generated a truncated variant, which co-eluted with the complete eC-P4H-I tetramer during poly-L-proline affinity chromatography, but which was separated from the tetramer during SEC (Supplementary Figure S1). According to SLS, the molecular mass of this eC-P4H-I minor peak was 148 kDa (Table 1). A strong band of 60 kDa was seen in the SDS-PAGE analysis of this sample; in addition, this sample also includes a fragment of 25 kDa (Supplementary Figure S1). Peptide mapping analysis by MS identified complete α and β subunits in the 60 kDa band, whereas the 25 kDa fragment corresponded to the N terminal region of the α subunit. Further MS analysis of this fragment revealed the presence of two peptides, with masses of 24008 Da and 24637 Da. The size of the fragments agrees with an α subunit cleavage site after Leu209, Leu210, or Leu214. This leucine-rich region locates in α 10 at the end of the PSB domain (Figure 2). These data indicate that the truncated eC-P4H-I assembly contains a full-length α subunit, a full-length β subunit, as well as a truncated N terminal half of a second α subunit, having its truncation site near the leucine-rich region of α 10 of the PSB domain (Figure 2). From now on, the truncated eC-P4H-I variant (the eC-P4H-I minor peak of the SEC step) is referred to as eC-P4H-I($\beta\alpha\alpha'$) and the complete eC-P4H-I tetramer is referred to as eC-P4H-I($\beta\alpha\alpha\beta$). In activity assays, eC-P4H-I($\beta\alpha\alpha'$) has similar specific activity as eC-P4H-I($\beta\alpha\alpha\beta$).

Purification and initial characterization of the complete human C-P4H-I tetramer and its truncated variants expressed in yeast and insect cells

Recombinant human C-P4H-I tetramers expressed in insect cells (referred to as iC-P4H-I) and in the yeast *P. pastoris* (referred to as yC-P4H-I) are modified by glycosylation [33,34]. To investigate whether glycosylation affects the C-P4H-I tetramer structure, C-P4H-I expressed in these eukaryotic systems was also studied. yC-P4H-I and iC-P4H-I were purified using the same protocol as eC-P4H-I. Like for the eC-P4H-I purification, in the final SEC step, two peaks were observed (Table 1; Supplementary Figures S2 and S3), indicating the presence of truncated C-P4H-I variants. yC-P4H-I($\beta\alpha\alpha\beta$) and iC-P4H-I($\beta\alpha\alpha\beta$) eluted in the major peak, with almost the same elution volume as eC-P4H-I($\beta\alpha\alpha\beta$) (Table 1). Because glycosylation increases the size of the α subunit in eukaryotic expression systems (as discussed below), the full-length α and β subunits are better separated in SDS-PAGE (Supplementary Figures S2 and S3). Several lower-molecular-weight bands are seen in SDS-PAGE for both samples. All these bands in iC-P4H-I were identified as degraded α subunit fragments in Western blot analysis or N terminal Edman sequencing. It should be noted that yC-P4H-I was expressed with an additional α MF sequence at the N terminus of the PDI/ β subunit [34,35]. This pre-sequence, which includes also one glycosylation site, is not completely removed during expression, causing additional sample heterogeneity (Supplementary Figure S2). yC-P4H-I($\beta\alpha\alpha\beta$) and iC-P4H-I($\beta\alpha\alpha\beta$) have the same specific activity as eC-P4H-I($\beta\alpha\alpha\beta$), but nevertheless the quality of the yC-P4H-I($\beta\alpha\alpha\beta$) and iC-P4H-I($\beta\alpha\alpha\beta$) is lower than that of eC-P4H-I($\beta\alpha\alpha\beta$), in particular due to proteolysis of the α subunit (Supplementary Figures S1-S3). Therefore, the eC-P4H-I($\beta\alpha\alpha\beta$) sample has been used for the detailed enzyme kinetic and SAXS modelling studies (below).

The truncated variants of yC-P4H-I and iC-P4H-I had a molecular mass of approximately 100 kDa (Table 1; Supplementary Figures S2 and S3). The full-length α subunit is hardly visible in SDS-PAGE of the minor SEC peak (Supplementary Figures S2

and S3). Instead, an additional band of 45 kDa is present. According to MS, this fragment comprises the C terminal part of the α subunit. N terminal sequencing of the iC-P4H-I sample identified a cleavage site between Gln121 and Asp122 (in $\alpha 5$ of the dimerization domain, Figure 2). These data propose that the minor peaks of yC-P4H-I and iC-P4H-I correspond to a dimeric C-P4H-I, having a full-length β subunit and a truncated α subunit (~45 kDa) starting at Asp122. These truncated variants of yCP4H-I and iC-P4H-I are referred to as yC-P4H-I($\beta\alpha'$) and iC-P4H-I($\beta\alpha'$), respectively. Unlike eC-P4H-I($\beta\alpha\alpha''$), yC-P4H-I($\beta\alpha'$) and iC-P4H-I($\beta\alpha'$) are completely inactive.

Solution structure of the complete C-P4H-I tetramer and its truncated variants

Synchrotron SAXS data were collected for the complete eC-P4H-I and yC-P4H-I tetramers and their truncated variants (Figure 3a,b; Table 2). The complete human C-P4H-I tetramer, regardless of the expression host, is an elongated, somewhat curved complex (Figure 3c,d). The distance distribution functions of the eC-P4H-I($\beta\alpha\alpha\beta$) and yC-P4H-I($\beta\alpha\alpha\beta$) tetramers are slightly different (Figure 3b). The maximal dimensions are 275 Å for the yC-P4H-I and 290 Å for eC-P4H-I, with R_g values slightly above 80 Å (Table 2). The *ab initio* models of eC-P4H-I($\beta\alpha\alpha\beta$) and yC-P4H-I($\beta\alpha\alpha\beta$) show that the enzyme has two-fold symmetry, with a thin middle part (waist) and bulkier lobes at both ends (Figure 3c,d). The eC-P4H-I($\beta\alpha\alpha''$) and yC-P4H-I($\beta\alpha'$) variants have a lower mass, R_g , and D_{max} than the complete heterotetramer (Table 2), and their shape is also asymmetric (Figure 3e,f).

To study the effect of proline-rich peptides on the C-P4H shape, eC-P4H-I($\beta\alpha\alpha\beta$) was studied by SAXS in the presence of a peptide substrate (PPG)₈ and an inhibitor P24 (Figure 4). Both peptides change slightly the shape of the C-P4H-I tetramer (Table 2; Figure 4a,b). C-P4H-I complexed with a proline-rich peptide ligand is more compact, having smaller R_g and D_{max} values compared to apo C-P4H-I (Table 2). This difference is also seen in the

dimensionless Kratky plot (Figure 4c), suggesting that upon peptide binding C-P4H-I becomes more compact and more rigid.

Hybrid modeling of C-P4H-I

The SAXS *ab initio* envelopes of eC-P4H-I and yC-P4H-I (Figure 3c,d) reveal the basic shape and dimensions of the C-P4H-I tetramer. The narrow central part of the SAXS envelope proposes the presence of only the coiled-coil dimerization motif of the α subunit in this region, as seen in the crystal structure of the DD dimer [20]. The shape proposes that the globular bulky lobes, at both ends of the envelope, are formed by the PDI/ β subunit, tightly interacting with the CAT domain of the α subunit (Figure 5b,c). This assembly fits with the MS data for the truncated $\beta\alpha'$ dimer (iC-P4H-I and yC-P4H-I) and the $\beta\alpha\alpha''$ trimer (eC-P4H-I), for which the SAXS shapes are asymmetric (Figures 3e,f and 5a,d).

To obtain a better insight into the C-P4H-I solution structure, hybrid modeling approaches using CORAL [44] and BUNCH [45] were employed, involving the use of rigid body models of the separate domains, supplemented by building of loops and termini as chain-like structures. First, a SAXS hybrid model was calculated for the yC-P4H-I($\beta\alpha'$), for which high-quality SAXS data were measured (Table 2; Figure 3a,b,f). As the α' chain of this truncated variant lacks the dimerization domain, only the Asp122-Glu238 region of the C-P4H-I DD crystal structure (including $\alpha 6$ of the dimerization domain and the whole PSB domain) [20] was included. Modeling using three rigid bodies (PSB, CAT, and PDI) suggests that the PDI and CAT domains form a tight $\beta\alpha'$ complex (Figure 5a). In the model, no interactions are seen between the PDI and PSB domains, whereas the PSB and CAT domains are in close contact (Figure 5a). The 70-residue linker region between the PSB and CAT domains, connects the PSB and CAT domains adopting a compact random coil structure. The C terminal residues of the PDI/ β subunit (32 residues including the KDEL ER retention signal) as well as the N terminal 74 residues of the α MF sequence of the PDI/ β subunit

(excluding the ER targeting signal of α MF [34]) are solvent-exposed (Figure 5a). This hybrid model fits well to the experimental scattering data collected from yC-P4H-I($\beta\alpha'$) (Figure 5e).

