*Journal of Computer-Aided Molecular Design*, **14:** 83–92, 2000.
*KLUWER/ESCOM*



# Refinement of modelled structures by knowledge-based energy profiles and secondary structure prediction: Application to the human procarboxypeptidase A2

Patrick Aloy, José M. Mas, Marc A. Martí-Renom, Enrique Querol, Francesc X. Avilés & Baldomero Oliva*
*Institut de Biologia Fonamental and Departament de Bioquímica, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain*

Received 7 December 1998; Accepted 1 April 1999

*Key words:* carboxypeptidases, comparative modelling, energy profiles, molecular modelling, secondary structure prediction

**Summary**

Knowledge-based energy profiles combined with secondary structure prediction have been applied to molecular modelling refinement. To check the procedure, three different models of human procarboxypeptidase A2 (hPCPA2) have been built using the 3D structures of procarboxypeptidase A1 (pPCPA1) and bovine procarboxypeptidase A (bPCPA) as templates. The results of the refinement can be tested against the X-ray structure of hPCPA2 which has been recently determined. Regions miss-modelled in the activation segment of hPCPA2 were detected by means of pseudo-energies using Prosa II and modified afterwards according to the secondary structure prediction. Moreover, models obtained by automated methods as COMPOSER, MODELLER and distance restraints have also been compared, where it was found possible to find out the best model by means of pseudo-energies. Two general conclusions can be elicited from this work: (1) on a given set of putative models it is possible to distinguish among them the one closest to the crystallographic structure, and (2) within a given structure it is possible to find by means of pseudo-energies those regions that have been defectively modelled.

*Abbreviations:* Protein Data Bank, PDB; Carboxypeptidase, CP; Porcine procarboxypeptidase B, pPCPB; Human procarboxypeptidase A2, hPCPA2; Bovine procarboxypeptidase A, bPCPA; Porcine procarboxypeptidase A1, pPCPA1; Root Mean Square Deviation, RMSD.

## Introduction

Comparative modelling of proteins has been used to build putative 3D structures of target proteins with unknown structure since the early 1980s [1, 2]. The methodology has been improved during the last 10 years by automated algorithms, either based on assembling the model from stretches of related structures [3–6], or by means of spatial restraints extracted from the set of homologous proteins that have to be satisfied by the target protein [7–9] (for a review see [10]). The methods have been successful when the target protein is close to the known structures and when the sequence percentage identity to the unknown is greater than 40%. Nevertheless, some aspects of the modelling remain still unsolved. In particular, the correct assignment, from the alignments, of the N and C terminal residues of the regular secondary structures and the conformation of loops, which usually present the lowest percentage of identity on the alignment. Another problem to tackle is how to evaluate the accuracy of a model or at least to select the best model among several possible conformations.

*To whom correspondence should be addressed. E-mail: baldo@pug.uab.es

The assessment of the accuracy of a model is similar to the problem of validating an experimentally determined structure. One of the approaches for such a purpose is 'the inverse protein folding' [11]. The goal of inverse protein folding is to identify the sequences in the data base which fit into a known fold and its guiding principle is the energy or fitness of sequence structure pairs. This principle is based on the folding postulate that guarantees that the pseudo-potential energy of a given sequence within its native conformation is a global minimum over all possible conformations. This principle acquiesces the profile method designed by Eisenberg and co-workers which proposes that it can be used to identify incorrect chain tracings and problematic structures [12]. The knowledge-based mean field (pseudo-potential energies) can be used in this topic to analyse the distribution of energies in experimentally determined structures, obtaining profiles which display native-like features of protein folds [13]. In this sense, the same methodology can be used to distinguish regions where comparative modelling may fail or to choose the best conformation when more than one model is considered [14].

In order to distinguish the accuracy of a model we have tested different methods for comparative modelling. The automated program COMPOSER [15] models a target protein by superimposing, as rigid bodies, the known structures of homologous proteins and aligning its sequence upon the consensus sequence of the known. On the other hand, the automated program MODELLER [16] uses spatial constraints extracted from the multiple alignment of the known homologous proteins to construct the target protein by satisfying these restraints. Both methods can be tested on the basis of the native-like profiles of the model.

The actual methods of secondary structure prediction (about 70% accuracy) [17–21] can be very helpful for checking a model built structure [18, 20, 21], mainly because the misalignments on the loop regions could be related with erroneous secondary structure assignment for the model. Therefore, the modelled structure of a target protein can be modified in order to agree with its secondary structure prediction, although this does not necessarily mean that the model has been improved. The validation of the modification can be assessed by pseudo-potential energies as formerly introduced.

As an example of this hypothetical methodology to improve the model and to check the accuracy of the structure the human procarboxypeptidase A2 has been chosen. This can be a good example because a previous model was proposed [22] and its actual conformation in crystals has been recently obtained by X-ray diffraction analysis [23]. In addition, in the PDB there are the structures for hPCPA2 [23], bPCPA, pPCPA1 [25, 26] and pPCPB [27]. Therefore, the hypothesis pursued here can be corroborated with this example. This is also a very interesting case with relevant biological significance. There are two isoforms of A pancreatic carboxypeptidases known as isoforms A1 and A2. A1 carboxypeptidases show preference for aliphatic C-terminal residues (e.g., bovine and porcine carboxypeptidases A and A1) whilst A2 isoforms show preference for aromatic C-terminal residues. Its zymogen forms (known as procarboxypeptidases) present a quite specific system of activation [24]. Therefore, comparative modelling of the structure of hPCPA2 was tackled in order to characterise the fold and the activation mechanism of this proenzyme [22, 28].

