Sequence homologies, hydrophobic profiles and secondary structures of cathepsins B, H and L: comparison with papain and actinidin

E. Dufour

Biochimie **70** (1988) 1335–1342

Review

Sequence homologies, hydrophobic profiles and secondary structures of cathepsins B, H and L: comparison with papain and actinidin

Eric DUFOUR

SRV, INRA de Theix, 63122 Ceyrat, France

(Received 18-1-1988, accepted after revision 25-2-1988)

Summary — The comparison of the amino acid sequences of 5 cysteine proteinases: papain, actinidin, rat cathepsins B and H and chicken cathepsin L, demonstrates a striking homology among their sequences. The N-terminal region (residues 1–70 in papain) and C-terminal region (residues 118–212 in papain) display the highest sequence homologies, whereas the lowest sequence homologies are observed in the middle region (residues 71–117 in papain); a segment where most insertions / deletions are observed. The highest sequence homology is observed between rat cathepsin H and chicken cathepsin L. As shown by X-ray studies, papain and actinidin have a clearly defined double domain structure. Each domain contains a core of non-polar side chains, which are retained in cathepsins B, H and L, except for the non-polar residue 203 of the core which is replaced by glutamic acid in cathepsin B. The percentage and the location of α-helix and β-sheets of cathepsins B, H and L, assessed using the methods of Garnier *et al.* (1978, *J. Mol. Biol.* 120, 97–120) and Chou and Fasman (1974, *Biochemistry* 13, 222–245), show that the main ordered structures in papain and actinidin are probably retained in cathepsins B, H and L. The differences observed occur essentially in the middle region, a place where sequences display the lowest homologies and which is far removed from the active site

cysteine proteinase / sequence homology / secondary structure

Introduction

The cysteine proteinase family includes mainly plant thiol proteases, papain [1] and actinidin [2], and animal lysosomal thiol proteases, cathepsins B, H and L [3]. Papain and actinidin, which are easily purified in large quantities, have been studied extensively. Their amino acid sequences have been determined and their molecular structures have been refined to a resolution of 1.65 Å for papain [4] and 1.7 Å for actinidin [5]. For cathepsins, only the primary structure is known: rat cathepsins B and H have been sequenced by Takio *et al.* [6] and the amino acid sequence of chicken cathepsin L has been reported recently [7, 8]. The cathepsins play an important role in intracellular degradation of proteins and, possibly, in the activation of some peptide hormones [9]. In addition, several papers report that cathepsin B-like [10] and L-like [11] enzymes are released from tumors and might be involved in tumor metastasis

The recent determination of the chicken cathepsin L sequence has been used to estimate structural similarities among cysteine proteinases using 3 criteria: sequence homologies, retention of the hydrophobic cores and secondary structure. Indeed, the studies about structure and conformation of cathepsins are rather rare and only cathepsin B [12] and cathepsin L secondary structures have been investigated by circular dichroism. As shown below, it appears that all the sequences of cysteine proteinases retain the active site residues, while the amino acids

Abbreviations: DC_H: decision constant for α-helix; DC_E: decision constant for β-sheet.

around the S_2 subsite 'pocket' differ and could be responsible for the enzyme specificity. Likewise, it appears that the non-polar amino acids forming the hydrophobic cores in actinidin and papain are grossly conserved in all cathepsin sequences. As for secondary structures, the methods of Garnier *et al.* [13] and Chou and Fasman [14] were used to predict the location of α-helix and β-sheets and suggest that the main ordered structures observed in papain and actinidin are maintained in cathepsins B, H and L. Nevertheless, cathepsins H and L may have a higher α-helix content than plant cysteine proteinases.