Subsequent modeling of the complete C-P4H-I tetramer was based on the assumption that the CAT-PDI/ β assembly is the same in the complete tetramer as in yC-P4H-I($\beta\alpha'$), but that the PSB domain may adopt another position in the complete tetramer. Therefore, only the CAT-PDI/ β part of the yC-P4H-I($\beta\alpha'$) hybrid model (excluding also the linker region of the α chain as well as the α MF sequence of the β chain) was subsequently used for hybrid modeling of the complete eC-P4H-I($\beta\alpha\alpha\beta$) tetramer. The SAXS *ab initio* model of the DD construct has the same shape as its various crystal structures [20], suggesting that the conformation of the DD region is well defined and will be preserved in the complete tetramer. Consequently, the SAXS hybrid model calculations were done using two rigid units, being the DD structure as well as the two CAT-PDI/ β assemblies, and using P2 symmetry. The resulting hybrid model of eC-P4H-I($\beta\alpha\alpha\beta$) fits well to the experimental SAXS data (Figure 5f) as well as to the corresponding *ab initio* shape (Figure 5b). In the model, the DD dimer defines the P2 symmetry of the complete tetramer and the N terminal coiled-coil helices are positioned between the two CAT-PDI/ β parts. The CAT-PDI/ β assemblies locate somewhat above the DD plane, causing a curved shape for the eC-P4H-I($\beta\alpha\alpha\beta$) heterotetramer (Figure 5b). The PSB-CAT linker region in the α subunit forms a compact loop structure (Figure 5b), and the N and C (with KDEL) termini of the PDI/ β subunit are solvent exposed, as in the yC-P4H-I($\beta\alpha'$) model (Figure 5a,b). The hybrid model of eC-P4H-I($\beta\alpha\alpha\beta$) agrees also with the SAXS data and the *ab initio* model of yC-P4H-I($\beta\alpha\alpha\beta$) (Figure 5c,g). After removing the second CAT-PDI/ β part from eC-P4H-I($\beta\alpha\alpha\beta$), the model fits also well to the eC-P4H-I($\beta\alpha\alpha''$) experimental data (Figure 5d,h).

Taken together, the SAXS experiments show that: i) the C-P4H-I tetramer is an elongated molecule with two-fold symmetry, with the CAT domains pointing away from each other, ii) the central part of the $\beta\alpha\alpha\beta$ tetramer is a thin region which accommodates the

dimerization region of the α subunits, iii) there is close interaction between the CAT domain and the PDI/ β subunit, and iv) the bound substrate and inhibitor peptides induce a conformational change in the C-P4H-I($\beta\alpha\alpha\beta$) tetramer, making it more compact.

Kinetic analysis of C-P4H-I with proline-rich peptides

To study the C-P4H-I reaction mechanism, kinetic assays were performed with eC-P4H-I($\beta\alpha\alpha\beta$) using various proline-rich peptides (Table 3), including hybrid peptides containing PPG (substrate) repeats as well as poly-L-proline (inhibitor) repeats of different length and different order. The longest peptide in these assays consists of four repeats of both, having altogether 24 residues, being either (PPG)₄(PPP)₄ or (PPP)₄(PPG)₄ (Table 3). The K_m values from two different measurements for (PPG)₄(PPP)₄ were 84 and 124 μ M, whereas the values were 265 and 428 μ M for (PPP)₄(PPG)₄. Thus, the hybrid peptide having the PPG repeat at the N terminal end has a higher affinity for eC-P4H-I($\beta\alpha\alpha\beta$) than the peptide with the same length but having the PPG repeat at the C terminal end. Typically, longer peptides are better substrates for C-P4Hs [4]. This is also seen in our activity measurements (Table 3). For the shortest hybrid peptides, (PPG)₂(PPP)₂P and P(PPP)₂(PPG)₂, the latter peptide is not a substrate, but instead acts as an inhibitor when (PPG)₁₀ is used as a substrate, whereas (PPG)₂(PPP)₂P is a weak substrate with K_m close to 1 mM (Table 3).

C-P4H-I is effectively inhibited by long poly-L-proline peptides [4]. However, the inhibition of C-P4H-I by short poly-L-prolines has not been studied in detail before. We showed before that P9 has high affinity (10 μ M) towards the PSB domain of C-P4H-I [20]. Therefore, this peptide was also included in the enzyme kinetic assays, the assumption being that it binds to the PSB domain and thereby affects the activity of the C-P4H-I tetramer. However, C-P4H-I was only inhibited (30% inhibition) at a very high concentration (4 mM) of P9, when (PPG)₁₀ was the substrate. In the presence of 2 mM P9, C-P4H-I was inhibited by

3-14%, when (PPG)₁₀ was used, but no inhibition was observed when a shorter peptide, (PPG)₃, was used as a substrate.

DISCUSSION

The βααβ assembly

The crystal structure of the DD region of the C-P4H α(I) subunit has important implications for understanding the C-P4H assembly, showing that the N terminal α helices of the α subunit are responsible for the α-α dimerization [20]. This finding provided also a reliable starting point for the SAXS modeling of the C-P4H-I tetramer structure (Figures 5 and 6). The two dimerization domains, as seen in the DD crystal structure, are at the center of the βααβ model of the tetramer, being extended at both ends by the PSB domain. The PDI/β subunit and the CAT domain of the α subunit are located at the ends of the bilobal complex. According to the model of γC-P4H-I(βα'), the PDI/β subunit closely interacts with the CAT domain. The PDI/β subunit is essential for keeping the C-P4H α subunit soluble [25], and indeed, the CAT domain is the only domain of the α subunit, which cannot be recombinantly expressed as an individual domain because of its poor solubility. The role of PDI/β in the assembly process of C-P4H βααβ is not clearly established [6]. Most likely, it is involved in chaperoning the folding and SS-bridge formation (oxidation) of the α subunit. If in the so-formed C-P4H complex, the subunits remain associated after assembly, then PDI/β will be in the reduced form (active sites in the SH form) and the α subunit will be in the oxidized (SS bridged) form. However, it has been reported that in the active C-P4H tetramer, both the α and PDI/β subunits are oxidized [51]. It is possible that the reduced PDI is exchanged for its oxidized form, or that the CAT domain is oxidized in some other way (for example by glutathione [50]) before forming a complex with oxidized PDI. Much structural information has been obtained recently for the PDI/β subunit due to the crystal structure determination of full-length PDI in its reduced and oxidized (having active sites SS bridged) forms [26,52,53]. In

the latter form, the **b'** and **a'** domains form an extensive hydrophobic surface, believed to play an important role in the chaperone activity of PDI [26,54]. In our SAXS model calculated using the oxidized PDI structure, the CAT domain interacts with both PDI/ β active site regions being bound in the cleft shaped by the **a** and **a'** domains of PDI (Figures 5 and 6), similarly to the ERp57-tapasin complex [55]. The CAT domain interacts with the **a**, **b'**, and **a'** domains of PDI/ β , whereas no interactions are seen with the **b** domain. This is in line with mutagenesis experiments that show that **a**, **b'** and **a'** domains are needed for the formation of the active C-P4H-I tetramer [56]. Interestingly, the active site of the **a** domain (Cys36 and Cys39) is close to Cys486 of the α subunit, which may form an intrachain disulfide bridge with Cys511 in the C-P4H-I tetramer (Figure 6). Although it has been shown that PDI activity is not needed for C-P4H-I tetramer formation or activity [25], the close location of the PDI active site to the possible disulfide bridge of the α subunit suggests that the active site of PDI could play a role in the formation and stabilization of disulfide bridges in the C-P4H α (I) subunit. In Cr-P4H-1, the long β IV- β V loop covers the cysteine residue corresponding to the C-P4H-I α (I) Cys486 [28]. A similarly elongated β IV- β V structure is not found in the α subunit of any tetrameric C-P4H (Figure 2), correlating with the notion that the β IV- β V loop is topologically replaced by PDI/ β in the C-P4Hs. Indeed, our model suggests that the **a** domain of PDI/ β interacts with this part of the α subunit in the tetramer. The KDEL region at the C terminus of the **a'** domain of PDI is solvent-exposed, in agreement with its role as an ER retention signal (Figure 6).