The known structures of bPCPA and pPCPA1 can be split in three regions. The main domain is the carboxypeptidase (CP), formed by around 307 residues, yielded by the tryptic cleavage of the proenzyme into the activation segment (N-terminal fragment) and the enzyme (CP). The N-terminal activation segment, formed by the first 92 or 93 residues of the proenzyme, produced by the tryptic cleavage of the proenzyme, is split in two regions: (1) the activation domain, formed by residues 1 to 80, which is responsible for the inhibition of the CP; and (2) the connecting segment which joins the activation segment and the enzyme, formed by residues 81 to 92, mainly structured in $\alpha$-helix conformation.

The sequence of human procarboxypeptidase A2 shows 64 and 63% identity with respect to bovine and porcine carboxypeptidases A and A1 respectively. This high percentage of identity facilitates the model building by homology, the structure of hPCPA2 upon bPCPA and pPCPA1, and the modelling can be split in three regions (CP, activation domain and connecting segment). The residues involved in these regions are: (1) the last 307 residues in the C-terminal fragment, forming the main enzyme domain (CP); (2) the first 80 residues forming the activation segment; and (3) between residues 81 and 95, forming the connecting segment.

The secondary structure prediction of the hPCPA2 shows significant differences with respect to the secondary structure of bPCPA and pPCPA1, mainly in the connecting segment region. Therefore, any

```
        10        20        30        40        50        60        70        80
       EEEEE  HHHHHHHHHHHHHHH   EEEE          EEEEE    HHHHHHHHHH   EEEEHHHHHHHH PCPA2h PHD
-ETFVGDQVLEIVPSNEEQIKNLLQLEAQEHLQLDFWKSPTTPGETAHVRVPFVNVQAVKVFLESQGIAYSIMIEDVQVL PCPA2h
KEDFVGHQVLRITAADEAEVQTVKELEDLEHLQLDFWRGPGQPGSPIDVRVPFPSLQAVKVFLEAHGIRYRIMIEDVQSL PCPA1b
KEDFVGHQVLRISVDDEAQVQKVKELEDLEHLQLDFWRGPARPGFPIDVRVPFPSIQAVKVFLEAHGIRYTIMIEDVQLL PCPA1p

        90       100       110       120       130       140       150       160
HHHHHHHHHHHHHH                                                                PCPA2h PHD
LDKENEEMLFNRRRERSGN-FNFGAYHTLEEISQEMDNLVAEHPGLVSKVNIGSSFENRPMNVLKFSTGG-DKPAIWLDA  PCPA2h
LDEEQEQMFASQSRARSTNTFNYATYHTLDEIYDFMDLLVAEHPQLVSKLQIGRSYEGRPIYVLKFSTGGSNRPAIWIDL  PCPA1b
LDEEQEQMFASQGRARTTSTFNYATYHTLEEIYDFMDILVAEHPALVSKLQIGRSYEGRPIYVLKFSTGGSNRPAIWIDS  PCPA1p

       170       180       190       200       210       220       230       240
GIHAREWVTQATALWTANKIVSDYGKDPSITSILDALDIFLLPVTNPDGYVFSQTKNRMWRKTRSKVSGSLCVGVDPNRN  PCPA2h
GIHSREWITQATGVWFAKKFTEDYGQDPSFTAILDSMDIFLEIVTNPDGFAFTHSQNRLWRKTRSVTSSSLCVGVDANRN  PCPA1b
GIXSRXWITQASGVWFAKKITENYGQNSSFTAILDSMDIFLEIVTNPNGFAFTHSDNRLWRKTRSKASGSLCVGSDSNRN  PCPA1p

       250       260       270       280       290       300       310       320
WDAGFGGPGASSNPCSDSYHGPSANSEVEVKSIVDFIKSHGKVKAFIILHSYSQLLMFPYGYKCTKLDDFDELSEVAQKA  PCPA2h
WDAGFGKAGASSSPCSETYHGKYANSEVEVKSIVDFVKDHGNFKAFLSIHSYSQLLLYPYGYTTQSIPDKTELNQVAKSA  PCPA1b
WDAGFGGAGASSSPCAETYHGKYPNSEVEVKSITDFVKNNGNIKAFISIXSYSQLLLYPYGYKTQSPADKSELNQIAKSA  PCPA1p

       330       340       350       360       370       380       390       400
AQSLRSLHGTKYKVGPICSVIYQASGGSIDWSYDYGIKYSFAFELRDTGRYGFLLPARQILPTAEETWLGLKAIMEHVRD  PCPA2h
VEALKSLYGTSYKYGSIITTIYQASGGSIDWSYNQGIKYSFTFELRDTGRYGFLLPASQIIPTAQETWLGVLTIMEHTLN  PCPA1b
VAALKSLYGTSYKYGSIITVIYQASGGVIDWTYNQGIKYSFSFELRDTGRRGFLLPASQIIPTAQETWLALLTIMEHTLN  PCPA1p

       410
HPY                                                                           PCPA2h
NLY                                                                           PCPA1b
NSX                                                                           PCPA1p
```

*Figure 1.* Multiple alignment of human pro-carboxypeptidase A2 (hPCPA2), bovine pro-carboxypeptidase A1 (bPCPA1) and the porcine pro-carboxypeptidase A1 (pPCPA1p). This alignment has been performed using the program PILEUP included in the GCG Package (Wisconsin Packcage Version 9.0, Genetics Computer Group, Madison, WI). The numbering used here has been performed sequentially from the N-terminus (number 1) to the end of these proteins. The relation between this numbering and that used at the PDB can be extracted from this alignment. A regular secondary structure prediction for the hPCPA2 has been performed by means of the program PHD [20] and it is also shown.

model of hPCPA2 based on its alignment with bPCPA and/or pPCPA1 misleads the correct structure in one of the most relevant regions of the procarboxypeptidase related with the activation of the enzyme. As a consequence, the comparative modelling of hPCPA2 becomes a very attractive example to test the improvement on modelling and to select the most accurate model.