Sequence homologies

In Fig. 1, the sequences of papain [1], actinidin [2], rat cathepsins B and H [6] and chicken cathepsin L [8] are aligned so as to achieve maximal homology. For comparison, the amino acid sequences of cysteine proteinases are arbitrarily divided into 3 regions [6]: an amino-terminal (or active-site cysteinyl) region, a central one and a carboxyl-terminal (or active-site histidyl) region. All the sequences in the N-terminal region (70 residues in papain) contain the cysteine-rich site (C−G−S−C−W), where Cys^{25} is the active site cysteine. In this region, homologies between actinidin-papain, cathepsin H and cathepsin L-papain are the highest (59, 57 and 55%, respectively), while they are the lowest when the cathepsin B sequence is compared with the other cysteine proteinases (Table I). In addition, the buried acidic residues (Glu^{35} and Glu^{50} are also conserved, except for cathepsin B. The side chain charges of Glu^{35} and Glu^{50} are compensated for in actinidin and papain, by the charges of Lys^{17} and Lys^{174} side chains [5]. Furthermore, these residues (Glu^{35}, Glu^{50}, Lys^{17} and Lys^{174}) are conserved in cathepsins H and L, suggesting the evolutionary conservation of an inter−domain electrostatic interaction in plant

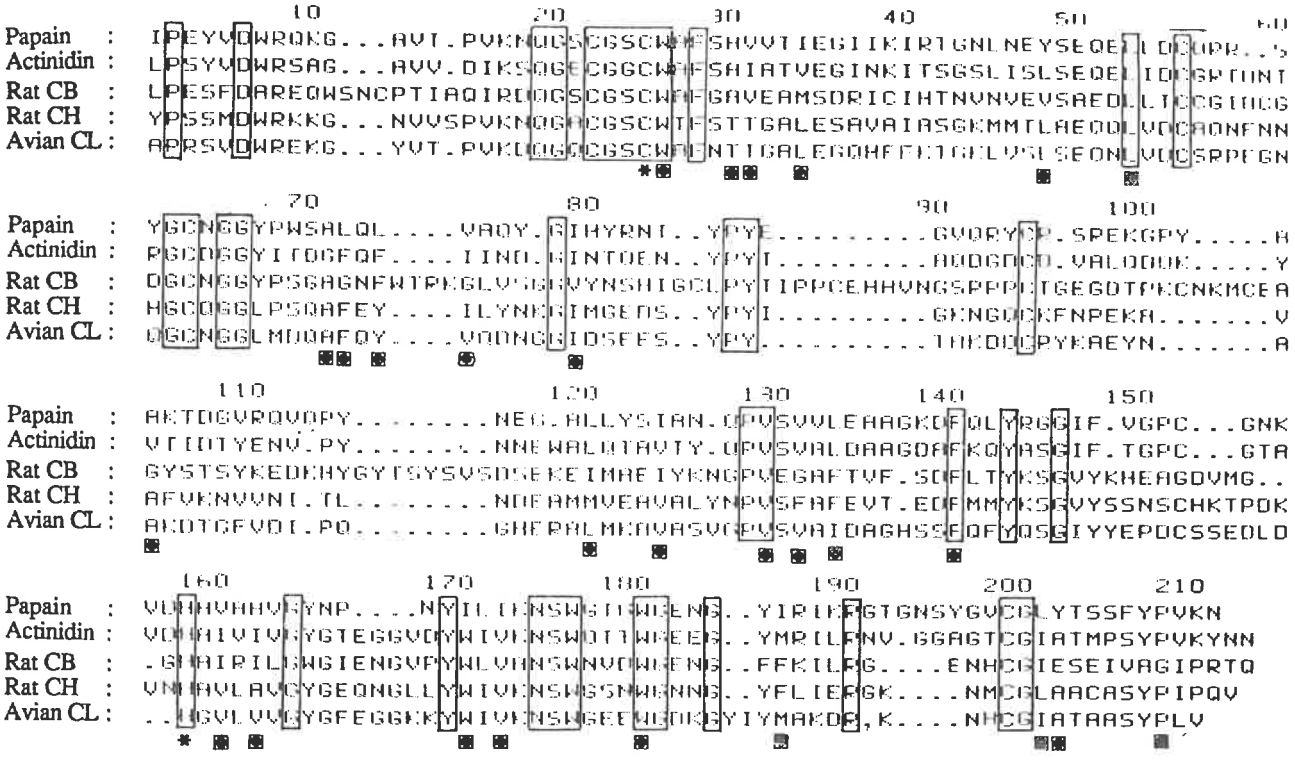

Fig. 1. Alignment of the sequences of papain, actinidin, rat cathepsin B (rat CB), rat cathepsin (rat CH) and chicken cathepsin L (avian CL). The residue numbers are those of papain. Identical residues are enclosed in boxes, ∗: active-site cysteine and histidine, ◆: amino acids forming papain and actinidin non-polar cores [18].