Human C-P4H-I has two N-linked glycosylation sites (Asn96 and Asn242), but the glycosylation is not required for the assembly or activity of the C-P4H-I tetramer [15]. The glycosylation of iC-P4H-I($\beta\alpha\alpha\beta$) and yC-P4H-I($\beta\alpha\alpha\beta$) has not been characterized in detail, but it has been shown with C-P4H-I purified from chicken embryos that when both sites are maximally glycosylated, both sites are occupied by N-linked eight-mannose oligosaccharide units [57]. In yC-P4H-I the level of glycosylation is likely to be somewhat higher than in the

native C-P4H-I, as hyperglycosylation of recombinant proteins expressed in yeast is known phenomenon [35]. In addition, the α MF sequence of yC-P4H-I($\beta\alpha\alpha\beta$) also contains one N-linked oligosaccharide [34]. Because of the glycosylation as well as the presence of the partially cleaved α MF fragment, yC-P4H-I($\beta\alpha\alpha\beta$) is expected to have a molecular mass ranging from 240 kDa to 260 kDa (the mass of non-glycosylated eC-P4H-I($\beta\alpha\alpha\beta$) is 230 kDa). The molecular weights for the complete tetramers, determined using SLS and SAXS, deviate somewhat from each other, however the R_g values and the SEC elution volumes (Tables 1 and 2) agree with the notion that yC-P4H-I($\beta\alpha\alpha\beta$) is larger in size than eC-P4H-I($\beta\alpha\alpha\beta$). The above-mentioned differences also correlate with the differences in the SAXS scattering profiles (Figures 3b and 4c) and the *ab initio* models (Figure 3c,d) of the yeast and *E. coli* expressed proteins, and the increase bulkiness of the two lobes, when compared to the eC-P4H-I($\beta\alpha\alpha\beta$), can be attributed to glycosylation in the case of yC-P4H-I($\beta\alpha\alpha\beta$). The resolution of the SAXS data does not allow to reliably model the carbohydrate units of yC-P4H-I($\beta\alpha\alpha\beta$), but both sites are accessible to bulk solvent in our model (Figure 6).

The truncated $\beta\alpha'$ and $\beta\alpha\alpha''$ variants

Human C-P4H-I purified from eukaryotic (yeast and insect) and prokaryotic (*E. coli*) expression systems consists of a mixture of complete tetramers and differently truncated variants. The complete tetramer and its truncated forms co-purified when using the poly-L-proline affinity column, which indicates that each of these variants has at least one functional PSB domain. The truncated iC-P4H-I and yC-P4H-I variants are $\beta\alpha'$ dimers, built up of a complete PDI/ β subunit complexed with an N terminally truncated α chain (α'). The truncated eC-P4H-I, a $\beta\alpha\alpha''$ trimer, instead, is built up of a complete $\beta\alpha$ unit, which is further dimerized with a truncated α subunit (α'') lacking the CAT domain. The $\beta\alpha'$ assembly is inactive, whereas the $\beta\alpha\alpha''$ is active. Apparently, dimerization between the α subunits has an important role in achieving a competent PSB-CAT assembly in the first α

subunit. Indeed, the N terminal dimerization domain of the second α subunit interacts with the PSB domain of the first subunit as seen in the DD crystal structure [20]. Helix $\alpha 5$ also plays an important role in the PSB-dimerization domain assembly [20]. This helix, with a high sequence conservation within the C-P4H family (Figure 2), is not present in the inactive $\beta\alpha'$ assembly. In addition to residues that form contacts between the dimerization and PSB domains, the $\alpha 5$ helix contains several other highly conserved residues such as Glu108, Asp109, Arg119, Asp122, and Tyr124 (Figure 2), which are fully solvent-exposed in the DD crystal structure [20]. Although these conserved residues are not important for the folding or stability of the DD region, the high sequence conservation suggests that they are important for the folding and/or function of the C-P4H tetramer. The PSB domain of the α'' subunit is probably not functional, because it lacks $\alpha 11$, which has several conserved residues important for peptide-substrate binding, such as Tyr230 and Tyr233 (Figure 2) [20,22].

The processivity and topology of the C-P4H-I-peptide substrate interactions

The substrate of C-P4H is a monomeric procollagen polypeptide, which in the case of the fibril-forming type I, II, and III collagens contains multiple X-P-G tripeptide motifs amounting to approximately 1000 residues. C-P4H is a very efficient enzyme when acting on its procollagen substrate. The k_{cat}/K_m ratio is much higher than the rate for diffusion-controlled encounter between an enzyme and its substrate, suggesting that C-P4H is a processive enzyme when hydroxylating procollagen [58]. Processive enzymes remain bound to their substrates, while performing multiple rounds of catalysis. Well-studied processive enzymes use electrostatic (*e.g.* DNA processing enzymes) or hydrophobic interactions (*e.g.* cellulases) for retaining their substrates [59,60]. Characteristically, these enzymes have relatively large binding surfaces between ligand and enzyme. The PSB and CAT domains of C-P4H contain many conserved solvent-exposed aromatic residues in their peptide binding sites (Figure 2), which can form stacking/C-H... π interactions with multiple proline residues

of collagenous peptides [20,31]. Early studies on the hydroxylation of a (PPG)₁₀ peptide have shown that hydroxylation is most prominent at the C terminal end of the peptide [61]. This observation is consistent with the notion that the catalytic machinery of C-P4H is processive, for example that it is effectively moving from the N to the C terminus of the procollagen substrate. The PSB domain has low affinity for the hydroxylated peptide product generated by the CAT domain [62]. This observation suggests that the hydroxylated peptide triplet in the CAT domain has to move away from the PSB domain, which implies that the CAT-PSB unit moves towards the C terminus of the monomeric procollagen chain (Figure 6). The PSB binding pocket interacts with its ligand in such a way that the C terminal helix ($\alpha 11$) interacts with the N terminus of the peptide [20]. All these observations predict that the PSB domain binds downstream, at the C terminal end of the procollagen peptide substrate with respect to the CAT domain (Figure 6). Experiments with hybrid peptides with mixed PPG and PPP regions indeed showed that peptides with the PPG region at the N terminus and the poly-L-proline part at the C terminus are higher affinity substrates for C-P4H-I than the peptides with a reverse order (Table 3). Peptides with an N terminal PPG part are apparently better positioned for catalytically competent interactions with the CAT active site, whenever the C terminal PPP part binds to the PSB domain (it should be noted that poly-L-proline peptides have higher affinity to PSB domain than (PPG)_n peptides [20,62]). Hybrid peptides with different lengths were tested to get some insight into the spatial separation of the peptide-binding pockets in the PSB and CAT domains. The longest hybrid peptide tested, having an N terminal PPG region, was the best substrate for C-P4H-I, indicating that the peptide is long enough for the PPG region to reach the CAT active site. Also (PPG)₃(PPP)₃ was a good substrate. However, the shortest hybrid peptide of our study with an N terminal PPG region, (PPG)₂(PPP)₂P, is only a weak substrate for C-P4H-I. Apparently, this peptide binds mainly to the PSB domain, and in this mode of binding, the PPG region is not optimally placed for competent interactions with the active site of the CAT domain.

The longest hybrid peptides with an N terminal PPP region, the ((PPP)₄(PPG)₄ and

(PPP)₃(PPG)₃), are substrates for C-P4H-I, but the K_m values are significantly higher than for the peptides of reversed order. Our data predict that the PPG regions of these peptides are hydroxylated by the CAT domain being bound in such a way that the PPP region cannot interact with the PSB domain (Table 3). The short peptides P(PPP)₂(PPG)₂ and P9 are weak inhibitors of C-P4H-I when tested with (PPG)₁₀ as substrate. We assume that they both bind to the PSB domain only, leaving the CAT domain free for catalysis.

The possible function of the PSB domain

The peptide binding studies raise an interesting question about the functional interaction of the PSB and CAT domains. Our enzymatic assays in the presence of P9 (which has high affinity for the PSB domain [20]) propose that C-P4H-I can hydroxylate synthetic (PPG)_n peptides *in vitro* without the substrate peptide binding to the PSB domain (Table 3). This is also in line with results obtained from complementary point mutation studies of the PSB domain and the complete C-P4H-I [22]. The *in vivo* function of the PSB domain may relate to capturing the long procollagen chain. It is noteworthy that procollagen chains have variable X-Y-G sequence repeats [1], and the PSB domain might be important in binding to specific regions (with very high affinity), in order to facilitate the hydroxylation of several PPG motifs up-stream, before the procollagen chain is released from the PSB domain. Concerning the reaction mechanism, it is interesting to note that the CAT domain binds the cosubstrate, 2OG, in a completely buried fashion, in a pocket shielded from bulk solvent by the bound proline-rich peptide [28,31]. During each reaction cycle, 2OG is converted into succinate (Figure 1). Therefore, completion of the reaction cycle requires that the product peptide must dissociate from its binding site, allowing succinate to be exchanged for a new 2OG molecule. This suggests a function for the PSB domain in holding on to the procollagen substrate, while a competent active site of the CAT domain is regenerated. The affinity of the PSB domain for its bound peptide determines its relative “residence” time, during which the catalytic domain can hydroxylate several PPG tripeptide repeats [58]. This hypothesis suggests that the number

of residues between the catalytic site interactions (labelled as C1 to C9 in Table 3) and the anchoring interactions (labelled as P1 to P9 in Table 3) can vary (referred to as “X” in Table 3). According to the SAXS model, the PSB and CAT domains closely interact. The smallest number of X residues is zero, meaning that a 9-residue peptide bound to the PSB binding pocket is 5-13 residues towards the C terminus from the hydroxylation site (Table 3). Indeed, our SAXS peptide binding studies showed that in the presence of the P24 and (PPG)₈ peptides, the C-P4H-I tetramer adopts a more compact shape (Table 2), suggesting that a peptide can bind to both the CAT and PSB domains simultaneously.