## Methods

The sequence of human procarboxypeptidase A2 [22] has been used as target for the comparative modelling. The three dimensional coordinates of hPCPA2 have been used to check the validity of the different models and the approach followed for its refinement. The sequence and three dimensional coordinates of bovine procarboxypeptidase A (bPCPA) and porcine procarboxypeptidase A1 (pPCPA1) were used as templates for the hPCPA2 model. The sequence of hPCPA2 was aligned with respect to bPCPA and pPCPA1 by means of a multiple alignment with the PILEUP program [29, 30], using the default values of the program for the gap penalty and the PAM matrix for the sequence comparison. The secondary structure prediction was obtained by means of the PHD program [20, 31].

The program COMPOSER [15] was used for the construction of the hPCPA2 structure by comparative modelling. Two more different methods based on spatial restraints were used: one extracts the distance restraints from the closest homologous protein with known three-dimensional structure. This model

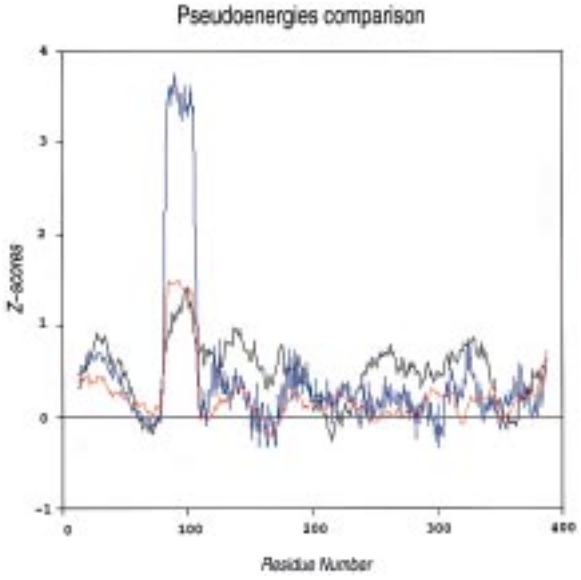


*Figure 2.* Differences in pseudo-energies calculated between the original models and the crystallographic structure. The values have been obtained using the PROSA II program [23]. The plot shows a large increase of pseudo-energy for the region that involves the connecting segment in all models. This increase is especially remarkable for the COMPOSER model. DR model in black, CMP model in blue and MDL model in red.

has already been published [22] and was applied using the program DIANA [40]. The other method is based on several spatial restraints extracted from multiple alignments previously described; this model has been obtained using the automatic program MODELLER [16]. The program Prosa II [13, 32] was used to check the quality of the model. The regions with non near-native fold are identified by the high positive values of pseudo-potential energy, independently of the crystallographic structure.

The secondary structure prediction was used to check those regions identified by Prosa II with non near native fold. A further re-modelling was applied for those regions where the predicted secondary structure was different to that found on the model. The program FRAZER, developed in our laboratory, was used to reconstruct these problematic regions by superimposition of a predicted secondary structure conformation with the corresponding sequence of hPCPA2. The program TURBO-FRODO [33] was used for the visual inspection of the model and the correction of the region. The pseudo-energy of the final model built was inspected again in order to identify the regions of the protein still ambiguous.

The RMSD and superimposition of the structure of the hPCPA2 models with respect to the crystallographic structure was also obtained with FRAZER. The whole conformation of the models was checked with the program DIFDIST, plotting by a color-scale gradient the difference ($\Delta d_{ij}$) of the Cα atoms distance ($d_{ij}$) between the structures of the model of hPCPA2 and the X-ray structure. The average of this difference is given as a whole parameter for the comparison of the two structures (DD = ΣΣ Δdij/2N), whilst the plot of the differences was used to identify the wrongly modelled regions (Distance Map plot). Both programs are available by FTP in ftp://luz.uab.es.

## Results

Three modelled structures of human procarboxypeptidase A2 (hPCPA2) have been analysed and afterwards modified according to the secondary structure prediction. The first model is taken from a previous work [22] obtained by applying specific distance restraints extracted from porcine procarboxypeptidase A1 (pPCPA1) which is homologous to hPCPA2; this model is named hereafter as 'original Distance Restraints model' (DR). The second model is automatically obtained with the program COMPOSER according to the alignment of hPCPA2 with bovine A and porcine A1 procarboxypeptidases (bPCPA and pPCPA1, respectively) and it is named 'original COMPOSER model' (CMP). The last model is also automatically obtained with the program MODELLER using the same alignment as for the original COMPOSER model, this model is analogously named 'original MODELLER model' (MDL). The three models analysed were modified according to the secondary structure prediction of the activation and connecting segments of the hPCPA2 and named with the extension 'SS' (i.e., DR-SS, CMP-SS, MDL-SS).

