Volume 303, number 1, 48–52 FEBS 11059 May 1992

Three-dimensional structural resemblance between leucine aminopeptidase and carboxypeptidase A revealed by graph-theoretical techniques

Peter J. Artymiuk[a], Helen M. Grindley[b], Julie E. Park[a], David W. Rice[a] and Peter Willett[b]

The Krebs Institute for Biomolecular Research, [a]Department of Molecular Biology and Biotechnology and [b]Department of Information Studies, The University, Sheffield, S10 2TN, UK

Received 17 March 1992; revised version received 7 April 1992

Using 3-D searching techniques based on algorithms derived from graph theory we have established a striking structural similarity between the structure of bovine carboxypeptidase A and that of the C-terminal domain of bovine leucine aminopeptidase. There is no significant sequence homology between the aminopeptidases and the carboxypeptidases but the strong structural relationship detected in this complex fold suggests that there may be a very remote divergent evolutionary relationship between these two enzyme classes.

Structural similarity; Graph theory; Carboxypeptidase; Aminopeptidase; Protein evolution

1. INTRODUCTION

Carboxypeptidase A is a monomeric peptidase of molecular weight 34,600 which binds one zinc ion and catalyses the hydrolysis of the carboxyl terminal residue from polypeptide chains [1]. The crystal structure of bovine carboxypeptidase A has been solved at 1.54 Å resolution [2], and detailed binding studies have revealed the nature of the substrate specificity pocket [3].

Leucine aminopeptidase (LAP) [4] belongs to a family of widely distributed exopeptidases that catalyse the hydrolysis of amino acids from the N-terminus of polypeptide chains. The molecule is a hexamer of identical subunits, each of which has MW 54,000 and binds two zinc ions. The crystal structure of LAP, both in the uncomplexed form and with the inhibitor bestatin bound, has recently been solved [5,6]. Each subunit consists of an N-terminal domain that is involved in trimer–trimer interactions, and a C-terminal domain which is involved in catalysis.

We have been engaged in the development of techniques derived from graph theory which enable the rapid detection of structural similarities between proteins [7–9] in the Protein Data Bank (PDB) [10,11]. Previous work used one type of graph-matching procedure, a subgraph isomorphism algorithm, to identify all protein structures that contained a user-defined structure motif [7,8]. In this work, the graph representations of proteins are searched using a modification of the maximal common subgraph algorithm of Bron and Kerbosch [12]. The program (which is named PROTE) allows very fast searches of all the structures in the PDB for similarities, either partial or complete, to the search protein structure [9].

In this paper we use these techniques to show that there is a strong 3-D structural relationship between the families of carboxypeptidases and aminopeptidases.

Abbreviations: LAP, leucine aminopeptidase; CPA, carboxypeptidase A; PDB, Protein Data Bank (Brookhaven); SSE, secondary structure element

Correspondence address: P.J. Artymiuk, The Krebs Institute for Biomolecular Research, Department of Molecular Biology and Biotechnology, The University, Sheffield, S10 2TN, UK.

2. MATERIALS AND METHODS

2.1. *Linear representations of helices and strands*

Regions of helix and strand in proteins in the Protein Data Bank (August 1991 release) were assigned using the algorithm of Kabsch and Sander [13]. The position and direction of each secondary structure element (SSE) was then approximated by a vector in 3-D space which corresponds to the axis of an idealized helix or strand superposed on the real helix or strand by least squares. The torsional angles, closest approach distances and distances between midpoints of each pair of SSEs within each protein in the PDB are stored in a database as a labelled graph. The nodes of the graph are the linear representations of the SSEs, and the edges of the graph the distances and angles between them [7].

2.2. *Detection of subgraph isomorphism*

The PROTE program [9] uses a maximal common subgraph algorithm [12] to match the query nodes, i.e. SSEs in the present context, to the structure nodes by looking at the relationships (graph edge values, within specified tolerances) between them. This permits the rapid location of any structural overlaps between the query structure and any of the other proteins in the PDB, and output is interfaced to the FRODO graphics program [14]. The program, which is now distributed by Tripos Associates Inc., has been extensively tested against

Published by Elsevier Science Publishers B.V.