Table I. Amino acid sequence homologies (percentage) in cysteine proteinases.

Cysteine proteinase[a]	Region of comparison			
	N-terminal (1–70)[b]	central (71–117)[b]	C-terminal (118–212)[b]	whole protein (1–212)[b]
Avian CL / rat CH	57.5	35.7	51.7	50.5
Avian CL / rat CB	37.0	16.6	35.2	32.0
Avian CL / actinidin	45.8	28.0	55.2	46.7
Avian CL / papain	54.9	30.0	44.3	45.2
Rat CB / rat CH	31.1	13.6	39.8	31.5
Rat CB / papain	43.7	13.0	28.4	30.2
Rat CB / actinidin	32.8	18.2	36.6	31.3
Rat CH / papain	47.9	21.4	42.9	40.2
Rat CH / actinidin	45.7	22.7	41.6	39.0
Papain / actinidin	58.7	23.9	50.7	48.8

[a]Avian CL: chicken cathepsin L; rat CB: rat cathepsin B; rat CH: rat cathepsin H.
[b]Residues in papain.

cysteine proteinases and cathepsins H and L.

Homologies in the central region (residues 71–117 in papain) are low, especially between cathepsin B (insertion of 28 residues in this segment) and the other cysteine proteinases. However, 36% and 30% identities are observed between cathepsin L–cathepsin H and cathepsin L–papain, respectively. In the C-terminal region (residues 118–212 in papain), cathepsin L and actinidin show the highest degree of identity (55%). Only in this region does cathepsin B display comparable identity to the other cysteine proteinases, *i.e.*, cathepsin B–cathepsin H and cathepsin B–actinidin display 40 and 37% homologies, respectively. However, the segment after His^{159}, the active-site histidyl residue, has a lower degree of identity than the region surrounding the active-site cysteinyl residue, although the hydrophobic character is maintained in all the cysteine proteinases (Fig. 1). Among the 5 complete amino acid sequences, 20% of the residues are identical. Of the 37 invariant residues, 6 are cysteines (Fig. 1), including the active-site cysteinyl residue, while the others are probably involved in disulfide bridges, *i.e.*, according to X-ray data on papain [4], between Cys^{22}—Cys^{63} and Cys^{56}—Cys^{95}. However, Cys^{153}, which is retained in all sequences except in cathepsin B, suggests that the Cys^{153}–Cys^{200} disulfide bridge in papain does not exist in cathepsin B.

It is well known that papain and actinidin have broad pH optima ranging from 5.0 to 7.0 [2, 15], while cathepsins display maximum activity at pH 5.0–6.0 and are irreversibly inactivated above pH 7.0 [3]. For chicken cathepsin L, the profile of the curve of inactivation suggests that histidine residues could be involved in the process (Dufour *et al.*, submitted). Looking at the histidine content of cysteine proteinases, it appears that cathepsins B, H and L exhibit a higher histidine content than papain and actinidin. If we except the active-site histidine, with an atypical pK_a of 8.5–9.0 [16] and which is found in all cysteine proteinases, it is interesting to note that plant cysteine proteinases contain no (actinidin) or 1 (papain) histidine residue, while cathepsins B, H and L exhibit 7, 2 and 4 additional histidine residues in their sequences, respectively. Recently we showed by circular dichroism (Dufour *et al.*, submitted) that a modification of the pH from 5.8 to 7.0 produces a strong lowering of the α-helix content in chicken cathepsin L (40% at pH 5.8 and 17% at pH 7.0) and that the histidine residues are found in or near the α-helix regions (Dufour *et al.*, submitted) predicted by the algorithm of Garnier *et al.* We hypothesize that the inactivation of chicken cathepsin L observed at pH 7.0 results mainly in a change of the ionization state of one or more histidine residues.