CONCLUSIONS

The proposed quaternary structure of C-P4H-I provides a functional model for the organization of all subunits and domains in the complete C-P4H tetramer assembly (Figure 6). The unique function of each domain of the α subunit can now be proposed: the N terminal coiled-coil motif of the α subunit provides the stability of the α_2 dimer, and the PSB domain anchors the procollagen substrate, such that several catalytic cycles by the CAT domain are possible before the release of the procollagen substrate. The CAT domain forms the major site of interaction between the α chain and the PDI/ β chain. The finding that the C-P4H-I tetramer is an elongated symmetric molecule with a $\beta\alpha\alpha\beta$ architecture also defines the topology of the competent enzyme-substrate complex, such that effectively the CAT-PSB unit moves over the procollagen polypeptide from the N towards the C terminus (Figure 6). The model also suggests that the two CAT-PSB units of the tetramer are located far apart and function independently, which is in good agreement with the observation that the eC-P4H-I($\beta\alpha\alpha'$) variant is fully active. Interestingly, however, inactivation of one of the catalytic sites in the *Caenorhabditis elegans* C-P4H tetramer with dissimilar α subunits reduced the total activity by more than half, indicating some co-operativity [63]. The observation that the $\beta\alpha'$ variant of human C-P4H-I is inactive highlights the importance of the α - α dimerization interactions for

the proper assembly and functioning of this C-P4H. We have initiated further structural studies in order to establish the structural details of the C-P4H tetramer assembly at high resolution, as well as to elucidate the precise interactions of the PSB and CAT domains with the procollagen substrate.

AUTHOR CONTRIBUTIONS

M.K.K., J.A., A.V.M. and P.D. did all the protein purifications and sample preparations for various assays. J.M. and A.V.M. carried out the Western blot and activity assays of the purified C-P4H samples. J.M. expressed iC-P4H and yC-P4H. The SLS analyses were done by M.K.K. J.A. and A.V.M. and all the SAXS experiments and analyses were done by P.K. Initial model of C-P4H was generated by M.K.K. and the model was further refined by J.A., P.K., A.V.M. and R.K.W. Mass spectrometry measurements and analyses were done by U.B., M.K.K., J.A. and A.V.M. The experiments were conceived, and the manuscript was written by M.K.K., J.A., P.K., J.M. and R.W. All authors provided intellectual content to and approved of the final version of the manuscript.

FUNDING

This work was supported by the EU-BIOXHIT grant to RKW, and by the Academy of Finland grants to RKW (200966), to MKK (115124), to JM (200471, 202469 and the Center of Excellence Grant 251314), and to PK (108423, 214317, 124241, 218045). We also acknowledge the Biocenter Finland support to the Biocenter Oulu Protein Crystallography Core Facility coordinated by RKW. The work was also funded by a fellowship from the CIMO organization to JA, grants from the Sigrid Jusélius Foundation to MKK/RW, PK and JM, a Jane and Aatos Erkko Foundation grant to JM, and funding from the Research and Science Foundation of the City of Hamburg (PK). The research leading to these results has also received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under BioStruct-X (grant agreement N°283570).

ACKNOWLEDGEMENTS

We thank Dr. Antje Neubauer for many valuable suggestions concerning the expression and purification of eC-P4H-I. Ville Ratas, Eeva Lehtimäki and Merja Nissilä are acknowledged for excellent technical assistance, M.Sc Ábris Bendes for SAXS data collection of eC-P4H-I($\beta\alpha\alpha'$), and Dr. Hongmin Tu for the N terminal sequencing analysis. We also wish to acknowledge the excellent support of the staff at the beamlines of MAX-Lab, ESRF, and EMBL/DESY.

REFERENCES

1. Piez, K.A. (1976) Biochemistry of collagens (eds. Ramachandran G.N. and Reddi, A.H.) 1–44 (Plenum Publishing Corporation, New York)
2. Myllyharju, J. and Kivirikko, K.I. (2004) Collagens, modifying enzymes and their mutations in human, flies and worms. *Trends Genet.* **20**, 33–43
3. Shoulders, M.D. and Raines, R.T. (2009) Collagen structure and stability. *Annu. Rev. Biochem.* **78**, 929–958
4. Myllyharju, J. (2003) Prolyl 4-hydroxylases, the key enzymes of collagen biosynthesis. *Matrix Biol.* **22**, 15–24
5. Myllyharju, J. (2008) Prolyl 4-hydroxylases, key enzymes in the synthesis of collagens and regulation of the response to hypoxia, and their roles as treatment targets. *Ann. Med.* **40**, 402–417
6. Gorres, K.L. and Raines, R.T. (2010) Prolyl 4-hydroxylases. *Crit. Rev. Biochem. Mol. Biol.* **45**, 106–124
7. Myllylä, R., Majamaa, K., Günzler, V., Hanauske-Abel, H.M. and Kivirikko, K.I. (1984) Ascorbate is consumed stoichiometrically in the uncoupled reactions catalyzed by prolyl 4-hydroxylase and lysyl hydroxylase. *J. Biol. Chem.* **259**:5403-5405
8. Franklin, T.J. (1997) Therapeutic approaches to organ fibrosis. *Int. J. Biochem. Cell Biol.*

29, 79–89

9. Gilkes, D.M., Chaturvedi, P., Baipai, S., Wong, C.C., Wei, H., Pitcairn, S., Hubbi, M.E., Wirtz, D., Semenza, G.L. (2013) Collagen prolyl hydroxylases are essential for breast cancer metastasis. *Cancer Res.* **73**, 3285-3296
10. Xiong, G., Deng, L., Zhu, J., Rychahou, P.G and Xu, R. (2014) Prolyl-4-hydroxylase α subunit 2 promotes breast cancer progression and metastasis by regulating collagen deposition. *BMC Cancer* **14**:1. Doi:10.1186/1471-2471-14-1
11. Qi, H.H., Ongusaha, P.P., Myllyharju, J., Cheng, D., Pakkanen, O., Shi, Y., Lee, S.W., Peng, J. and Shi, Y. (2008) Prolyl 4-hydroxylation regulates argonaute 2 stability. *Nature*, **455**, 421–423
12. Vasta, J.D. and Raines, R.T. (2015) Selective inhibition of prolyl 4-hydroxylases by bipyridinedicarboxylates. *Bioorg. Med. Chem.* **23**:3081-3090. doi:10.1016/j.bmc.2015.05.003
13. Helaakoski, T., Annunen, P., Vuori, K., MacNeil, I.A., Pihlajaniemi, T. and Kivirikko, K.I. (1995) Cloning, baculovirus expression, and characterization of a second mouse prolyl 4-hydroxylase α -subunit isoform: formation of an $\alpha_2\beta_2$ tetramer with the protein disulfide-isomerase/ β subunit. *Proc. Natl. Acad. Sci. U. S. A.* **92**, 4427–4431
14. Kukkola, L., Hieta, R., Kivirikko, K.I. and Myllyharju, J. (2003) Identification and characterization of a third human, rat, and mouse collagen prolyl 4-hydroxylase isoenzyme. *J. Biol. Chem.* **278**, 47685–47693
15. Lamberg, A., Pihlajaniemi, T. and Kivirikko, K.I. (1995) Site-directed mutagenesis of the α subunit of human prolyl 4-hydroxylase. Identification of three histidine residues critical for catalytic activity. *J. Biol. Chem.* **270**, 9926–9931
16. Myllyharju, J. and Kivirikko, K.I. (1997) Characterization of the iron- and 2-oxoglutarate-binding sites of human prolyl 4-hydroxylase. *EMBO J.* **16**, 1173–1180
17. Gorres, K.L., Pua, K.H. and Raines, R.T. (2009) Stringency of the 2-His-1-Asp active-site motif in prolyl 4-hydroxylase. *PLoS One* **4**(11):e7635. doi:

10.1371/journal.pone.0007635.

