The Crystal Structure of the Inhibitor-complexed Carboxypeptidase D Domain II and the Modeling of Regulatory Carboxypeptidases*

Received for publication, December 19, 2000, and in revised form, February 14, 2001
Published, JBC Papers in Press, February 14, 2001, DOI 10.1074/jbc.M011457200

Patrick Aloy‡§, Verónica Companyṣ‡§, Josep Vendrell‡, Francesc X. Aviles‡§, Lloyd D. Fricker**i, Miquel Coll‡‡, and F. Xavier Gosim-Rüth‡‡

From the ‡Institut de Biologia Fonamental and Departament de Bioquímica i Biologia Molecular, Unitat de Ciències, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain, †Department of Molecular Pharmacology, Albert Einstein College of Medicine, Bronx, New York 10461, and ‡‡Institut de Biologia Molecular de Barcelona, Centre d’Investigació i Desenvolupament-Consejo Superior de Investigaciones Científicas, Jordi Girona, 18-26, 08034 Barcelona, Spain

The three-dimensional crystal structure of duck carboxypeptidase D domain II has been solved in a complex with the peptidomimetic inhibitor, guanidinoethylmercaptosuccinic acid, occupying the specificity pocket. This structure allows a clear definition of the substrate binding sites and the substrate funnel-like access. The structure of domain II is the only one available from the regulatory carboxypeptidase family and can be used as a general template for its members. Here, it has been used to model the structures of domains I and III from the former protein and of human carboxypeptidase E. The models obtained show that the overall topology is similar in all cases, the main differences being local and because of insertions in non-regular loops. In both carboxypeptidase D domain I and carboxypeptidase E slightly different shapes of the access to the active site are predicted, implying some kind of structural selection of protein or peptide substrates. Furthermore, emplacement of the inhibitor structure in the active site of the constructed models showed that the inhibitor fits very well in all of them and that the relevant interactions observed with domain II are conserved in domain I and carboxypeptidase E but not in the non-active domain III because of the absence of catalytically indispensable residues in the latter protein. However, in domain III some of the residues potentially involved in substrate binding are well preserved, together with others of unknown roles, which also are highly conserved among all carboxypeptidases. These observations, taken together with others, suggest that domain III might play a role in the binding and presentation of proteins or peptide substrates, such as the pre-S domain of the large envelope protein of duck hepatitis B virus.

Carboxypeptidases (CPs) are enzymes that catalyze the cleavage of C-terminal peptide bonds in proteins and peptides. From a mechanistic point of view, CPs can be classified in two groups, metalloCPs and serine CPs. The metalloCPs possess a Zn²⁺ cofactor in the active site. In mammals, this family currently contains 13 members subdivided into two subfamilies, the digestive enzymes and the regulatory enzymes (1–5). Whereas the biological function of the digestive CPs is to contribute to protein degradation, the regulatory ones are generally involved in physiological processes that require a higher specificity. Within each group, the members have 25–63% amino acid sequence identity, but it decreases to only 15–25% when comparison is performed between subfamilies. This low overall homology between subfamilies implies that they diverged early in time.

The digestive CPs are soluble, non-glycosylated proteins that are synthesized as inactive precursors containing a 90–95-amino acid N-terminal pro-segment (5, 6). The regulatory CPs have been purified and characterized from biological fluids and tissues, where they are found in soluble or in membrane-attached forms, in minor quantities. This subfamily includes CPD, CPE, CPM, CPN, CPZ, and novel proteins with an unknown function designated adipocyte enhancer-binding protein 1, CPX-1, and CPX-2 (3, 7–11). These proteins perform a variety of important physiological functions, including neuropeptide and prohormone processing, regulation of peptide hormone activity, and alteration of protein-protein or protein-cell interactions (2, 3, 12).

CPE, also known as enkephalin convertase or CPH (EC 3.4.17.10), is a CPB-like enzyme associated with the biosynthesis of many peptide neurotransmitters and hormones. It was purified for the first time from bovine brain (13). Later, cDNAs corresponding to CPEs from cattle (14), rat (15, 16), human (17), Aplysia californica (18), and the fish Lophius americanus (19) were cloned and sequenced. The amino acid sequence homology among vertebrate species is greater than 80%. The molecular mass of CPE is 55 kDa, and it is formed by 476 residues, of which 25 correspond to the signal peptide, and 17 correspond to the pro-segment. However, and in contrast with digestive CPs, scission of the pro-segment is not necessary for expression of the activity (20). Also, in contrast with the great majority of metalloCPs, whose optimum pH value is around neutrality, CPE has its maximum activity at an acidic pH value, between 5 and 5.5 (21), coincident with the internal pH of neutrality, CPE has its maximum activity at an acidic pH value, between 5 and 5.5 (21), coincident with the internal pH of the large envelope protein of duck hepatitis B virus.