The active site

The amino acids involved in catalysis, *e.g.*, Cys^{25} and His^{159}, as well as Gln^{19}, Asn^{175} and Trp^{177}

(Figs. 1 and 2), are retained in all cysteine proteinases. In fact, cathepsins B, H and L share with papain and actinidin the sequence Asn–Ser–Trp (residues 175–177 in papain). According to X-ray data of plant cysteine proteinases, Asn^{175} is hydrogen to the active-site histidine and Trp^{177} shields this bond from the solvent. Kamphuis *et al.* [17] have noted for papain and actinidin that the residues near the active site and those actually involved in the catalytic process surimpose to an extent that approaches the atomic coordinate accuracy of both structures. In view of this high degree of similarity in papain and actinidin active-site residues and the sequence homologies between plant and animal cysteine proteinases, papain, actinidin and cathepsins B, H and L probably have similar catalytic mechanisms. The changes in the active site do occur in the hydrophobic specificity 'pocket', the S_2 subsite [17]. For Baker [5], the replacement of Met^{205} in actinidin by Ser in papain (Table II) results in different surface characteristics of the S_2 subsite binding-site, explaining the preferential binding of aromatic side chains in papain to this subsite. Likewise, in the cathepsin B sequence, the residue 205 is a glutamic acid whose side chain may interact with 1 arginine side chain of the highly specific substrate Z–Arg–Arg–NMec [18].

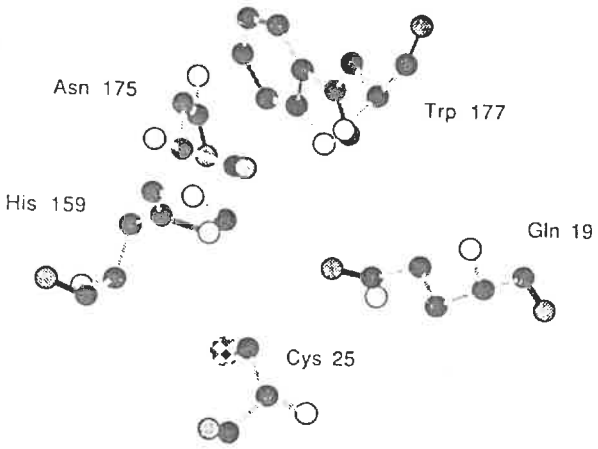

Fig. 2. Active site amino acid residues in cysteine proteinases.
O: nitrogen; ●: carbon; O: oxygen; ●: sulfur.
Atomic co-ordinates are those of papain and are from the Brookhaven Protein Data Bank (9PAP).

Using papain as a model, the characteristics of the cysteine proteinase active-site have been investigated by Schechter and Berger [19] who have shown that the enzyme could bind a peptide along a length of 7 residues. In addition, investigations of cysteine proteinase specificity [20–23] have shown that these enzymes preferentially cleave the peptide bonds with an amino acid residue, such as Phe, Ile or Leu, in position P_2, whereas the other subsites do not show a clear specificity. Nevertheless, specific substrates have been developed for cathepsin B (Z–Arg–Arg–NMec) and cathepsin H (Arg–NMec), but, there is still no specific substrate for cathepsin L.

The hydrophobic cores and the hydropathic profiles

As shown by X-ray studies, papain and actinidin molecules have a clearly defined double domain structure with the active site lying between the 2 domains. Each domain contains a core of buried non-polar side chains [17]. These are mostly Val, Leu or Ile, but some aromatic residues contribute too [5]. Nevertheless, these latter ones lie around the aliphatic residues and often one side of the aromatic ring is in contact with the core,

Table II. Amino acids around the S_2 subsite 'pocket' of cysteine proteinases.