### *Sequence alignment and secondary structure prediction*

The sequence of human procarboxypeptidase A2 (hPCPA2) was aligned with respect to bovine (bPCPA) and porcine (pPCPA) procarboxypeptidases A, for whom the 3D structures were known [27, 34]. The model structure of hPCPA2 was obtained by means of this alignment (Figure 1). The secondary structure prediction was also obtained for the hPCPA2 sequence by means of PHD [20, 21]. Figure 1 shows

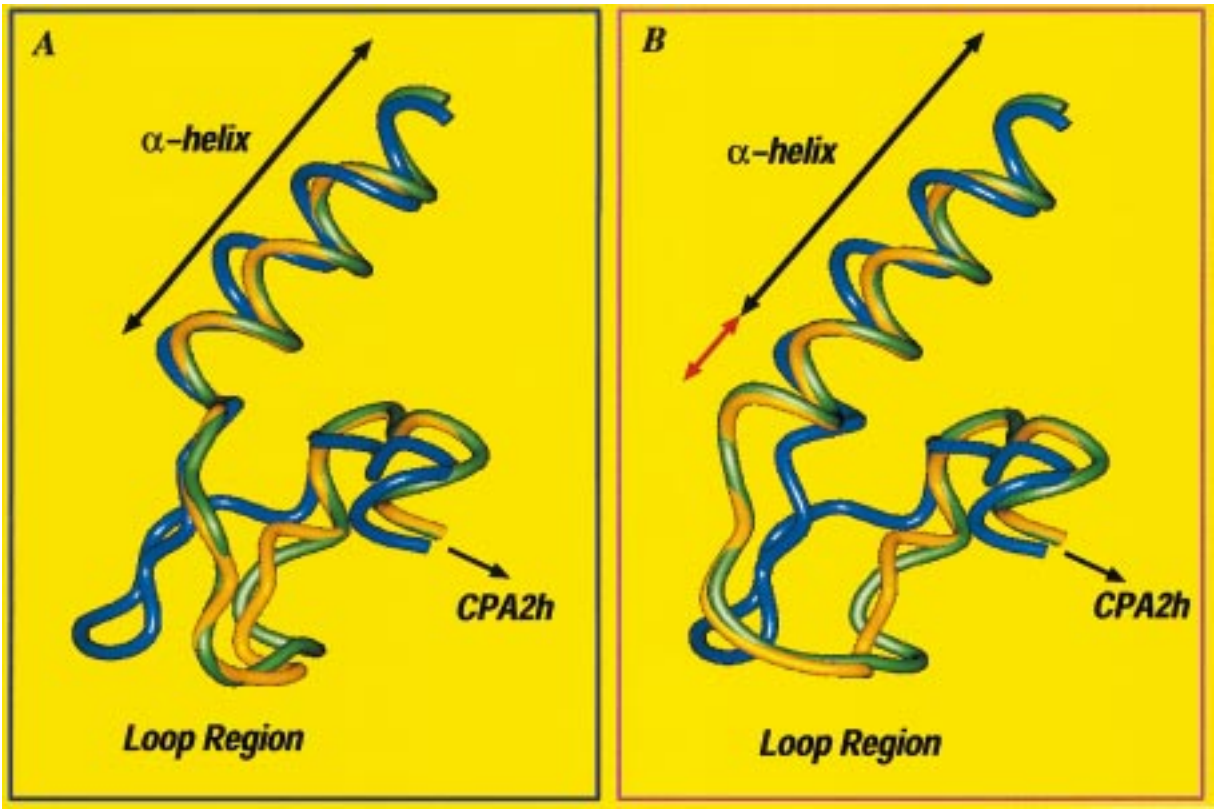


*Figure 3.* Backbone trace for the comparison of the different models of the loop region 'bridging' the connecting segment and the enzyme region. (A) shows the models DR (cyan), MDL (orange) and CMP (green) while (B), with the expanded connecting segment, marked with a red arrow, shows the models DR-SS (cyan), MDL-SS (orange) and CMP-SS (green).

the secondary structure prediction of the pro-segment of hPCPA2. The main difference between bPCPA and pPCPA1 secondary structures and the secondary structure predicted for hPCPA2 is found in the connecting segment and the first α-helix of the pro-segment. The secondary structure predicted for the connecting segment of hPCPA2 shows two extra α-helix turns (4-turn in DSSP nomenclature) with respect to the connecting segment of pPCPA1 and bPCPA in the C-terminal cap of the α-helix. The prediction for the first α-helix of the pro-segment also shows one more turn than found on pPCPA1 and bPCPA1. However, only for the connecting segment this difference is also corroborated by the fact that three arginines are found at the end of the α-helix at its C-terminal side, stabilising the dipolar momentum of the α-helix and its possible extension at the C-cap. Moreover, the sequence of a Shellman motif can be identified in the C-cap region by extension of this helix. Nevertheless, the current automated methods of protein modelling by homology do not automatically recognise this possibility, hence being neglected in the original models of hPCPA2 [22].

*Detection of possible wrongly modelled regions*

The pseudo-energies of the original models are calculated with PROSA II (Figure 2) in order to identify their incorrect chain tracings. It is remarkable that the region at the end of the connecting segment and the loop that links it with CPA2 (around residues 90 to 100) shows the highest energy. In this case we could also assume that the α-helix should be extended in the C-terminal cap because of the secondary structure prediction and also because of the over-stability produced by the last three arginines on the total dipolar momentum of the α-helix. This is an important region for the enzyme activation, therefore special attention should be given to the modelling of its conformation.