* This work was supported in part by Grants BIO98-0362, BIO2000-1659, PB98-1631, and 2FD97-0518 from the Ministerio de Educación y Cultura (Spain), by Grant 1999SGR-188 and the Center de Referència en Biotecnologia (both from the Generalitat de Catalunya), by Grants DA-00194 and DK-51271 from the National Institutes of Health, and by the United States-Spain Science and Technology Collaborative Program, 1999. V. C. is a predoctoral fellow of the Universitat Autònoma de Barcelona, and P. A. is a postdoctoral fellow of the Ministerio de Educación y Cultura (Spain). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

‡ First author.
‡‡ To whom correspondence may be addressed: Institut de Biologia Fonamental, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain. Tel.: 34-93-581-1315; Fax: 34-93-581-2011; E-mail: fricker@aecom.yu.edu.

The abbreviations used are: CP(s), carboxypeptidase(s); GEMSA, guanidinoethylmercaptosuccinic acid; 3D, three-dimensional; RMSD(s), root mean square deviation(s).

This paper is available on line at http://www.jbc.org
value of the secretory granules. It has also been shown that its activity is regulated by the presence of Co^{2+} (1). Several analogs of arginine and lysine, which were originally designed as active site-directed inhibitors of CPB and CPN, were found to be potent inhibitors of CPE (13). Two of these compounds, guanidinoethylercaptosuccinic acid (GEMSA) and aminopropylmercaptosuccinic acid, are several hundred-fold more potent as inhibitors of CPE than of either CPB or CPN (13).

It has been shown that mice with the mutation Cpe^{fat}/Cpe^{fat} have deficient proinsulin processing because of the absence of CPE activity in the pancreatic islets and the pituitary, caused by a point mutation S202P (22). Mice containing such mutations in the Cpe gene also show a reduced ability to process other hormones (23). However, the observation that Cpe^{fat}/Cpe^{fat} mice are still able to process a small quantity of insulin suggested that another CP was also involved in peptide processing.

A search for additional CPE-like enzymes led to the discovery of CPD (EC 3.4.17.22) (8). CPD is a 180-kDa protein containing a signal peptide, three CPE-like domains of ~390 residues separated by short bridge regions, a transmembrane domain, and a 60-residue C-terminal cytosolic tail (24–26). The cDNAs corresponding to CPD of human (25), rat (26), mouse (27), duck (24), Drosophila melanogaster (28), and A. californica (29) have been cloned. All species contain three CPE-like domains (here named CPD-I, CPD-II, and CPD-III), suggesting that their distinct physiological functions are important. The characterization of the first and second domains of CPD has shown that both possess catalytic activity and have somewhat complementary activities. Specifically, the first domain is optimally active at pH 6.3–7.5 and prefers substrates with C-terminal Arg, whereas the second domain is optimally active at pH 5.0–6.5 and prefers substrates with C-terminal Lys (30). In contrast, the third domain is inactive toward a variety of standard CP substrates (30, 31). Duck CPD, also named gp180, was identified by its ability to bind the pre-S domain of the large envelope protein of duck hepatitis B virus particles (24). A comparison of human and duck CPD reveals 66, 83, and 82% amino acid sequence identity among the first, second, and third CP repeats, respectively. Recent studies with mutants lacking the first, second, or third CP-like domains have shown that the third domain of duck CPD is responsible for binding to the pre-S domain of the large envelope protein of hepatitis B virus and that this binding does not require CP activity (31). Despite the absence of activity in the third domain, the fact that it is highly conserved among duck and mammals suggests the existence of a biological function for it.

Crystallization of both CPE and the complete three-domain CPD has been attempted. However, low yields in the protein recovery and the occurrence of glycosylations, together with the fact that the interdomain linker peptides in CPD are probably highly flexible, have precluded direct 3D structure determination. The only crystal structure from the regulatory metalloCP subfamily that has been solved is that of the second domain of duck CPD (32). It displays a 300-residue N-terminal α/β-hydrolase with overall topological similarity to pancreatic CPA. This subdomain is followed by a C-terminal 80-residue β-sandwich subdomain, unique for these regulatory metalloenzymes and topologically related to transthyretin and sugar-binding proteins (32). To further investigate and better define the enzyme substrate pocket and to provide a basis for the rational design of specific inhibitors of regulatory CPs, we have solved the crystal structure of CPD domain II in complex with the peptideimid mimetic inhibitor GEMSA. Based on this structure, overall models and detailed ones of the respective active sites have been built for human CPE and domains I and III of duck CPD.

These models permit hypotheses about the structural basis of enzyme specificity and biological activity.