18. McDonough, M.A., Loenarz, C., Chowdurry, R., Clifton, I.J. and Schofield, C.J. (2010) Structural studies on human 2-oxoglutarate dependent oxygenases. *Cur. Opinion in Struct. Biol.* **20**, 659-672
19. Markolovic, S, Wilkins, S.E. and Schofield, C.J. (2015) Protein hydroxylation catalyzed by 2-oxoglutarate-dependent oxygenases. *J. Biol. Chem.* **290**, 20712-20722
20. Anantharajan, J., Koski, M.K., Kursula, P., Hieta, R., Bergmann, U., Myllyharju, J. and Wierenga, R.K. (2013) The structural motifs for substrate binding and dimerization of the α subunit of collagen prolyl 4-hydroxylase. *Structure* **21**, 2107-2118
21. Myllyharju, J. and Kivirikko, K.I. Identification of a novel proline-rich peptide-binding domain in prolyl 4-hydroxylase. *EMBO J.* **18**, 306–312 (1999).
22. Pekkala, M., Hieta, R., Bergmann, U., Kivirikko, K.I., Wierenga, R.K. and Myllyharju, J. (2004) The peptide-substrate-binding domain of collagen prolyl 4-hydroxylase is a tetratricopeptide repeat domain with functional aromatic residues. *J. Biol. Chem.* **279**, 52255–52261
23. Kozlov, G., Määttänen, P., Thomas, D.Y. and Gehring, K.A. (2010) Structural overview of the PDI family of proteins. *FEBS J.* **277**, 3924–3936
24. Khan, H.A. and Mutus, B. (2014) Protein disulfide isomerase a multifunctional protein with multiple physiological roles. *Front Chem.* **2**:70. doi: 10.3389/fchem.2014.00070
25. Vuori, K., Pihlajaniemi, T., Myllylä, R. and Kivirikko, K.I. (1992) Site-directed mutagenesis of human protein disulphide isomerase: effect on the assembly, activity and endoplasmic reticulum retention of human prolyl 4-hydroxylase in *Spodoptera frugiperda* insect cells. *EMBO J.* **11**, 4213-4217
26. Wang, C., Li, W., Ren, J., Fang, J., Ke, H., Gong, W., Feng, W. and Wang, C.C. (2013) Structural insights into the redox-regulated dynamic conformations of human protein disulfide isomerase. *Antioxid. Redox Signaling*, **19**, 36-45
27. McDonough, M.A., Li, V., Flashman, E., Chowdhury, R., Mohr, C., Liénard, B.M.,

- Zondlo, J., Oldham, N.J., Clifton, I.J., Lewis, J., McNeill, L.A., Kurzeja, R.J., Hewitson, K.S., Yang, E., Jordan, S., Syed, R.S. and Schofield, C.J. (2006) Cellular oxygen sensing: Crystal structure of hypoxia-inducible factor prolyl hydroxylase (PHD2). *Proc. Natl. Acad. Sci. U S A* **103**, 9814-9819
28. Koski, M.K., Hieta, R., Böllner, C., Kivirikko, K.I., Myllyharju, J. and Wierenga, R.K. (2007) The active site of an algal prolyl 4-hydroxylase has a large structural plasticity. *J. Biol. Chem.* **282**, 37112–37123
29. Culpepper, M.A., Scott, E.E. and Limburg, J. (2010) Crystal structure of prolyl 4-hydroxylase from *Bacillus anthracis*. *Biochemistry* **49**, 124-133
30. Longbotham, J.E., Levy, C., Johannissen, L.O., Tarhonskaya, H., Jiang, S., Loenarz, C., Flashman, E., Hay, S., Schofield, C.J. and Scrutton, N.S. (2015) Structure and mechanism of a viral collagen prolyl hydroxylase. *Biochemistry* **54**, 6093-6105
31. Koski, M.K., Hieta, R., Hirsilä, M., Rönkä, A., Myllyharju, J. and Wierenga, R.K. (2009) The crystal structure of an algal prolyl 4-hydroxylase complexed with a proline-rich peptide reveals a novel buried tripeptide binding motif. *J. Biol. Chem.* **284**, 25290–25301
32. Neubauer, A., Neubauer, P. and Myllyharju, J. (2005) High-level production of human collagen prolyl 4-hydroxylase in *Escherichia coli*. *Matrix Biol.* **24**, 59–68
33. Vuori, K., Pihlajaniemi, T., Marttila, M. and Kivirikko, K.I. (1992) Characterization of the human prolyl 4-hydroxylase tetramer and its multifunctional protein disulfide-isomerase subunit synthesized in a baculovirus expression system. *Proc. Natl. Acad. Sci. U.S.A.* **89**, 7467–7470
34. Vuorela, A., Myllyharju, J., Nissi, R., Pihlajaniemi, T. and Kivirikko, K.I. (1997) Assembly of human prolyl 4-hydroxylase and type III collagen in the yeast *Pichia pastoris*: formation of a stable enzyme tetramer requires coexpression with collagen and assembly of a stable collagen requires coexpression with prolyl 4-hydroxylase. *EMBO J.* **16**, 6702-6712
35. Ahmad, M., Hirz, M., Pichler, H. and Schwab, H. (2014) Protein expression in *Pichia*

- pastoris*: recent achievements and perspectives for heterologous protein production. *Appl. Microbiol. Biotechnol.* **98**, 5301-5317
36. Aro, E., Khatri, R., Gerard-O'Riley, R., Mangiavini, L., Myllyharju, J. and Schipani, E. (2012) Hypoxia-inducible factor-1 (HIF-1) but not HIF-2 is essential for hypoxic induction of collagen prolyl 4-hydroxylases in primary newborn mouse epiphyseal growth plate chondrocytes. *J. Biol. Chem.* **287**, 37134–37144
 37. Kivirikko, K.I. and Myllylä, R. (1982) Posttranslational enzymes in the biosynthesis of collagen: intracellular enzymes. *Methods Enzymol.* **82**, 245–304
 38. Blanchet, C.E., Spilotros, A., Schwemmer, F., Graewert, M.A., Kikhney, A.G., Jeffries, C.M., Franke, D., Mark, D., Zengerle, R., Cipriani, F., Fiedler, S., Roessle, M. and Svergun, D.I. (2015) Versatile sample environments and automation for biological solution X-ray scattering experiments at the P12 beamline (PETRA III, DESY). *J. Appl. Crystallogr.* **48**, 431-443
 39. Konarev, P.V., Petoukhov, M.V., Volkov, V.V. and Svergun, D.I. (2006) ATSAS 2.1 program package for small-angle scattering data analysis. *J. Appl. Crystallogr.* **39**, 277–286
 40. Majava, V., Wang, C., Myllykoski, M., Kangas, S.M., Kang, S.U., Hayashi, N., Baumgärtel, P., Heape, A.M., Lubec, G. and Kursula, P. (2010) Structural analysis of the complex between calmodulin and full-length myelin basic protein, an intrinsically disordered molecule. *Amino Acids* **39**, 59–71
 41. Konarev, V.V., Volkov, A.V., Sokolova, A.V., Koch, M.H.J. and Svergun, D.I. (2003) PRIMUS – a windows-PC based system for small-angle scattering data analysis. *J. Appl. Crystallogr.* **36**, 1277-1282
 42. Svergun, D.I. (1992) Determination of the regularization parameter in indirect-transform methods using perceptual criteria. *J. Appl. Crystallogr.* **25**, 495-503
 43. Volkov, V.V. and Svergun, D.I. (2003) Uniqueness of ab-initio shape determination in small-angle scattering. *J. Appl. Crystallogr.* **36**, 860-864

44. Petaukhov, M.V., Franke, D., Shkumato, A.V., Tria, G., Kikhney, A.G., Gajda, M., Gorba, C., Mertens, H.D.T., Konarev, P.V. and Svergun, D.I. (2012) New developments in the ATSAS program package for small-angle scattering data analysis. *J. Appl. Crystallogr.* **45**, 342–350
45. Petoukhov, M.V. and Svergun, D.I. (2005) Global rigid body modeling of macromolecular complexes against small-angle scattering data. *Biophys. J.* **89**:1237-1250
46. Svergun, D., Barberato, C. and Koch, M. (1995) CRY SOL – a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. Crystallogr.* **28**, 768–773
47. Eswar, N., Webb, B., Marti-Renom, M.A., Madhusudhan, M.S., Eramian, D., Shen, M.Y., Pieper, U. and Sali, A. (2007) Comparative protein structure modeling with MODELLER. *Curr. Protoc. Protein Sci.* **2**:, doi: 10.1002/0471140864.ps0209s50
48. Drozdetskiy, A., Cole, C., Procter, J. and Barton, G.J. (2015) JPred4: a protein secondary structure prediction server. *Nucl. Acids Res.* doi: 10.1093/nar/gkv332
49. Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N. and Sternberg, M.J.E. (2015) The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols* **10**: 845-858
50. John, D.C.A. and Bulleid, N.J. (1994) Prolyl 4-hydroxylase: defective assembly of α -subunit mutants indicates that assembled α -subunits are intramolecularly disulfide bonded. *Biochemistry* **33**, 14018–14025
51. Lumb, R.A. and Bulleid, N.J. (2002) Is protein disulfide isomerase a redox-dependent molecular chaperone? *EMBO J.* **21**, 6763-6770
52. Tian, G., Xiang, S., Noiva, R., Lennarz, W.J. and Schindelin, H. (2006) The crystal structure of yeast protein disulfide isomerase suggests cooperativity between its active sites. *Cell* **124**, 61–73
53. Tian, G., Kober, F.-X., Lewandrowski, U., Sickmann, A. and Lennarz, W.J. (2008) The catalytic activity of protein-disulfide isomerase requires a conformational flexible