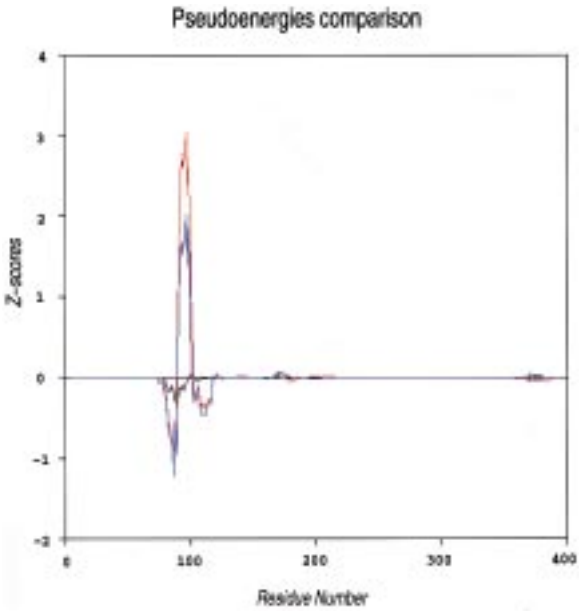


*Figure 4.* Difference in pseudo-energies calculated between the original models and the modified structure by means of their secondary structure prediction. The values have been obtained using the PROSA II program. The plot shows a large increase of pseudo-energy for the region that involves the connecting segment in all modified models and a clear decrease in the pseudo-energies at the end of the connecting α-helix. DR-SS model in black, CMP-SS model in red and MDL-SS model in blue.

### *Refinement of the model by secondary structure prediction*

According to the secondary structure prediction of hPCPA2, the first α-helix of the pro-segment and the α-helix of the connecting segment are longer than in the original model. Therefore, three new models of hPCPA2 are built by extending both α-helix conformations in their C-terminal cap (Figure 3).

Figure 4 shows for each model the difference in pseudo-energies between the original and its modified version. This indicates that the elongation of the first α-helix of the pro-segment does not improve the results in terms of pseudo-energies. On the contrary, the modification in the C-cap of the connecting segment helix improves the pseudo-potential energy of this region for the three-modelled structures. However, energies at the N-terminal side of the loop that connects the α-helix of the connecting segment and the enzyme (residues 95–100) are not improved for the CMP-SS and MDL-SS models. Interestingly, the whole region from residues 90 to 100 is improved in the DR-SS model.

The actual conformation of the first α-helix of the pro-segment of hPCPA2 is not clear and thus it is risky to accept the possibility of extending this α-helix unless more information is known. On the other hand, the local conformation of the modelled 'bridging' loop presents an important increase in pseudo-energy, which indicates the failure of the new model in this region. Therefore, it could be necessary to choose between rearranging the secondary structure to ameliorate the pseudo-energy of the modified fraction (although giving a worse energy for the unmodified loop) or to maintain the model unchanged, knowing that probably the loop will be wrongly modelled anyway.

Unfortunately, extra data is required in order to improve the local conformation of the 'bridging' loop. The local conformation of loops is a complex subject which we have not addressed in this paper. Work is in progress on this matter [35] by means of detailed analyses of loop classification [36, 37]; by now there is not enough data to rearrange a long loop (larger than 10 residues) to be modelled from scratch.

### *Comparison between the models and the X-ray structure*

The crystallographic structure of hPCPA2 has recently been obtained [23] and we have had access to the 3D coordinates previous to their availability in the PDB. Therefore, we have been able to check the accuracy of the previous models of the structure of hPCPA2 and analyze the disagreements. The accuracy of the original models has been tested by means of the RMSD of the overall hPCPA2 structure (Table 1). The structure of hPCPA2 has been split in three main regions for which the RMSD has been calculated: the activation domain (residues 1 to 76), the connecting segment (residues 77 to 93) and the enzyme domain (hCPA2). The two former regions constitute the complete pro-segment. The Distance Map plot of the overall modelled structure compared with the crystallographic structure (Figure 5) shows the regions where the original modelling has failed.

In Table 1 the RMSD results are shown for the comparison between the models and the crystallographic structure (overall and in regions). The MDL model obtained from scratch for the whole hPCPA2 shows the smaller RMSD of the backbone (1.2 Å). Also COMPOSER gives a close model to the crystallographic structure (RMSD of the backbone is 1.3 Å), but the DR model [22] gives the worst result (RMSD