Cysteine proteinase	Amino acid number in papain							
	67	68	131	133	157	160	162	205
Papain	Tyr	Pro	Ser	Val	Val	Ala	Ala	Ser
Actinidin	Tyr	Ile	Ser	Ala	Val	Ala	Val	Met
Rat cathepsin B	Tyr	Pro	Glu	Ala	Gly	Ala	Arg	Glu
Rat cathepsin H	Leu	Pro	Ser	Ala	Val	Ala	Leu	Cys
Chicken cathepsin L	Leu	Met	Ser	Ala	Leu	Gly	Leu	Ala

while the other side is accessible to the solvent [5]. In Fig. 1, the retention of hydrophobic cores in cysteine proteinases is studied using the sequence alignment described above. It appears that the non-polar residues forming the hydrophobic cores in plant cysteine proteinases are retained in cathepsins, except for cathepsin B which exhibits Glu^{203} instead of Ala^{203} in actinidin (Fig. 1). In addition, non-polar residues 72, 75, 105, 132 and 209 are replaced in the cathepsin B sequence by glycines. Nevertheless, even if the hydrophobic cores are conserved in cysteine proteinases, amino acid residues differ, thus residue 134 is Leu, Phe or Ile in actinidin, cathepsin H and cathepsin L sequences, respectively. The packing of the side chains in these areas allows for mutations conserving the hydrophobic character [24].

The hydrophobic profiles of cysteine proteinases (Fig. 3) evaluated by the method of Kyte and Doolittle [25] confirm these hypotheses. Indeed, in the case of soluble, globular proteins, there is a remarkable correspondence between the interior portions of their sequence and the region appearing on the hydrophobic side, as well as the exterior portions and the regions on the hydrophilic side [25]. In Fig. 3, it appears that the hydrophobic and hydrophilic regions in cysteine proteinase sequences are well conserved, *e.g.*, the segments located after the active-site cysteine (residue 25 in papain) and the active-site histidine (residue 159 in papain and 197 in cathepsin B) are hydrophobic in all proteins, while the middle regions (residues 80–120 in papain and 100–160 in cathepsin B) are largely hydrophilic. It has been shown for papain and actinidin that this region lies exposed on the protein surface. It should be noted that each of the principal helices has a polar and a non-polar side, *e.g.*, the central α-helix, residues 24–43 in actinidin, is completely buried, but its polar side chains form part of the distinctly polar interface between domains I and II, while the non-polar side chains are involved in the non-polar core of domain I [5]. On the other hand, the residues participating in the β-sheet structures are mostly hydrophobic.

Prediction of the secondary structures

The 2 algorithms of Garnier *et al.* [13] and Chou and Fasman [14] are used to predict the locations of α-helix and β-sheet segments in cathepsins B, H and L. The results, reported in Fig. 4, are compared with papain and actinidin secondary structure assignments from X-ray data [4, 5]. The percentage and the location of α-helices and β-sheets have been initially assessed using the algorithm of Garnier *et al.* (or GOR method) and taking the decision constants as being equal to zero. In reference to X-ray data, the method gives a low percentage of α-helices for papain and actinidin, *i.e.*, 5 and 13% of α-helices instead of 27 and 32%, respectively. Similarly for cathepsins B, H and L, the GOR method predicts 14, 24 and 16% helical content, respectively, whereas circular dichroism measurements give 33 and 40% α-helices for cathepsin B [12] and cathepsin L (Dufour *et al.*, submitted), respectively. Considering the disagreement observed between the predictive method of Garnier *et al.* and experimental data, an optimization of decision constants has been performed for cathepsins B, H and L, as suggested by Garnier *et al.* [13], using α-helix and β-sheet contents of cathepsins B and L deduced from circular dichroism investigations. Applied to papain, the optimized method of Garnier *et al.* raises the accuracy of the prediction for α-helices from 19 to 58% (Dufour *et al.*, submitted). The secondary structure of cathepsin H has not yet been investigated, nevertheless, cathepsin H and cathepsin L sequences are very similar and with decision constants equal to zero, the GOR method predicts that cathepsins H and L have the highest α-helix content among cysteine proteinases. One should note too that the optimized decision constants for cathepsins B and L are similar (cathepsin B: $DC_H = 100$ and $DC_E = 45$; cathepsin L: $DC_H = 130$ and $DC_E = 45$). This explains why cathepsin L decision constants have been applied to cathepsin H. Using these decision constants, the GOR method predicts 43% α-helices in the cathepsin H sequence; the highest helical content in cysteine proteinases. The method of Chou and Fasman gives similar results, *i.e.*, 27, 35 and 30% helical content for cathepsins B, H and L, respectively.