**EXPERIMENTAL PROCEDURES**

**Crystallization**—Crystals of native CPD-II were produced as mentioned previously (32). The CPD-II-GEMSA complex was obtained by soaking native crystals in a 2.5 M solution of ammonium sulfate, buffered with 0.15 M sodium acetate to pH 5.2, and containing 10 mM GEMSA (purchased from Calbiochem), for 3 days. Diffraction data from 2.6-Å resolution were collected from a single N$_{2}$-cryocooled complex crystal that belongs to the same spacegroup (P2$_1$2$_1$2$_1$) as the native ones at the EMBL synchrotron beamline BW7B (Deutsches Elektronensynchrotron, Hamburg, Germany). Data were processed with MOSFLM, version 6.0.1 (33) and SCALA from the CCP4 suite (Collaborative ComputationalProject 4, 1994). The coordinates of native CPD-II (after removal of solvent molecules and sulfate anion 998 located in the active site cleft; see Ref. 32) were used for initial rigid body refinement. Positional/temperature refinement employing the program CNS, version 0.9 (34) and using maximum likelihood as minimization criterion followed and omit maps (αK-weighted 2F\_obs - F\_calc and F\_obs - F\_calc) were computed. The difference density map clearly revealed the location of the bound inhibitor (Fig. 1) allowing its model building using the Turbo-Frodo program (35). The complex was submitted to further positional/temperature refinement after setup of appropriate inhibitor parameter and topology files. The refinement of the occupancy of the latter revealed 100% presence, in accordance with the very high affinity of the inhibitor (in the nM range). The final model comprises residues Gln^Thr^40 of the deduced sequence, 195 solvent molecules (labeled 601–785), one zinc cation (residue 999), one sulfate anion (998) with partial occupancy, and the 15-atom inhibitor GEMSA (designated Gem801). Three asparaginyl residues were found to be glycosylated (Asn^{Thr}201, Asn^{Ser}221, and Asn^{Ser}277). One peptide bond (Pro^{Thr}Ph^251) has been found in the cis conformation. Table I provides a summary of the data processing and final model refinement. The coordinates of the complex structure have been deposited with the Protein Data Bank (access code 1b51).

**Model Building**—A preliminary multiple alignment was performed by means of the program PILEUP (36, 37) for the three duck CP structural repeats (domains). This alignment was used as a “seed” to build a hidden Markov model profile with the program HMMER (38) that was used to align eight additional homologous sequences. Expert knowledge and experimental information were also used to improve the quality of the alignment in several segments. The primary and 3D structures of duck CPD-II were used as a template to build the models. A segment of 30 residues from CPT (PDB access code 1b0r) was aligned to CPE to model a 25-residues insertion (residues 202–224; see footnote to Table II for the conventions used on numbering the different sequences). Finally, a 25-residues stretch from the sequence of adenovirus coat protein (1dxh) was also aligned to the insertion observed in CPD-I (residues 96–124). Using the multiple alignment for the three cp domains (CPD-I, CPD-II, and CPD-III) and CPE as a starting point, a method of comparative modeling by satisfaction of spatial constraints was used to build the 3D structure of CPD-I, CPD-III, and CPE. This method is implemented in the program MODELLER (39). The spatial constraints are derived by transferring the spatial features from the structures of known proteins to the sequence of the unknown ones. The program PROSA-II (40) was used to check the quality of the models as described in a previous work (41). The regions with non-near-native fold were identified by the high positive values of pseudo-potential energy, independently of the crystallographic structure. Once the three models were built automatically, manual intervention was required for re-modeling those regions identified by PROSA-II with non-native fold. The program FRAZER, developed in our laboratory was used to reconstruct the problematic regions. The overall RMSD calculations and superimposition of the three modeled structures with respect to the crystallographic one (CPD-II) were obtained according to the structural alignment given by the program SSAP (42). The active sites superposition and GEMSA inhibitor replacement in the three models were also performed with FRAZER. The coordinates of the CP models, in PDB format, are available upon request.

**RESULTS**

**Structure of the Complex CPD-II-GEMSA**—The CPD-II polypeptide chain in the complex is folded into two distinct subdomains, a 300-residue catalytic CP subdomain displaying

2 Unpublished information.
the α/β-hydrolase fold reminiscent of the CPA structure and an 80-residue C-terminal subdomain of all-β pre-albumin-like β-sandwich folding topology, the so-called transthyretin subdomain (32). The complex structure displays no significant deviations from the native protein (32), as denoted by a RMSD of 0.19 Å for all Cα atoms. Only the catalytic zinc ion is somewhat moved away (0.7 Å) from its position in the non-complexed domain structure forced by the presence of the inhibitor. This movement is accompanied by a similar displacement (0.7 Å) of one of the coordinating residues, His74. Interestingly, the catalytic solvent molecule (601) attached to the zinc ion in the unliganded structure is moved 2.3 Å away upon inhibitor binding (solvent molecule 684 in the present structure; see Fig. 2). The peptidomimetic GEMSA molecule occupies the primed side of the catalytic active site cleft, emulating a bound C-terminal amino acid after proteolytic cleavage of a substrate. The gua-