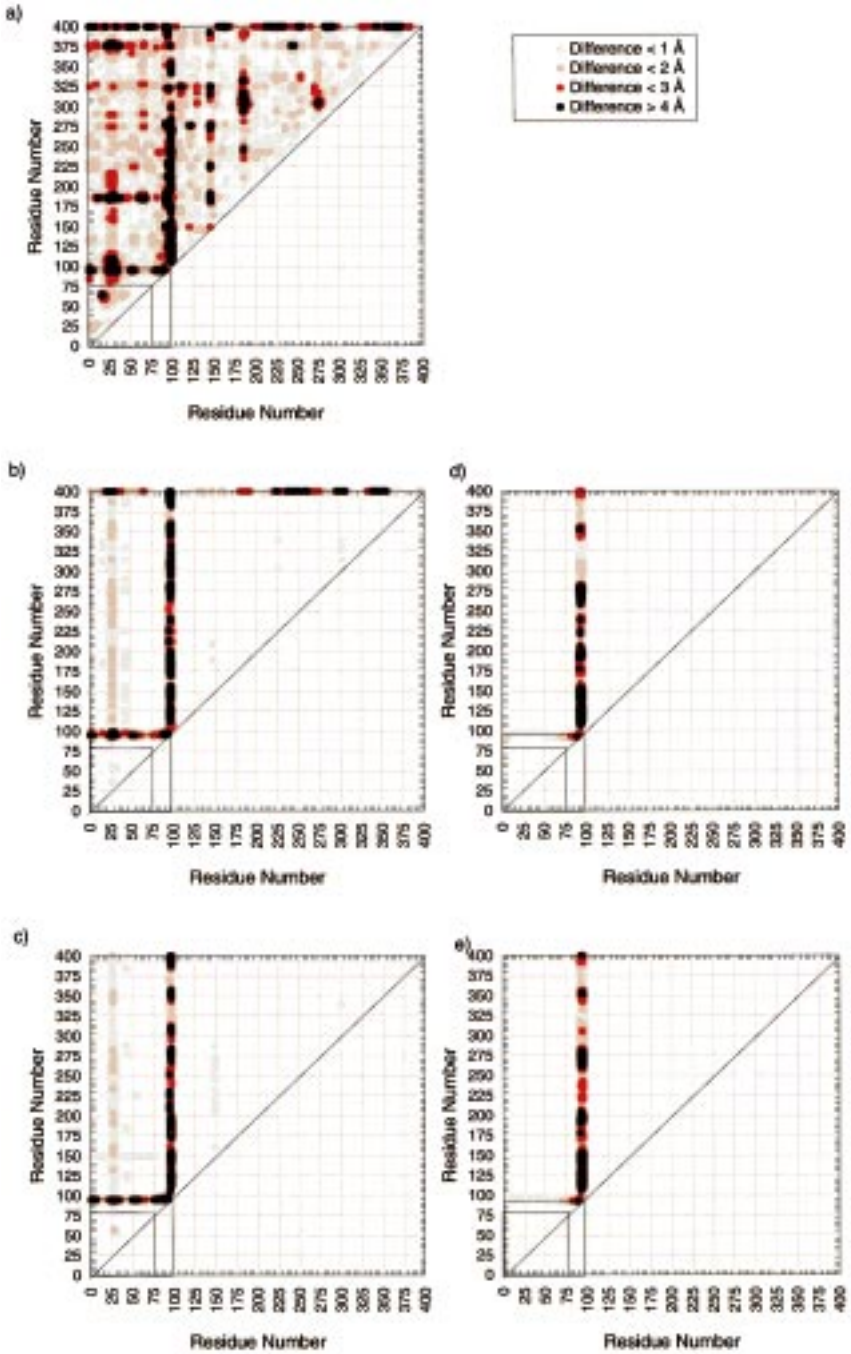


*Figure 5.* (a–c) Distance Map plots of the overall modelled structures compared with the crystallographic structure. While the DR method shows the worst results for the complete proteins, three regions are not properly modelled by the automatic methods (MDL and CMP). These regions are the C-terminus, between the residues 25 and 30 (the first alpha helix of the activation domain) and between residues 90 and 100 (the 'bridging loop', after the connecting segment). (d, e) Difference plots between the original models and the ones that have been modified according to the pseudo-energy profiles and the secondary structure prediction.

*Table 1.* RMSD and average distances (Å) between X-ray and modelled structures (DR-MC, CMP-MC, MDL-MC, DR-SS, CMP-SS, MDL-SS, DR-SS-MC,CMP-SS-MC, MDL-SS-MC) (see text). Calculations have been performed for the backbone and all atoms using the program FRAZER. These values are split in four groups: the pro-enzyme, the activation domain, the connecting segment and the enzyme region

| Region | Atoms | DR model | DR-SS model | CMP model | CMD-SS model | MDL model | MDL-SS model |
|---|---|---|---|---|---|---|---|
| PCPA2h | Backbone | 2.25 | 2.25 | 1.31 | 1.09 | 1.20 | 0.96 |
| | All | 3.06 | 3.06 | 2.03 | 1.81 | 1.89 | 1.68 |
| Activation segment | Backbone | 1.64 | 1.64 | 0.87 | 0.87 | 0.81 | 0.81 |
| | All | 2.53 | 2.53 | 1.52 | 1.52 | 1.53 | 1.53 |
| Connecting segment | Backbone | 1.93 | 1.70 | 1.96 | 0.53 | 1.87 | 0.37 |
| | All | 3.74 | 3.73 | 3.74 | 3.08 | 3.88 | 2.16 |
| CPA2h | Backbone | 2.06 | 2.06 | 1.27 | 1.07 | 1.15 | 0.92 |
| | All | 2.90 | 2.90 | 1.91 | 1.83 | 1.69 | 1.65 |
| PCPA2h | DD[a] | 1.20 | 1.20 | 0.52 | 0.48 | 0.47 | 0.43 |

[a]Average values of the distance difference between Ca atom pairs of the modelled structures and the X-ray.

of the backbone is 2.2 Å). Also the average distance difference (DD) is presented in Table 1, which leads to the same conclusion as in the RMSD analyses. On the other hand, when the different regions are analyzed, the connecting segment region shows the largest RMSD for the three original models (around 1.9 Å for the RMSD of the backbone), whilst for the activation domain region the RMSD is smaller than 1 Å except for the DR model (1.6 Å). By visual inspection the hypothesis is corroborated that the α-helix forming the connecting segment of hPCPA2 is two turns longer than for the pCPA1 and bPCPA1. Interestingly, the CMP-SS and the MDL-SS models show a much better structural agreement with respect to the crystallographic structure than its original models (CMP and MDL). It is remarkable that the connecting segment for all the modified models shows smaller RMSD than in the original models, whilst for the activation domain the modelled structures remain under the same accuracy with respect to the crystallographic data (RMSD does not vary).