As reported in Fig. 4, the ordered structures predicted by the 2 methods are globally similar and show that the main α-helix and β-sheet segments of papain and actinidin are retained in the 3 cathepsins, *e.g.*, the active site helix (residues 25–42 in papain) is predicted, for the 3 cathepsins by both algorithms, as well as helices 50–56 and 114–127 (papain numbering). Likewise, the β-sheet segments around the active site histidine (157–167, 170–175 and 185–191 in

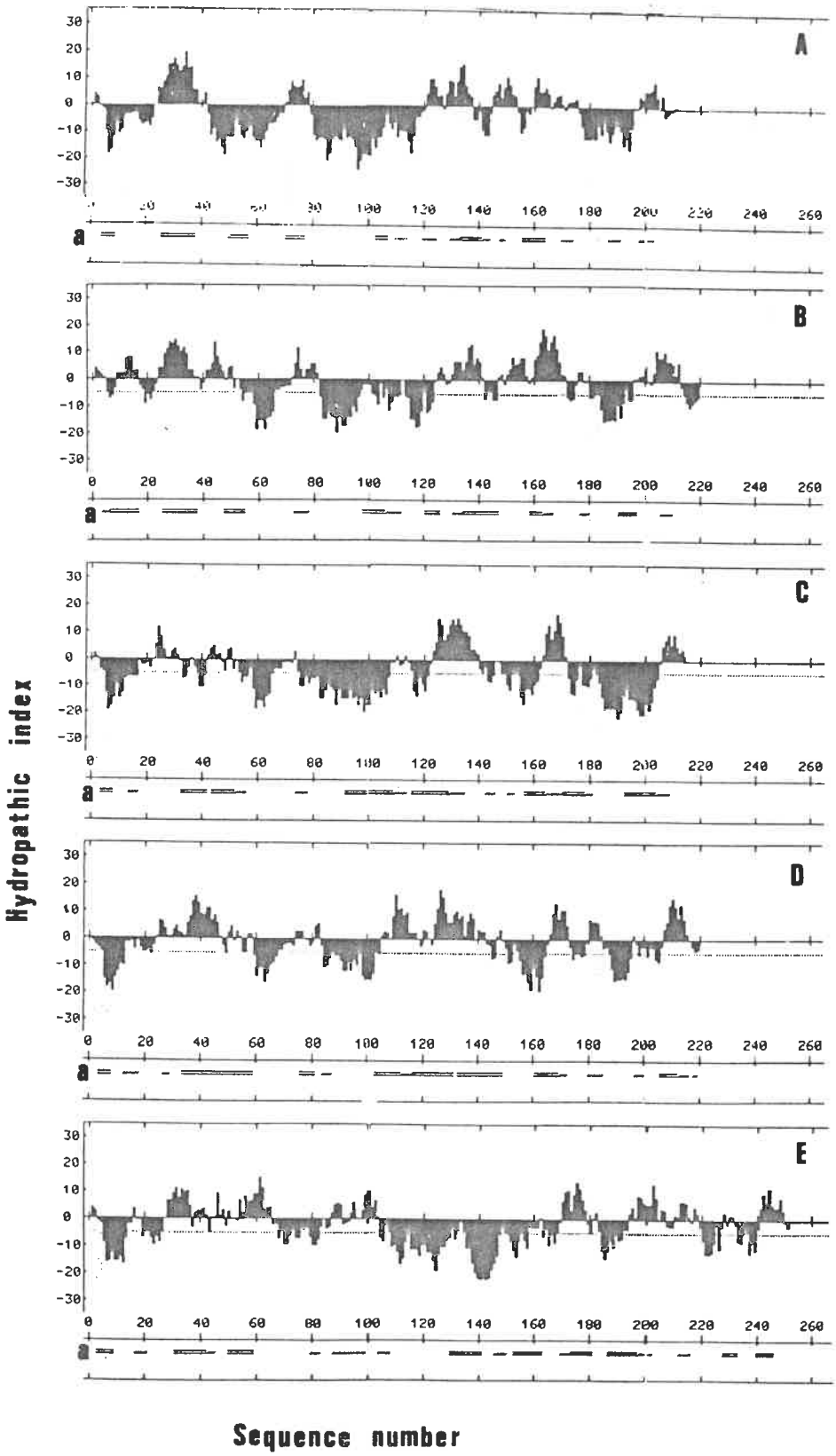

Fig. 3. Prediction of the hydropathic profiles of cysteine proteinases using the method of Kyte and Doolittle [25]: papain (A), actinidin (B), chicken cathepsin L (C), rat cathepsin H (D) and rat cathepsin B (E). All plots utilize a span setting of 9. [a]Secondary structures predicted by the method of Garnier *et al.*: ⩘: α-helix, =: β-sheet.