Table I

| Data collection, processing, and final model refinement |
|---------------------------------------------------------|
| Wavelength (Å)                                           | 0.8437 |
| Temperature                                              | N2 cryocooling |
| No. of measurements                                      | 362,974 |
| No. of unique reflections                                | 26,360 |
| Statistics                                              |    |
| Whole resolution range (Å)                              | 43.03–2.60 |
| Completeness (%)/Rmerge<sup>a</sup>                     | 100.0/8.3 |
| (I)/(σ(I))/average multiplicity                         | 7.8/13.7 |
| Last resolution shell (Å)                               | 2.74–2.60 |
| Completeness (%)/Rmerge<sup>a</sup>                     | 100.0/47.5 |
| (I)/(σ(I))/average multiplicity                         | 1.6/13.9 |
| Resolution range used for refinement (Å)                | 43.03–2.60 |
| No. of reflections/completeness (%)                      | 26,328/100.0 |
| Final crystallographic R-factor<sup>b</sup> (using all data) | 0.195 |
| R-factor/free R-factor<sup>b</sup> (penultimate cycle)  | 0.199/0.239 |
| Non-hydrogen protein atoms (including glucides and inhibitor) | 3.138 |
| Solvent molecules/ions<sup>c</sup>                      | 195/1 Zn<sup>2+</sup>, 1 SO<sub>4</sub><sup>2−</sup>, Asn<sup>136</sup>, Asn<sup>322</sup>, Asn<sup>377</sup> |
| Glycosylation sites                                      |    |
| r.m.s. deviation for the protein part (including glucides) from target values of |    |
| bonds (Å)                                                | 0.007 |
| angles (°)                                               | 1.34 |
| bonded temperature factors main chain/side chain (Å<sup>2</sup>) | 1.92/2.95 |
| Average B-factors (Å<sup>2</sup>)                        |    |
| Protein atoms                                            | 43.5 |
| Solvent molecules                                        | 48.6 |
| Sulfate anion                                            | 51.3 |
| Carbohydrate atoms (attached to Asn<sup>136</sup>, Asn<sup>321</sup>, and Asn<sup>377</sup>) | 49.3, 62.3, 61.3 |
| Zinc cation                                              | 33.3 |
| Inhibitor atoms                                          | 39.6 |

<sup>a</sup> Rmerge = Σhkli(l(hkl) – (I(hhl))/Σhkli l(hhl).
<sup>b</sup> R-factor = Σhkli|F<sub>obs</sub>| – k|F<sub>calc</sub>|/Σhkli|F<sub>calc</sub>|; free R-factor, same for a test set of 1950 (7%) reflections not used during refinement (till the penultimate cycle). This test set includes the same reflections as the native data set (32), further extended to cover the non-common resolution shell (2.70–2.60 Å).
The nidinoethylmercapto group is reminiscent of a substrate arginine side chain (CPD-II displays a CPB-like preference for basic residues in P1<sup>9</sup>) and occupies the same position in the specificity pocket. It is anchored through its atoms N<sub>h1</sub> and N<sub>h2</sub> to the side chain of Asp<sub>192</sub> and the main chain carbonyls of Gly<sub>246</sub> and Tyr<sub>250</sub>, the latter one present in the "down" conformation as in the native structure (32). This planar guanidinoethylmercapto group establishes an additional van der Waals' interaction (3.8 Å) with Val<sub>252</sub>. The inhibitor carboxylate group mimicking a peptide substrate C terminus is anchored to both Arg<sub>145</sub> and Arg<sub>135</sub>. The second carboxylate group is similar to a scissile carbonium ion in the transition state and coordinates the catalytic zinc ion in a bidentate manner. One of its oxygens is further bonded by Arg<sub>135</sub> and His<sub>74</sub> (see Fig. 2).

**Sequence Alignments, Model Building, and Refinement**—Fig. 3 shows the multiple sequence alignment of the three domains of duck CPD and human CPE. This alignment, performed as indicated under "Experimental Procedures," allows us to derive accurate models for these proteins. The alignment reveals 42% sequence identity between duck CPD-II and CPD-I and 32% sequence identity between duck CPD-II and CPD-III. The percentage identity for human CPE with respect to duck CPD-II is even higher, at 50%. These identity levels allow homology modeling of the three-dimensional structures of these proteins. The alignment in Fig. 3 shows that all the residues experimentally described as important (those at the active site, at the metal binding site, and at the substrate binding subsites) are essentially conserved among all the sequences, except for CPD-III. Several insertions and deletions, however, can be observed in the alignment. First of all, a large insertion (29 residues) in CPD-I can be detected. This inserted stretch is extremely charged, with 5 basic and 15 acidic residues, and the only sequences in the data banks that show a certain level of homology with it belong to proteins that interact with nucleic acid-binding proteins, which are largely unstructured in the absence of an interacting partner (43). In any case, only one of

---

**TABLE II**

| Duck CPD-II<sup>a</sup> | Full-length duck CPDB | Human CPE<sup>c</sup> | CPA<sup>d</sup> | Bovine CPE<sup>d</sup> |
|------------------------|-----------------------|----------------------|-----------------|----------------------|
| His<sup>74</sup>       | His<sup>139</sup>     | His<sup>574</sup>    | His<sup>397</sup> | His<sup>114</sup>    |
| Glu<sup>77</sup>       | Glu<sup>142</sup>     | Glu<sup>577</sup>    | Glu<sup>1000</sup>| Glu<sup>117</sup>    |
| Arg<sup>135</sup>      | Arg<sup>270</sup>     | Arg<sup>634</sup>    | Arg<sup>1058</sup>| Arg<sup>179</sup>    |
| Asn<sup>144</sup>      | Asn<sup>231</sup>     | Asn<sup>643</sup>    | Asn<sup>1058</sup>| Asn<sup>188</sup>    |
| Arg<sup>143</sup>      | Arg<sup>256</sup>     | Arg<sup>643</sup>    | Arg<sup>1067</sup>| Arg<sup>190</sup>    |
| Asn<sup>146</sup>      | Asn<sup>253</sup>     | Asn<sup>645</sup>    | Asn<sup>1069</sup>| Asn<sup>190</sup>    |
| His<sup>181</sup>      | His<sup>260</sup>     | His<sup>650</sup>    | Asp<sup>1105</sup>| His<sup>146</sup>    |
| Gly<sup>183</sup>      | Gly<sup>262</sup>     | Gly<sup>662</sup>    | Gly<sup>1107</sup>| Gly<sup>250</sup>    |
| Asn<sup>185</sup>      | Ser<sup>267</sup>     | Ser<sup>687</sup>    | Thr<sup>1112</sup>| Asp<sup>255</sup>    |
| Asp<sup>133</sup>      | Asp<sup>271</sup>     | Asp<sup>681</sup>    | Asp<sup>1118</sup>| Asp<sup>259</sup>    |
| Tyr<sup>250</sup>      | Tyr<sup>311</sup>     | Tyr<sup>249</sup>    | Asp<sup>1166</sup>| Tyr<sup>310</sup>    |
| Val<sup>252</sup>      | Val<sup>333</sup>     | Val<sup>711</sup>    | His<sup>1170</sup>| Val<sup>322</sup>    |
| Gln<sup>257</sup>      | Gln<sup>338</sup>     | Gln<sup>756</sup>    | Lys<sup>1175</sup>| Gln<sup>327</sup>    |
| Glu<sup>272</sup>      | Glu<sup>353</sup>     | Glu<sup>771</sup>    | Tyr<sup>1190</sup>| Glu<sup>342</sup>    |