The Distance Map plots obtained for the differences between the original models and the crystallographic structure of hPCPA2 (Figure 5) clearly show that the original DR model fails in almost the complete structure, whilst for the original models obtained by the automated programs COMPOSER and MODELLER only three regions are detected that have been wrongly modelled. These regions are: (1) between residues 25 and 30; (2) between residues 90 and 100; and (3) the last C-terminal residues of the enzyme moiety. The last residues on the C-terminal side are of minor importance because this region may have larger mobility than the rest of the structure. However, the region comprised between residues 90–100 plays a major role in the activation of hPCPA2 [38]. Figures 5d and 5e show how this region has been partially corrected after the modification of the helices in both models.

## Discussion

One of the aims of this work is to improve the molecular homology modelling and to find a reliable and simple procedure to predict the accuracy of the obtained models. Human procarboxypeptidase A2 is one case where the comparative modelling by homology fails to predict the secondary structure of the connecting segment region whilst the current predictive methods of secondary structure [20, 21] are more accurate. Although it is not possible to conclude with a general rule or principle, it seems reasonable to hypothesise the general validity of the method to improve modelling when additional information, as secondary structure predictions and pseudo-energy profiles, can be calculated. This procedure has been tested for three different approaches.

The present work has also shown that, although the current automated methods for comparative modelling are able to get accurate structural models for proteins with highly homologous sequences and to known 3D structures, still there can be regions were the lack of either homology or information leads to wrong answers. It is remarkable that inverse folding methods and pseudo-potential energies can at this point be of

much help to deduce the wrongly modelled regions and correct them by means of additional information. The recently attained accuracy on prediction of secondary structure, mostly for sequences homologous to a large family of proteins, shows to be of much help on giving the extra information necessary for the rearrangement of the model.

Despite the strong signal for secondary structure cappings, on average the ends of helices and strands are less accurately predicted than the core [41], generally being the helices predicted too long and the strands too short [42]. In our test system, we had different evidences to believe that the secondary structure prediction was right: the pseudo-energy profiles improved when the model was modified. On the other hand, by observation of the sequence, a Schellman motif was found in the C-cap of the predicted helix [43], and also a microdipole (EE—RRR) that could stabilise the helix [44]. Finally, it is widely known that the charged residues at the end of the helices neutralise the effect of the electrostatic macrodipole [45] and consequently we could consider the extension of the original helix according to the secondary structure prediction.

Some particular rules can be deduced from this work that might prove useful for the general case of modelling by homology. Two postulates are proposed in order to test the improvement of the model building: first of all, it is possible to distinguish the best conformation of a region between two or more proposed conformers by means of pseudo-potential energies of the local region; and, second, it is also possible to select the most accurate model using the same pseudo-potential energies of the overall backbone.

The next step in the refinement of a model is the improvement in the conformation of local regions where the pseudo-energies have shown a possible failure. To do this, extra information is required which can be obtained from the secondary structure prediction, experimental data, stabilisation of C and/or N terminal caps of α-helices, etc. In this study we have checked the improvement of our model according to the secondary structure prediction, concluding that the local conformation of one model obtained by comparative modelling has to be modified in order to agree with its secondary structure prediction whenever the pseudo-potential energy of the modified fraction of the local region (where different secondary structure is predicted) becomes improved.

A side-chain refinement using multy-copy techniques was performed (data not shown), but it did not improve the models. At this level of sequence identity, most of the side chains maintain the same conformation, and thus this sort of refinement is unnecessary. Furthermore, the use of limited energy minimization can be beneficial in order to improve the stereochemical quality of the model, but the RMSD often becomes worse as was suggested after CASP1 results [39].

## Acknowledgements

This work has been supported by grants BIO98-036, BIO97-0511 and IN94-0347 from CICYT (Ministerio de Educación y Ciencia, Spain) and by the CERBA (Centre de Referència en Biotecnologia) de la Generalitat de Catalunya. P.A. is a predoctoral fellowship recipient of the Ministerio de Educación y Ciencia, J.M.M has a predoctoral fellowship from CIRIT-CERBA (Generalitat de Catalunya) and M.A.M-R. is a predoctoral fellowship recipient from Universitat Autònoma de Barcelona (FI-DGR/UAB).