- molecule. *J. Biol. Chem.* **283**, 33630–33640
- 54.** Wang, C., Yu, J., Huo, L., Wang, L., Feng, W. and Wang, C.C. (2012) Human protein-disulfide isomerase is a redox-regulated chaperon activated by oxidation of domain a'. *J. Biol. Chem.* **287**, 1139–1149
- 55.** Dong, G., Wearsch, P.A., Peaper, D.R., Cresswell, P. and Reinisch, K.M. (2009) Insights into MHC class I peptide loading from the structure of the tapasin-ERp57 thiol oxidoreductase heterodimer. *Immunity* **30**, 21-32
- 56.** Koivunen, P., Salo, K.E., Myllyharju, J. and Ruddock, L.W. (2005) Three binding sites in protein-disulfide isomerase cooperate in collagen prolyl 4-hydroxylase tetramer assembly. *J. Biol. Chem.* **280**, 5227–5235
- 57.** Kedersha, N.L., Tkacz, J.S. and Berg, R.A. (1985) Characterization of the oligosaccharides of prolyl hydroxylase, a microsomal glycoprotein. *Biochemistry* **24**, 5952-5960
- 58.** de Jong, L., van der Kraan, I. and de Waal, A. (1991) The kinetics of the hydroxylation of procollagen by prolyl 4-hydroxylase. Proposal for a processive mechanism of binding of the dimeric hydroxylating enzyme in relation to the high k_{cat}/K_m ratio and a conformational requirement for hydroxylation of -X-Pro-Gly- sequences. *Biochim. Biophys. Acta* **1079**, 103-111
- 59.** Breyer, W.A. and Matthews, B.W. (2001) A structural basis for processivity. *Protein Sci.* **10**, 1699–1711
- 60.** Beckham, G.T., Matthews, J.F., Bomble, Y.J., Bu, L., Adney, W.S., Himmel, M.E., Nimlos, M.R. and Crowley, M.F. (2010) Identification of amino acids responsible for processivity in a family I carbohydrate-binding module from a fungal cellulose. *J. Phys. Chem. B.* **114**, 1447–1453
- 61.** Berg, R.A., Kishide, Y., Sakakibara, S. and Prockop, D.J. (1977) Hydroxylation of (Pro-Pro-Gly)₅ and (Pro-Pro-Gly)₁₀ by prolyl hydroxylase. Evidence for an asymmetric active site in the enzyme. *Biochemistry* **16**, 1615–1621

- 62.** Hieta, R., Kukkola, L., Permi, P., Pirilä, P., Kivirikko, K.I., Kilpeläinen, I. and Myllyharju, J. (2003) The peptide-substrate-binding domain of human collagen prolyl 4-hydroxylases. Backbone assignments, secondary structure, and binding of proline-rich peptides. *J. Biol. Chem.* **278**, 34966–34974
- 63.** Kukkola, L., Koivunen, P., Pakkanen, O., Page, A.P. and Myllyharju, J. (2004) Collagen prolyl 4-hydroxylase tetramers and dimers show identical decreases in K_m values for peptide substrates with increasing chain length: mutation of one of the two catalytic sites in the tetramer inactivates the enzyme by more than half. *J. Biol. Chem.* **279**, 18656-18661

FIGURE LEGENDS

Figure 1

The catalytic reaction and the domain structure of human C-P4H-I. **(a)** C-P4H requires Fe^{2+} , 2-oxoglutarate (2OG), and O_2 for its reaction. 2OG is stoichiometrically decarboxylated during hydroxylation, with one atom from molecular oxygen being incorporated into 2OG, producing succinate and CO_2 , whereas the other oxygen atom is incorporated as the 4(R)-hydroxy group of the peptidyl hydroxyproline product. **(b)** Domain structure of the α and PDI/ β subunits of human C-P4H-I. The active site residues of the α subunit important for the prolyl 4-hydroxylation reaction, as well as the catalytic cysteines of the PDI/ β subunit, are indicated. The two proposed intrachain disulfide bridges of the α subunit are also shown. The linker regions between the PSB and CAT domains of the α subunit, and those between the **b'** and **a'** domains of the β subunit, are labeled as L and X, respectively. The N terminal ER targeting signal sequence is not included in the numbering scheme of the subunits.

Figure 2

Sequence alignment of the human C-P4H $\alpha(I)$ – $\alpha(III)$ subunits (Hs- $\alpha(I-III)$) and the C-P4H $\alpha(I)$ subunits of chicken (Gg- $\alpha(I)$), fruit fly (Dm- $\alpha(I)$), zebrafish (Dr- $\alpha(I)$), and *C. elegans* (PHY-1, Ce- $\alpha(I)$). The secondary structure elements of the N terminal dimerization domain (orange) and the middle PSB domain (magenta) are based on the crystal structure of the DD of human C-P4H $\alpha(I)$ subunit (PDB code 4BTB) and are indicated above the sequences. For residues 239–308 (the linker region) of the human C-P4H $\alpha(I)$ subunit (green dashed line), there is no homologous model. The alignment is further supplemented with the Cr-P4H-1 sequence, which shows high similarity with the C terminal CAT domain of the C-P4H α subunits. The secondary structure elements of Cr-P4H-1 (red) are shown above its sequence. The predicted secondary structure elements for the CAT domain of C-P4Hs are indicated by dashed shapes. The two α helices of the CAT domain, $\alpha 13$ and $\alpha 14$, correspond to $\alpha 1$ and $\alpha 2$ of the Cr-P4H-1 structure, respectively. The eight β strands predicted to form the jelly-roll core fold of the CAT domain are numbered using Roman numerals I–VIII. The $\beta 5$ – $\beta 6$ loop and the βII – βIII loop correspond to the flexible $\beta 3$ – $\beta 4$ and βII – βIII active site loops of Cr-P4H-1 (highlighted by a dashed line above the sequences). These two loops are important in forming the closed peptide substrate-binding cavity of the CAT domain. The residues of the CAT domain binding Fe^{2+} and 2OG are highlighted with red arrowheads and black squares, respectively, whereas the residues predicted to be important for the peptide-substrate binding are indicated by light blue dots. The disulfide bond-forming cysteine residues (Cys276, Cys293, Cys486, Cys511) and the glycosylation sites (Asn96, Asn242) are also labeled. The main truncation sites (Asp122 in $\alpha 5$ and the Leu-rich region near Leu210 in $\alpha 10$) are indicated by a lightning bolt.

Figure 3

3D solution structures of the C-P4H-I tetramer and its truncated variants. (a) X-ray scattering curves, (b) distance distribution functions, and 3D *ab initio* models of (c) eC-P4H-I($\beta\alpha\alpha\beta$),

(d) yC-P4H-I($\beta\alpha\alpha\beta$), (e) eC-P4H-I($\beta\alpha\alpha'$) and (f) yC-P4H-I($\beta\alpha'$). The shape information for the eC-P4H-I and yC-P4H-I tetramers is shown in two different orientations. The first view is down the 2-fold axis, and in the second view, the 2-fold axis points vertically after 90° rotation about the horizontal axis. All *ab initio* models are shown on the same scale.

Figure 4

SAXS analysis concerning the effect of peptide binding to eC-P4H-I. (a) X-ray scattering curves and (b) distance distribution functions of eC-P4H-I($\beta\alpha\alpha\beta$) complexed with Zn^{2+} and PDC (green), with Zn^{2+} , PDC, and 300 μM (PPG)₈ substrate (red), and with Zn^{2+} , PDC, and 300 μM P24 inhibitor (black). (c) Dimensionless Kratky plot based on the SAXS data from eC-P4H-I($\beta\alpha\alpha\beta$) (green) and eC-P4H-I($\beta\alpha\alpha\beta$)- Zn^{2+} -PDC complex (green), eC-P4H-I($\beta\alpha\alpha\beta$) complexed with (PPG)₈ substrate (red) and P24 inhibitor (black). Also included are yC-P4H-I($\beta\alpha\alpha\beta$) (dark blue), eC-P4H-I($\beta\alpha\alpha'$) (orange), and yC-P4H-I($\beta\alpha'$) (light blue). The black cross marks the theoretical position of a maximum for a perfectly globular particle.