<sup>a</sup>The numbering system is based on the construct of CPD-II alone that was expressed in *Pichia pastoris* and used for the determination of the crystal structure (32).

<sup>b</sup>The numbering system is based on the full-length amino acid sequence deduced from the cDNA sequence (24). In this system, amino acid number 1 is assigned to the initiation Met of the signal peptide.

<sup>c</sup>The numbering system is based on the full-length amino acid sequence deduced from the cDNA sequence (17). In this system, amino acid number 1 is assigned to the initiation Met of the signal peptide.

<sup>d</sup>This numbering system is the standard for pancreatic CPs and assigns amino acid number 1 to the first residue of the mature form after removal of the signal peptide and pro-segment.

<sup>e</sup>Residues interacting with the inhibitor GEMSA.

---

**FIG. 2**. Stereo plot of CPD-II in complex with GEMSA as a ball and stick model displaying the hydrogen bond network around the active site. The catalytic zinc ion (999) is displayed as a cyan sphere, the inhibitor moiety is displayed in violet, and the solvent molecules are displayed as red spheres. Residues that conform the active site and interact with the inhibitor are labeled.
these related sequences has its 3D structure determined (adenovirus type 2 hexon, PDB code 1dhx), and the low percentage of identity observed is not sufficient to model this stretch. Another large insertion (23 residues) can be found in human CPE. This insertion is present in all species of CPE, as well as most other members of the CP family, although the length varies from 14–15 residues for CPA, CPB, and the bacterial CP to 27 residues for CPX-1, CPX-2, and adipocyte enhancer-binding protein 1. Because the three-dimensional structure of this loop in CPA, CPB, and CPT is known, this region of CPD can be modeled using the crystal structure and alignment of the other CPs. The rest of indels are much shorter and can be modeled with reasonable confidence by energy optimization.

The pseudo-energies of the original models were calculated with PROSA-II (40) to identify the incorrect chain tracings. As expected, the regions that presented higher energy were those of the indels (data not shown). In the case of the large highly charged insertion in duck CPD-I, the energy tended to be infinite, and consequently, the loop was removed from the model. For the insertion in human CPE, the pseudo-energy was corrected to acceptable values by manual modification of the CPE model and energy minimization.

The overall RMSDs calculations between the 3D structure of CPD-II and the three different models gave the following values: 0.5 Å for CPD-I (once the non-modeled loop was removed), 1.3 Å for CPD-III, and 0.9 Å for CPE. Taking into account that only one crystal structure was used to model the three sequences, the RMSDs correlate well with the percentage of sequence identity obtained in the multiple alignment. The RMSD was also calculated for the three models using only the active site residues. In this case, the results were 0.1 Å for CPD-I and CPE and 0.5 Å for CPD-III.

Fig. 4 shows the modeled 3D-structures of CPD-I, CPD-III, and CPE compared with the crystal structure of CPD-II. The RMSD values indicate that, although a number of local differences are obviously present (discussed below) the models share a common topology, and the relative positions of the two subdomains are maintained in all of them. A close inspection of the models also shows that the major structural features in CPD-II that suggest a different selectivity of regulatory CPs toward large protein substrates as compared with the pancreatic enzymes are also present in the other members of the regulatory family studied here. Thus, those loops in the funnel-like access to the active site, which are probably responsible for the different selectivity of the two families of CPs, conform an opening of the solvent-exposed surface, which, beyond individual characteristics that will be discussed below, is very similar in all cases.

Conserved Interactions between GEMSA and the Different Models—After superimposing the four active sites, the GEMSA inhibitor was emplaced in the three models to find and rationalize its possible interactions with the enzymes. The fit was excellent in all three cases. The residues in the x-ray structure of CPD-II interacting with the GEMSA inhibitor and their equivalents in the three modeled structures are shown in Table II. As can be seen, in CPD-I and CPE all the hydrogen bonds found in the co-crystal are conserved in the complex between the inhibitor and the protein residues. In contrast, in CPD-III several critical interactions are lost because of the different residues found at the active site.