## References

1. Warme, P.K., Momany, F.A., Rumball, S.V., Tuttle, R.W. and Scheraga, H.A., Biochemistry, 13 (1974) 768.
2. Browne, W.J., North, A.C.T., Phillips, D.C., Brew, K., Vanaman, T.C. and Hill, R.C., J. Mol. Biol., 42 (1969) 63.
3. Unger, R., Harel, D., Wherland, S. and Sussman, J.L. (1989) Proteins, 5 (1989) 355.
4. Blundell, T.L., Sibanda, B.L., Sternberg, M.J. and Thornton, J.M., Nature, 326 (1987) 347.
5. Blundell, T.L., Barlow, B., Sibanda, B.L., Thornton, J.M., Taylor, W.R., Tickle, I.J., Sternberg, M.J., Pitts, J.E., Haneef, I. and Hemmings, A.M., Phil. Trans. Roy. Soc. Lond., A317 (1986) 333.
6. Blundell, T.L. and Sternberg, M.J., Trends Biotechnol., 3 (1985) 228.
7. Srinivasan, S., March, C.J. and Sudarsanam, S., Protein Sci., 2 (1993) 227.
8. Havel, T.F. and Snow, M.E., J. Mol. Biol., 217 (1991) 1.
9. Sali, A., Overington, J.P., Johnson, M.S. and Blundell, T.L., Trends Biochem. Sci., 15 (1990) 235.
10. Sánchez, R. and Sali, A., Curr. Opin. Struct. Biol., 7 (1997) 206.
11. Godzik, A. and Skolnick, J., J. Mol. Biol., 227 (1992) 227.
12. Lütthy, R., Bowie, J.U. and Eisernberg, D., Nature, 356 (1992) 83.
13. Sippl, M.J. and Weitckus, S., Proteins, 13 (1992) 258.
14. Sippl, M.J., Proteins, 17 (1993) 355.
15. Topham, C.M., Thomas, P., Overington, J., Johnson, M.S., Eisenmenger, F. and Blundell, T.L., Biochem. Soc. Symp., 57 (1990) 1.
16. Sali, A. and Blundell, T.L., J. Mol. Biol., 234 (1993) 779.
17. Fasman, G.D., Prediction of Protein Structure and the Principles of Protein Conformations, Plenum Press, New York, NY, 1998.

18. King, R.D., Saqi, M., Sayle, R. and Sternberg, M.J., Comput. Appl. Biosci., 13 (1997) 473.
19. Clotet, J., Cedano, E. and Querol, E., Comput. Appl. Biosci., 10 (1994) 495.
20. Rost, B., Sander, C. and Schneider, R., Comput. Appl. Biosci., 10 (1994) 53.
21. Rost, B. and Sander, C., J. Mol. Biol., 232 (1993) 584.
22. Catasus, L., Vendrell, J., Aviles, F.X., Carreira, S., Puigserver, A. and Billeter, M., J. Biol. Chem., 270 (1995) 6651.
23. García-Sáez, I., Reverter, D., Vendrell, J., Aviles, F.X. and Coll, M., EMBO J., 16 (1997) 6906.
24. Avilés, F.X., Vendrell, J., Guasch, A., Coll, M. and Huber, R., Eur. J. Biochem., 211 (1993) 381.
25. Pascual, R., Burgos, F.J., Salva, M., Soriano, F., Mendez, E. and Aviles, F.X., Eur. J. Biochem., 179 (1989) 609.
26. Coll, M., Guasch, A., Aviles, F.X. and Huber, R., EMBO J., 10 (1991) 1.
27. Guasch, A., Coll, M., Aviles, F.X. and Huber, R., J. Mol. Biol., 224 (1992) 141.
28. Aloy, P., Catasus, L., Villegas, V., Reverter, D., Vendrell, J. and Avilés, F.X., Biol. Chem., 379 (1998) 149.
29. Higgins, D.G. and Sharp, P.M., Comput. Appl. Biosci., 5 (1989) 151.
30. Fenger, D.F. and Doolittle, R.F., J. Mol. Evol., 25 (1987) 251.
31. Sutcliffe, M.J., Haneef, I., Carney, D. and Blundell, T.L., Protein Eng., 1 (1987) 377.
32. Sippl, M.J., J. Mol. Biol., 213 (1990) 859.
33. Roussell, A., Inisan, A.G. and Knoops-Mouthy, E., TURBO FRODO (version 5.0a) Manual, Technopole de Chateaux-Gombert, Marseille (1994).
34. Gomis-Rüth, F.X., Gómez-Ortiz, M., Vendrell, J., Ventura, S., Bode, W., Huber, R. and Avilés, F.X., J. Mol. Biol., 269 (1997) 861.
35. Oliva, B., Bates, P., Querol, E., Avilés, F.X. and Sternberg, M.J., J. Mol. Biol., 279 (1998) 1193.
36. Oliva, B., Bates, P., Querol, E., Avilés, F.X. and Sternberg, M.J., J. Mol. Biol., 266 (1997) 814.
37. Rufino, S.D., Donate, L.E., Canard, L.H.J. and Blundell, T.L., J. Mol. Biol., 267 (1997) 352.
38. Reverter, D., Villegas, V., Ventura, S., Vendrell, J. and Avilés, F.X., J. Biol. Chem., 273 (1998) 3535.
39. Mosimann, S., Meleshko, R. and James, M.N., Proteins Struct. Funct. Genet., 23 (1995) 301.
40. Güntert, P., Braun, W. and Wüthrich, K., J. Mol. Biol., 217 (1991) 517.
41. Rost, B. and Sander, C., In Bohr, H. and Brunak, S. (Eds), Protein Structure by Distance Analysis, IOS Press, Amsterdam, 1994, pp. 257–276.
42. Rost, B., Proteins Struct. Funct. Genet. Supplement 1 (1997) 192.
43. Aurora, R. and Rose, G.D., Protein Sci., 7 (1998) 21.
44. Negrete, J.A., Viñuales, Y. and Palau, J., Protein Sci., 7 (1998) 1368.
45. Villegas, V., Viguera, A.R., Avilés, F.X. and Serrano, L., Folding Design, 1 (1995) 29.