Figure 5

Ab initio models and superimposed hybrid models of unliganded C-P4H-I($\beta\alpha\alpha\beta$) and its truncated variants. Shown are (a) yC-P4H-I($\beta\alpha'$), (b) eC-P4H-I($\beta\alpha\alpha\beta$), (c) yC-P4H-I($\beta\alpha\alpha\beta$), and (d) eC-P4H-I($\beta\alpha\alpha'$). The *ab initio* models (all on the same scale) are shown as transparent surfaces (gray) and the superimposed hybrid models as cartoons. The first view (left) visualizes the hybrid model looking down the DD dimer 2-fold axis. The second view (middle) is obtained from the first view by 90° rotation around *x*. The hybrid models are colored as follows: dimerization domain, orange; PSB domain, magenta; CAT domain, red; PDI/ β subunit, blue. The chain-like models are colored as: αMF sequence of the PDI/ β , cyan (only in a); PSB-CAT linker region, green; C terminus of CAT, black; C terminus of PDI,

light blue. On the right are the corresponding schematic models of each variant (view down the 2-fold axis). Each subunit is colored differently in these schematic models: α_1 , magenta; α_2 , orange; β_1 , light blue; β_2 dark blue. In panels **(b)**, **(c)**, and **(d)**, the hybrid model calculated using eC-P4H-I($\beta\alpha\alpha\beta$) is used. In panel **(d)**, the CAT domain of the second α subunit as well as the whole β_2 subunit are removed, as they are not present in this truncated variant. **(e)** The fit between the experimental scattering curve of yC-P4H-I($\beta\alpha'$) and its hybrid model (same as in panel **a**). **(f)** The fit between the experimental scattering curve of eC-P4H-I($\beta\alpha\alpha\beta$) and its hybrid model (same as used in panels **b** and **c**). **(g)** The fit between the experimental scattering curve of yC-P4H-I($\beta\alpha\alpha\beta$) and the hybrid model of eC-P4H-I($\beta\alpha\alpha\beta$) (as shown in panels **b** and **c**). **(h)** The fit between the experimental scattering curve of eC-P4H-I($\beta\alpha\alpha''$) and the truncated hybrid model of eC-P4H-I($\beta\alpha\alpha\beta$) (as shown in panel **d**).

Figure 6

Schematic model for the domain assembly of the C-P4H-I $\beta\alpha\alpha\beta$ heterotetramer. The predicted mode of binding of a partially hydroxylated procollagen peptide is also shown. This topology is such that the unhydroxylated peptide forms an initial enzyme-substrate complex with the PSB domain. The region at its N terminal end extends to the catalytic site (black asterisk) and becomes hydroxylated. “X” refers to the variable region of the procollagen chain, as defined in Table 3. The model visualizes also the proposed disulfide bonds Cys276–Cys293 and Cys486–Cys511 in the α subunit [15,50]. The active sites of the PDI/ β subunits are indicated by red asterisks. The location of the known glycosylation sites of the α subunits are marked as gray polygons. The label CC identifies the N terminal coiled-coil dimerization domain.

Table 1. Characterization of the C-P4H-I variants by SEC-SLS.

variant	SLS: Mw (kDa) / polydispersity (Mw/Mn) ¹	SEC: elution volume (ml)
eC-P4H-I($\beta\alpha\alpha\beta$)	250 (3%) / 1.001 (4%)	9.94
eC-P4H-I($\beta\alpha\alpha'$)	148 (2%) / 1.001 (3%)	11.12
iC-P4H-I($\beta\alpha\alpha\beta$)	251 (0.8%) / 1.015 (1%)	9.90
iC-P4H-I($\beta\alpha'$)	104 (0.7%) / 1.004 (1%)	12.15
yC-P4H-I($\beta\alpha\alpha\beta$)	-	9.35
yC-P4H-I($\beta\alpha'$)	100 (2%) / 1.000 (3%)	11.50
PDI ²	-	12.58

¹Mw is the mass weighted molecular weight and Mn refers to the number weighted molecular weight. For a perfect monodisperse sample, the Mw/Mn ratio should be 1. The error estimates are given in parentheses.

²PDI is found in the eC-P4H-I purification (3rd peak in Supplementary Figure S1), being a dissociation product of the C-P4H-I tetramer.

Table 2. Summary of the SAXS analysis of C-P4H-I

sample	peptide	Mw ¹ (kDa)	R _g ² (Å)	D _{max} ³ (Å)
eC-P4H-I(β α α β)	-	305	81.0	290
eC-P4H-I (β α α '')	-	110	53.0	220
yC-P4H-I(β α α β)	-	270	83.5	275
yC-P4H-I (β α '')	-	103	38.2	140
eCP4H-I(β α α β) ⁴	-	324	82.3	290
eC-P4H-I(β α α β) ⁴	100 μ M (PPG) ₈	283	76.9	270
eC-P4H-I(β α α β) ⁴	300 μ M (PPG) ₈	280	78.3	270
eC-P4H-I(β α α β) ⁴	100 μ M (P24)	263	74.9	270
eC-P4H-I(β α α β) ⁴	300 μ M (P24)	269	74.9	270



¹The molecular weights were calculated from the observed forward scattering intensity.

²The radius of gyration, R_g, was calculated from the Guinier plot.

³The maximum distance, D_{max}, was estimated from the distance distribution function.

⁴The peptide-binding studies were done in the presence of 50 μ M Zn²⁺ and 60 μ M PDC.

Table 3. Peptide affinity data and the expected preferred mode of binding of substrates and inhibitors for the human C-P4H-I tetramer.

				catalytic centre of the CAT domain ¹										anchoring prolines in the PSB domains ¹								
																						
Peptide	inhibitor/substrate	K _m (μM)	IC50 (μM)	C1	C2	C3	C4	C5	C6	C7	C8	C9	X ²	P1	P2	P3	P4	P5	P6	P7	P8	P9
(PPG) ₁₀	good substrate	15-20 ³	.	(Pro- Pro -Gly)-(Pro- Pro -Gly)-(Pro- Pro -Gly)-(Pro- Pro -Gly)									- (Pro- Pro -Gly)-(Pro- Pro -Gly)-(Pro- Pro -Gly)-(Pro- Pro -Gly)...									
(PPG) ₅	substrate	170 ⁴		(Pro- Pro -Gly)-(Pro- Pro -Gly)-(Pro- Pro -Gly)									- (Pro- Pro -Gly)-(Pro- Pro -Gly)									
(PPG) ₃	very weak substrate	>2000 ⁵		(Pro- Pro -Gly)-(Pro- Pro -Gly)-(Pro- Pro -Gly)																		
(PPG) ₂	very weak substrate	2600 ⁴		(Pro- Pro -Gly)-(Pro- Pro -Gly)																		
(PPG) ₂ (PPP) ₂ P	substrate	1029 902	2000 ⁶ 2000	(Pro- Pro -Gly)-(Pro- Pro -Gly)									- Pro-Pro-Pro-Pro-Pro-Pro-Pro									
P(PPP) ₂ (PPG) ₂	weak inhibitor												Pro-Pro-Pro-Pro-Pro-Pro-Pro-(Pro- Pro -Gly)-(Pro- Pro -Gly)									
(PPG) ₃ (PPP) ₃	substrate	171 181		(Pro- Pro -Gly)-(Pro- Pro -Gly)-(Pro- Pro -Gly)									- Pro-Pro-Pro-Pro-Pro-Pro-Pro-Pro-Pro									
(PPP) ₃ (PPG) ₃	weak substrate	573 2525		Pro-Pro-Pro-Pro-Pro-Pro-Pro-Pro-Pro-(Pro- Pro -Gly)-(Pro- Pro -Gly)-(Pro- Pro -Gly)																		
(PPG) ₄ (PPP) ₄	substrate	84 124	428	(Pro- Pro -Gly)-(Pro- Pro -Gly)-(Pro- Pro -Gly)-(Pro- Pro -Gly)									- Pro-Pro-Pro-Pro-Pro-Pro-Pro-Pro-Pro-Pro-Pro									
(PPP) ₄ (PPG) ₄	weak substrate	265	-Pro-Pro-Pro-Pro-Pro-Pro-Pro-Pro-Pro-(Pro- Pro -Gly)-(Pro- Pro -Gly)-(Pro- Pro -Gly)									- (Pro- Pro -Gly)									
P9	weak inhibitor												Pro-Pro-Pro-Pro-Pro-Pro-Pro-Pro-Pro-Pro									

¹ From crystallographic binding studies [20,31] it is known that both the catalytic binding pocket (C) and the PSB binding groove (P) have nine binding positions. Position C5 identifies the proline that is hydroxylated by the catalytic centre. Positions P5 and P8 of the PSB domain correspond to the inward pointing proline residues that are deeply buried in the PSB binding groove.