DISCUSSION

Structural Basis of the Inhibitor Action—The CP inhibitor GEMSA has been frequently used as a potent inhibitor of regulatory CPs (CPN, CPE, and CPD). The Ki values determined to date fall in the low nanomolar range, 4 nM for duck CPD-I, 34 nM for duck CPD-II (30), and 8 nM for bovine CPE (13). This is the first time that a crystal structure of its complex with one of these enzymes has been reported. Such structure clearly explains the powerful action of this inhibitor; the catalytic water and the essential Zn$^{2+}$ are both displaced, the latter
one being bound in a bidentate manner by one of the carboxylate groups of the inhibitor. In addition, the inhibitor is bound to residues of CPD-II that are essential for substrate binding and polarization, Tyr$^{259}$, Arg$^{145}$, and Arg$^{135}$. Therefore, several structural elements indispensable for the catalytic action of the enzyme are perturbed or shielded by the inhibitor.

Taking into account the similarity of CPD-II to the modeled structures and the easy way in which GEMSA has been fitted on them, it is expected that the inhibitor binds in a very similar way to CPD-I and CPE. However, it is unlikely that GEMSA binds CPD-III because of the absence of critical residues (discussed below).

**Overall Comparison of the Models**—The derived models of CPD-I, CPD-III, and CPE show an overall similarity with the recently described crystal structure of CPD-II (32). In all models two subdomains are clearly visible, the CP subdomain and the C-terminal subdomain, which shares topological similarity and connectivity with transthyretin (32). The CP subdomain shows the α/β-hydrolase fold common to many proteases from the cysteine, serine, and metalloprotease families. It is formed by a doubly wound eight-stranded β-sheet flanked on both sides by three and six helices, respectively. Meanwhile, the C-terminal subdomain displays a rod-like shape forming a β-barrel or β-sandwich of pre-albumin-like folding topology made up by seven strands connected by short loops. This is valid for all the models studied here.

As in CPD-II, the interactions between both subdomains are mainly of a hydrophobic nature in all models. Most of the van der Waals’ interactions described for CPD-II are also found in CPD-I and CPE. Albeit containing a smaller number of such interactions, CPD-III still conserves the most significant ones. A number of hydrogen bonds also contribute to subdomain interactions in CPD-I, CPD-III, and CPE, most of them being exactly conserved in CPD-I and CPE versus CPD-II, and greater differences being found for CPD-III. It is worth mentioning that the only salt bridge between subdomains described for CPD-II, Asp$^{206}$-Arg$^{243}$, is also conserved in the modeled structures between pairs of Asp/Arg at equivalent positions in CPD-I and CPE and between Glu$^{123}$ and His$^{128}$ at equivalent positions in CPD-III. Also, the disulfide bond in CPD-II between Cys$^{239}$ and Cys$^{775}$ is also predicted in the three models; an additional disulfide between Cys$^{770}$ and Cys$^{132}$ in CPE (already detected from biochemical measurements) is also predicted in the model built for this form.

CPE is the protein with the highest homology to CPD-II and also one whose model has the lowest RMSD value with the experimental 3D structure. However, it should be taken into account that the RMSD value was calculated only on the structurally equivalent residues given by the alignment performed with the program SSAP. This means, for instance, that the 23-residues insertion in CPE, spanning from Glu$^{156}$ to Lys$^{189}$, was not considered. As compared with CPD-II, pancreatic and bacterial CPs also have a 14-amino acid insertion in this region, forming a loop that shapes one side of the entrance to the active site and establishing cross-connections to an adjacent loop. This feature is considered to be one of the distinctive determinants of specificity between regulatory and pancreatic CPs. The 23 extra residues in CPE form a turn-rich region rather exposed to the solvent in the model built, according to the well defined structure of this loop in *Thermoactinomyces vulgaris* CPT (1obr).

The main difference between CPD-I and CPD-II is the above mentioned long insertion of 29 residues in CPD-I that contains 20 net charges. This sequence was eliminated from the calculations as no homologous sequences and 3D structures were found to model it with a sufficient degree of confidence. A further significant difference is a glycine-rich insertion of nine residues in one of the loops that shape the active site entrance (residues 165 to 173 in Fig. 1). This insertion does not generate a substantial change in the surface of the active site cleft in our model and is folded inwards over the molecular body of the enzyme.

The study of the model of CPD-III is particularly important because of its lack of enzymatic activity (31), probably because of the absence of key residues for CP catalysis. However, alignment of the sequences and superimposition of the 3D structures shows that other residues with yet unknown function are highly conserved. When comparing the models and the structure, three categories of residues can be defined. The first one is formed by the residues essential for catalysis. In CPD-II, these essential residues are His$^{74}$, Glu$^{77}$, and His$^{181}$ (coordinators of Zn$^{2+}$), Glu$^{272}$, and Arg$^{135}$. Only the first His is conserved in CPD-III, whereas the other residues are replaced by Ala, Asp, Tyr, and His, respectively. The enzymatic machinery of CPD-III is therefore disabled, because neither proper coordinators of Zn$^{2+}$ nor a general base or a polarizing residue are present (6), respectively.