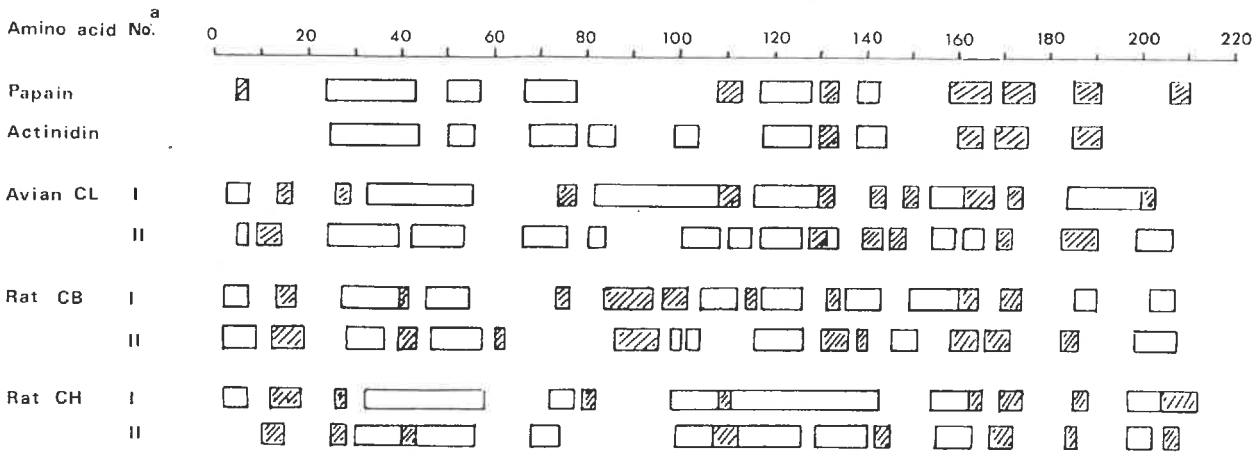

Fig. 4. Prediction of cathepsins B (rat CB), H (rat CH) and L (avian CL) secondary structures using the algorithms of Garnier *et al.* [13] (I) and Chou and Fasman [14] (II). Comparison with papain and actinidin secondary structure assignments from X-ray data [4, 5]. □ : α-helix, ▨ : β-sheet. Values of the optimized decision constants are: $DC_H = 100$ and $DC_E = 45$ for cathepsin B and $DC_H = 130$ and $DC_E = 45$ for cathepsins H and L.
[a]Residue numbers are those of papain.

papain) are well conserved in all the sequences. Furthermore, the most important changes in secondary structures occur in the middle region of cathepsins. As reported previously, the largest change in the backbone conformation for papain and actinidin occurs in segment 96–107, which has virtually no homology in sequence and lies exposed on the protein surface [17]. Indeed, most insertions / deletions and substitutions in cysteine proteinase sequences appear in the middle region (residues 71–117), suggesting that the main conformational differences occur in this region. It is thus clear that the insertion here of 28 residues in the cathepsin B sequence can occur without major changes in the overall molecular conformational and without affecting the conformational organization of active site residues [17]. In this region, both algorithms predict an additional α-helix for cathepsins H and L. When compared to papain, this additional helix (residues 85–108 in the cathepsin L sequence), predicted by the optimized method of Garnier *et al.*, agrees with the higher helical content of cathepsin L (40% instead of 27% in papain) deduced by circular dichroism investigations (Dufour *et al.*, submitted), whereas, for cathepsin H, the methods of Garnier *et al.* and Chou and Fasman predict a large α-helix in the region encompassing residues 103–150 (Fig. 4). It appears that chicken cathepsin L and rat cathepsin H, which have the highest percentages of sequence homologies among the 5 cysteine proteinases, display similar secondary structures.