²X identifies the unknown number of residues of the procollagen chain located between the 9-residue peptides that bind to the catalytic site and the PSB binding pocket, respectively.

³References [21, 22, 62]

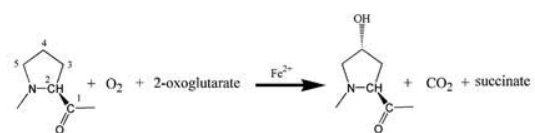
⁴Reference [62]

⁵Too high to be accurately measured

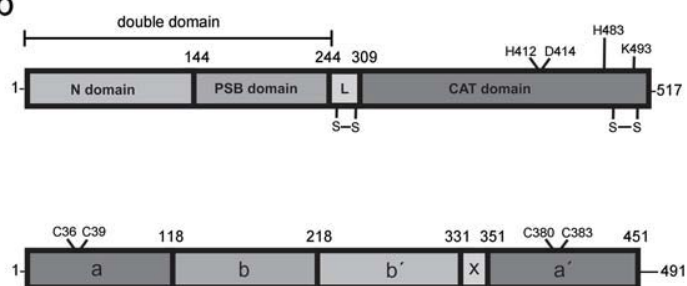
⁶Using (PPG)₁₀ as a substrate (when using (PPG)₃ as substrate: 28% inhibition at 2 mM P(PPP)₂(PPG)₂)

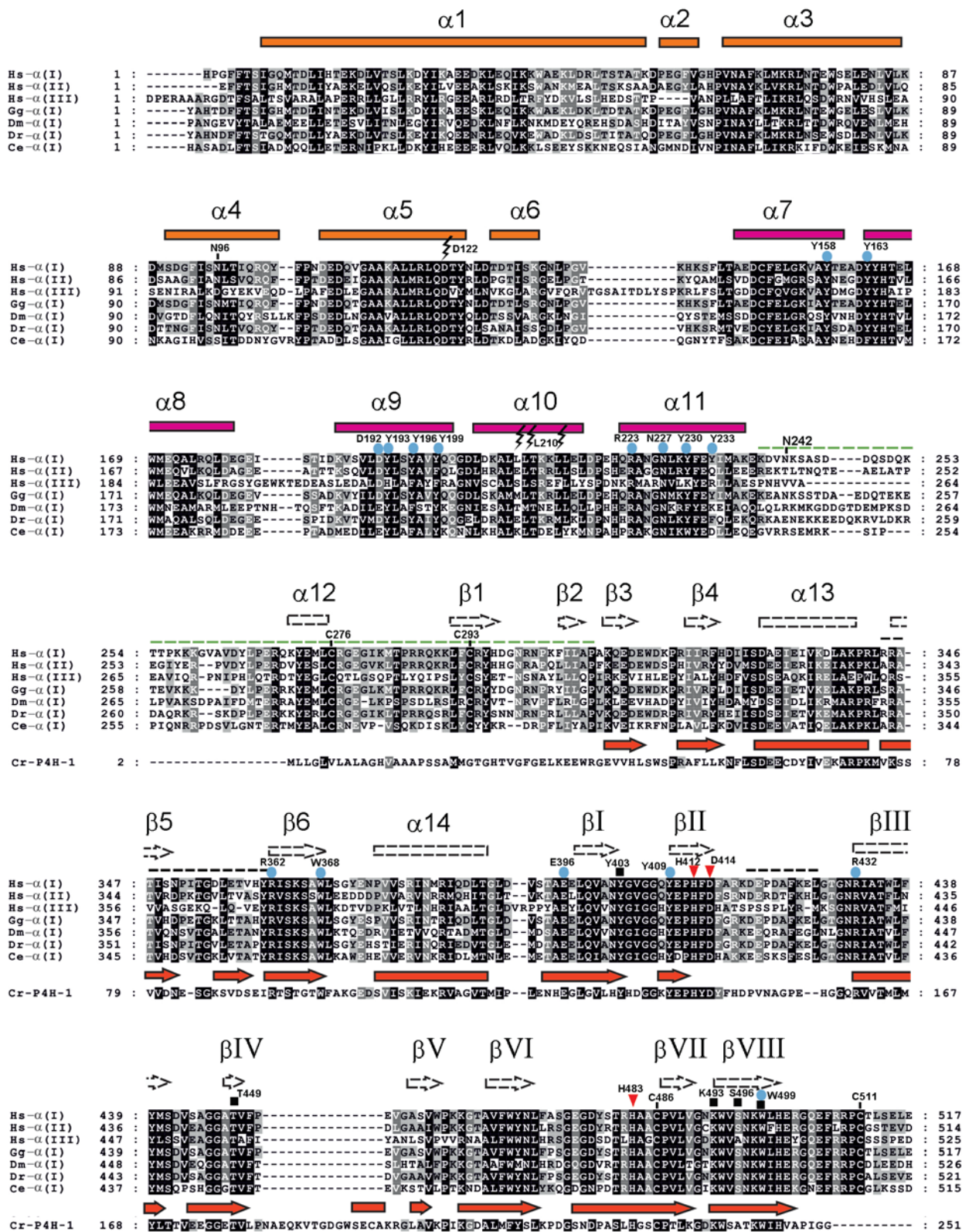
⁷ Using (PPG)₁₀ as substrate (when using (PPG)₃ as a substrate: no inhibition at 2 mM P9).

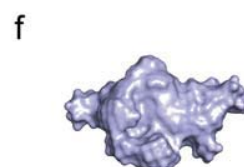
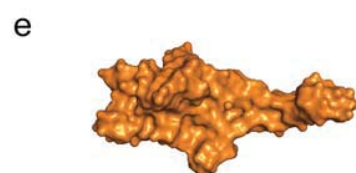
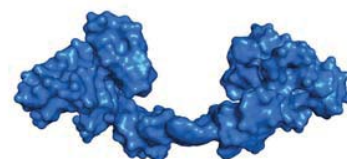
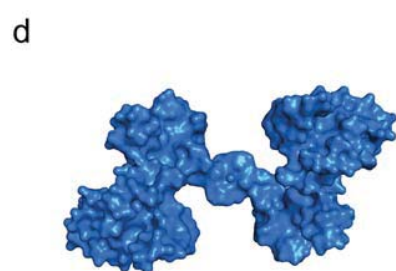
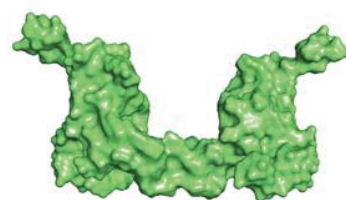
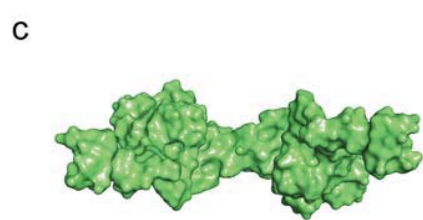
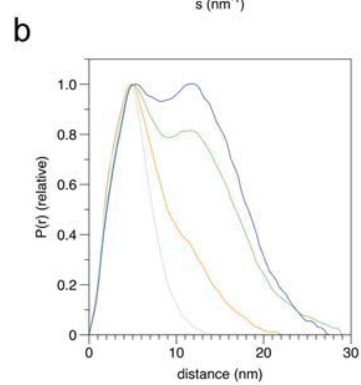
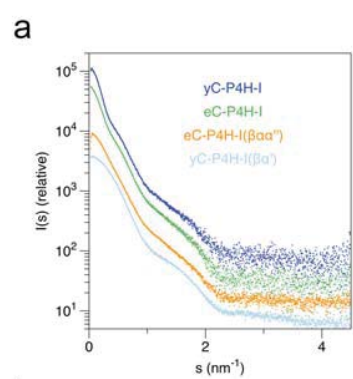
a

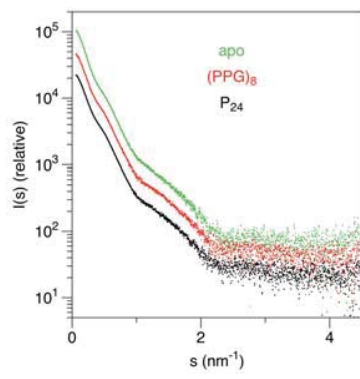
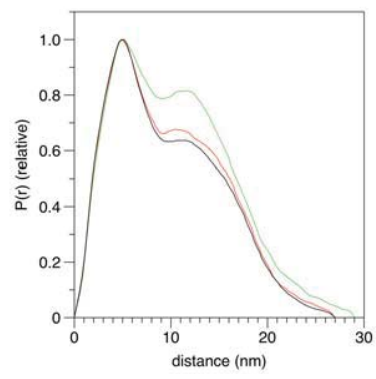


b







a**b****c**