Those residues that are necessary for substrate binding are
included in the second category. The triad Asn^{144}, Arg^{145}, and Asn^{154}, which is responsible for the anchoring of the C-terminal carboxylate (COO⁻) of the substrate, is generally conserved in all CPs that are enzymatically active toward peptides, including the pancreatic and bacterial CPs (Table II). In CPD-III, this triad is replaced by Asp, Thr, and Asp, rendering a domain that has lost the ability to anchor the carboxyl group. Interestingly, a peptidase in the bacterium Bacillus sphaericus is a distant member of the metalloCP family that also lacks this Asn-Arg-Asn triad (44). Instead of cleaving substrates with a C-terminal carboxylate group, as in other CPs, the B. sphaericus peptidase hydrolyzes C-terminal meso-diaminopimelic acid. This substrate has an amino group in place of the carboxylate of a typical peptide, consistent with the replacement of the Asn-Arg-Asn with an Asn-Asp-Gln. Thus, the differences in this sequence between CPD-II and CPD-III are predicted to be critical for defining the binding specificity of each domain.

Some other residues necessary for substrate binding in CPD-II are also different in CPD-III. For example, Tyr^{250}, Val^{252}, and Gly^{255} of CPD-II are replaced by His, His, and Ser, respectively, in CPD-III. However, despite these replacements, when the model of CPD-III and the crystal structure of CPD-II are superimposed, it can be observed that the different residues in CPD-III occupy exactly the same position of their homologues in CPD-II. Taken together, it is unlikely that CPD-III binds with high affinity to peptides that are substrates of the other two domains.

The rest of the residues that are highly conserved in almost all CPs, either regulatory or pancreatic, like Gly^{60}, Asn^{117}, Gly^{120}, and Pro^{190} (numbering system of CPD-II) would belong to the third category. None of them has been related to any specific function, and their role is more likely purely structural.

Thus, to summarize, the catalytic machinery of CPD-III has been suppressed by replacement of the key residues for CP activity, and there are also substantial differences in the residues responsible for substrate binding. The high conservation of sequence and structure in the enzymatically incompetent CPD-III suggests another biological function, possibly related to the binding of proteins or other molecules.

**Active Site and Substrate Binding Subsites**—All residues involved in metal binding and catalysis are conserved in CPD-I, CPD-II, and CPE. CPD-III is the already commented exception, because it lacks most of the residues involved in Zn^{2⁺} binding and the Arg that binds the terminal carboxylate (here a Thr) and polarizes the scissile peptide bond (a His in CPD-III). Also, the general base (Glu^{272} in CPD-II) has its position occupied by a Tyr in CPD-III.

The loops that form the specificity pocket in CPD-II (S₁ subsite) (Asn^{188}, Asp^{192}, Gly^{246}, Gln^{257}, and Phe^{267}, Thr^{270}) have the same length in all the models: amino acid residue identity in these loops is high for CPD-I and CPE and low for CPD-III. There is also low identity between these loops of CPD-II and those of the pancreatic CPs. On the other hand, it is worth noting that Tyr^{250} (equivalent to Tyr^{244} in pancreatic CPs, the one that caps the active site, facilitating the proper location of the substrate over it, and that fluctuates between two conformations depending on substrate binding), is replaced by His in CPD-III, supporting the idea that this domain is unable to catalyze peptide bond hydrolysis.

A key residue that is essential for the specificity of digestive CPB for C-terminal basic residues is Asp^{255} (6, 45), which is replaced by an Ile or Leu in the digestive enzymes that prefers C-terminal aliphatic and aromatic residues (Table II). In CPD-I, CPD-II, and CPE, which all are highly specific for basic C-terminal amino acids, the residue in a position sequentially equivalent to this Asp^{255} of CPB is a Gln (Table II), which is functionally unable to perform a similar role as the Asp. Instead, the electronegative character required for the selectivity for C-terminal Lys and Arg residues is provided by Asp^{192}, located in a spatially comparable position. This Asp^{192} is conserved in all regulatory CP, including CPD-III. However, in CPD-III a Lys residue is found in the position equivalent to Asp^{255} of CPB (Table II). In the model built, this Lys residue is not directed toward the substrate-binding pocket, as it adopts a conformation similar to the side chain to which it has been modeled (i.e. Gln^{257} in CPD-II). However, if we consider the presence of this Lys residue, together with the above-mentioned substitution of the very conserved triad Asn-Arg-Asn at the bottom of the specificity pocket by Asp-Thr-Asp in CPD-III, it is tempting to envisage that CPD-III could be able to show a fully reversed selectivity and bind positive terminal charges linked to acidic side chains. Clearly, only a crystal structure of CPD-III in complex with a yet unknown putative substrate would shed light into the question.

The relevant residues at the S₂ subsite in CPD-II are also found in equivalent positions in the three models of CPD-I, CPD-III, and CPE. The residues that line this subsite are considerably different in pancreatic CPs, suggesting that a general specificity for either sequence or volume of the substrates is shared in all regulatory enzymes, including the inactive CPD-III. As an example, Gly^{182} and Gly^{183} (CPD-II numbering) are present in all models of the regulatory forms at the same positions found in the crystal structure, whereas the equivalent residues in pancreatic CPs are Ser^{197} and Tyr^{198}, both highly conserved in such pancreatic enzymes.