In conclusion, the ordered structures in cysteine proteinases are probably well conserved in the N-terminal and C-terminal regions, although the differences occur essentially in the middle region, a place where sequences display the lowest homologies and which is far removed from the active site. As shown by Kamphuis *et al.* [17], the insertions and deletions in papain and actinidin sequences disturb the conformational homologies over a very limited range of 2–3 residues and apparently do not influence the major conformational characteristics. However, they can appreciably change the molecular surface properties of the proteins, *e.g.*, charge and solubility [8, 17, 26].

References

1 Cohen W.L., Coghlan V.M. & Dihel L.C. (1986) *Gene* 48, 219–227
2 McDowall M.A. (1970) *Eur. J. Biochem.* 14, 214–221
3 Kirschke H., Langner J., Riemann S., Wiederanders B., Ansorge S. & Bohley P. (1980) *in: Protein Degradation in Health and Disease* Excerpta Medica, Amsterdam pp. 15–35
4 Kamphuis I.G., Kalk K.H., Swarte M.B.A. &

Drenth J. (1984) *J. Mol. Biol.* 179, 233–256
5 Baker E.N. (1980) *J. Mol. Biol.* 141, 441–484
6 Takio K., Towatari K., Katunuma N., Teller D.C. & Titani K. (1983) *Proc. Natl. Acad. Sci. USA* 80, 3666–3670
7 Wada K., Takai T. & Tanabe T (1987) *Eur. J. Biochem.* 167, 13–18
8 Dufour E., Obled A., Valin C., Bechet D, Ribadeau-Dumas B. & Huet J.C. (1987) *Biochemistry* 26, 5689–5695
9 Docherty K., Carrell R.J. & Steiner F.D. (1982) *Proc. Natl. Acad. Sci. USA* 79, 4613–4617
10 Ryan R.E., Crissman J.D., Honn K.V. & Sloane B.F. (1985) *Cancer Res.* 45, 1–5
11 Portnoy D.A., Erickson A.H., Kochan J., Ravetch J.V. & Unkeless J.C. (1986) *J. Biol. Chem.* 261, 14697–14703
12 Bansal R., Singh S. & Kidway J.R. (1981) *Indian J. Biochem. Biophys.* 18, 110–113
13 Garnier J., Osguthorpe D.J. & Robson B. (1978) *J. Mol. Biol.* 120, 97–120
14 Chou Y.P. & Fasman G.D. (1974) *Biochemistry* 13, 222–245
15 Drenth J., Jansonius J.N., Kœkœk J.N. & Wolthers B.G. (1971) *Adv. Protein Chem.* 25, 79–115
16 Brocklehurst K., Baines B.S. & Kiersten M.P.J. (1981) *Top. Enzyme Ferment. Biotechnol.* 5, 262—335
17 Kamphuis I.G., Drenth J. & Baker E.N. (1985) *J. Mol. Biol.* 182, 317–329
18 Polgar L. & Csoma C. (1987) *J. Biol. Chem.* 262, 14448–14453
19 Schechter I. & Berger A. (1967) *Biochem. Biophys. Res. Commun.* 27, 157–162
20 Otto K. (1971) *in: Tissue Proteinases* (Barrett A.J. & Dingle J.T., eds.), North Holland Publishing Co., Amsterdam, pp. 181–207
21 Aronson N.N. & Barrett A.J. (1978) *Biochem. J.* 171, 759–765
22 Katanuma N., Towatari T., Tamai M. & Hanada K. (1983) *J. Biochem.* 93, 1129–1135
23 Bromme D., Bescherer K., Kirschke H. & Fittkau S. (1987) *Biochem. J.* 245, 381–385
24 Baker E.N. (1981) *in: Structural Studies on Molecules of Biological Interest* (Dodson G., Glusker J.P. & Sayre D., eds.), Clarendon Press, Oxford, pp. 339–349
25 Kyte J. & Doolittle F.R. (1982) *J. Mol. Biol.* 157, 105–132
26 Bechet D., Obled A. & Deval C. (1986) *Biosci. Rep.* 6, 991–997