Variation is also observed in all proteins for those residues putatively involved in subsite S₃. However, one remarkable difference involves Lys^{277}, conserved in CPD-I and CPE, and that was putatively involved in P₃ carbonyl oxygen binding in CPD-II (32), which is replaced by a Tyr in CPD-III.

**Accessibility of the Active Site**—One of the most significant structural differences between the crystal structure of CPD-II and pancreatic CPs is the long insertion Tyr^{225}–His^{241} that shapes the border of the funnel that leads to the active site and that hinders the binding of potato CP inhibitor to CPD. Potato CP inhibitor is a 39-residue peptide that potently inhibits several of the digestive CPs including CPA, CPB, and CPU (see 32). Although particular residues are not conserved, the loop is present in all models suggesting that restrictions in specificity may be common to all regulatory enzymes. However, two further loops are also critical in shaping the funnel border, and significant differences are observed in these cases (Fig. 4). The insertion of a 9-residue Gly-rich sequence at loop Ser^{124}–Val^{133} (CPD-II numbering) does not seem to affect the accessible surface in CPD-I. In all cases the loop is longer than that observed in pancreatic CPs and is folded inwards, partially covering the access to the active site. CPE has a much longer insertion between residues 157 and 158 of CPD-II (Fig. 3) that coincides with an equivalent, albeit shorter, insertion in pancreatic CPA and CPB. Taken together, all these observations suggest that, within a general frame of specificity, regulatory CPs have developed variations in the structural determinants that lead to selection of substrates that are far more sophisticated than the mere selectivity of C-terminal residues observed in the pancreatic enzymes. Work is in progress to test this hypothesis.

The information collected or derived in the present study might facilitate the understanding of the differential biological roles of regulatory CPs and the design of specific inhibitors for them. These would be interesting tools to experimentally analyze the properties and roles of these enzymes, and to produce lead compounds for drug design, given the potential biotechno-
logical and biomedical interest in the modulation of their activities.

Acknowledgments—The support provided by the Training and Mobility of Researchers/Access to Large Scale Facilities program to the EMBL Hamburg Outstation (reference ERBFMGECT980154) is gratefully acknowledged.

REFERENCES
1. Fricker, L. D. (1988) Annu. Rev. Physiol. 50, 309–321
2. Skidgel, R. A. (1988) Trends Pharmacol. Sci. 9, 299–304
3. Skidgel, R. A. (1990) in Zinc Metalloproteases in Health and Disease (Hooper, N. M., ed), pp. 241–283, Taylor and Francis, London
4. Rawlings, N. D., and Barret, A. J. (1995) Methods Enzymol. 248, 183–228
5. Vandrell, J., Querol, E., Aviles, and F. X. (2000) Biochin. Biophys. Acta 1477, 284–298
6. Aviles, F. X., Vandrell, J., Guasch, A., Coll, M., and Huber, R. (1993) Eur. J. Biochem. 211, 381–389
7. He, G. P., Muise, A., Li, A. W., and Ro, H. S. (1995) Biochim. Biophys. Acta 1297, 175–185
8. Song, L., and Fricker, L. D. (1995) J. Biol. Chem. 270, 30619–30624
9. Song, L., and Fricker, L. D. (1997) J. Biol. Chem. 272, 10543–10550
10. Xin, X., Varlamov, O., Day, R., Dong, W., Bridget, M. M., Leiter, E. H., and Fricker, L. D. (1997) DNA Cell Biol. 16, 897–905
11. Lei, Y., Xin, X., Morgan, D., Pintar, J. E., and Fricker, L. D. (1999) DNA Cell Biol. 18, 175–185
12. Fricker, L. D. (1991) in Peptide Biosynthesis and Processing (Fricker, L. D., ed), pp. 199–230, CRC Press, Inc., Boca Raton, FL
13. Fricker, L. D., and Snyder, S. H. (1988) J. Biol. Chem. 258, 10950–10955
14. Fricker, L. D., Evans, C. J., Esch, F. S., and Herbert, E. (1986) Nature 323, 461–464
15. Fricker, L. D., Adelman, J. P., Douglass, J., Thompson, R. C., von Strandmann, R. P., and Hutton, J. (1989) Mol. Endocrinol. 3, 666–673
16. Rodriguez, C., Brayton, K. A., Brownstein, M., and Dixon, J. E. (1989) J. Biol. Chem. 264, 5968–5995
17. Manser, E., Fernández, D., Loo, L., Goh, P. Y., Menfres, C., Hall, C., and Lim, L. (1990) Biochem. J. 267, 517–525
18. Fan, X., and Nagle, G. T. (1996) DNA Cell Biol. 15, 937–945
19. Roth, W. W., Mackin, R. B., Spiess, J., Goodman, R. E., and Noe, B. D. (1991) Mol. Cell Endocrinol. 78, 171–178
20. Parkinson, D. (1990) J. Biol. Chem. 265, 17101–17105
21. Gsellin, J., Das, B., and Fricker, L. D. (1992) Biochem. J. 285, 613–618
22. Naggert, J. K., Fricker, L. D., Varlamov, O., Nishina, P. M., Rouille, Y.,