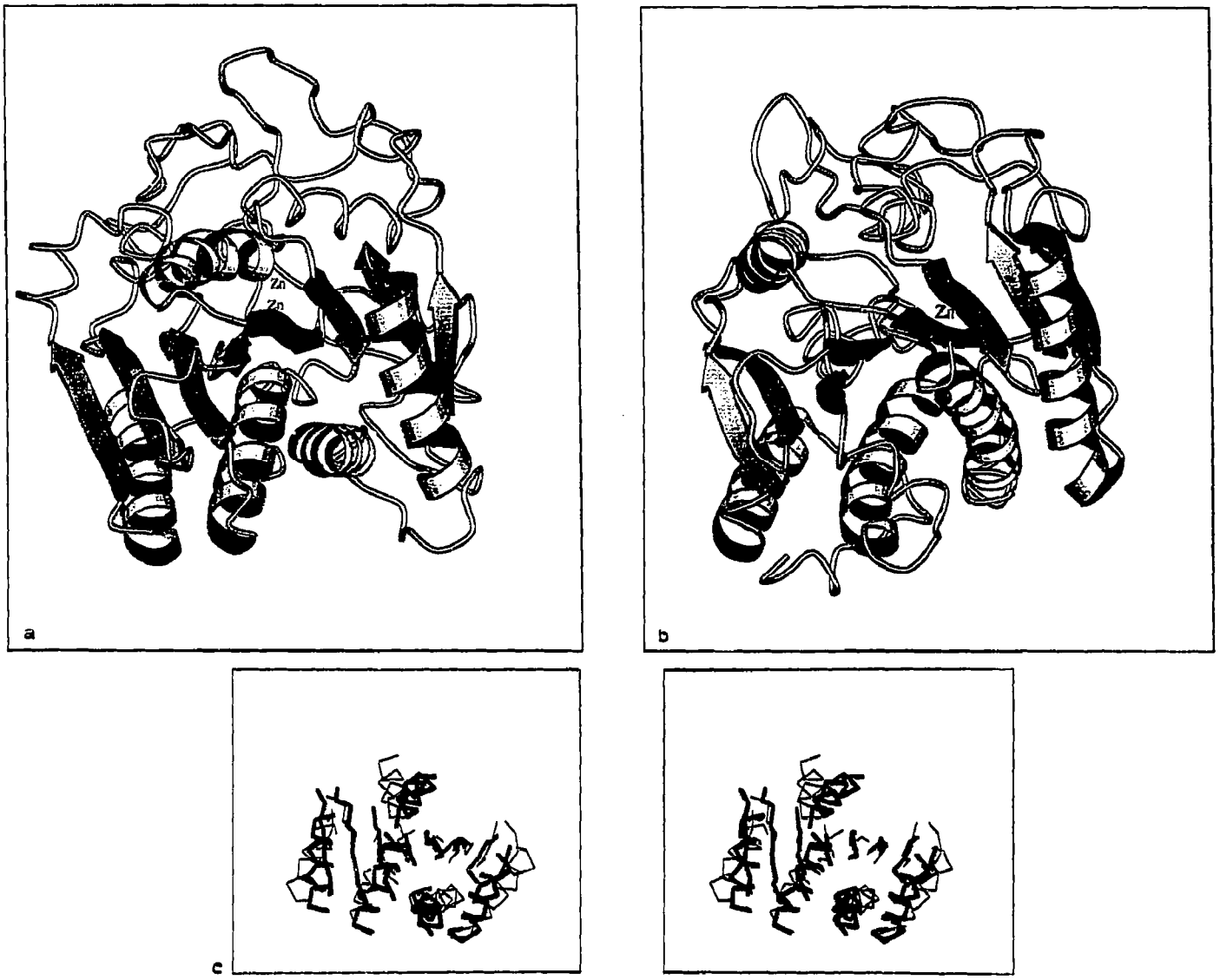

Fig. 1. C-α traces of (a) the C-terminal domain of LAP and (b) CPA produced using Kraulis' MOLSCRIPT program [18]. The helices and strands equivalenced in our study are represented as coiled ribbons and arrows, respectively; the non-equivalent parts of the structure are shown as a smoothed C-α trace. Figure (c) is a stereodiagram showing the superposition of C-αs of LAP (thick lines) and CPA (thin lines) in the equivalenced regions of the two structures. The view approximates to that used in the schematic diagram in Fig. 2.

a variety of known motifs, including the trypsin, azurin and globin families, and shown to operate correctly and effectively [9].

3. RESULTS

A search pattern consisting of the 42 helix and strand elements in the leucine aminopeptidase structure (PDB code 1LAP, from the October 1991 release of the PDB) was prepared and a search conducted of the August 1991 release of the Protein Data Bank. CPU time for the search was 40 min. on a Silicon Graphics 4D/210 VGX workstation. We monitored all structural similarities consisting of at least seven SSEs arranged in 3-D space within tolerances of 30° on inter-SSE angles and of the lesser of 40% or 4 Å on inter-SSE closest approach distances. As a further constraint, only those structural matches where SSEs occurred in the same sequence order in both the search protein and the PDB proteins were monitored. The most significant hits were in the three carboxypeptidase A structures (PDB codes: 3CPA, 4CPA, 5CPA) and carboxypeptidase B (1CPB). The hits in the carboxypeptidases A were the same in each case, and could be clustered together to form a region of similarity involving all 8 β strands in the main β sheet and 5 of the 10 α helices in the C-terminal domain of LAP (see Table I). For the purposes of the following discussion we consider only the overlap of 1LAP with 5CPA: the overlaps of 1LAP with 3CPA and 4CPA are essentially identical. Carboxypeptidase B shares 49% sequence identity with CPA [16], and the hit

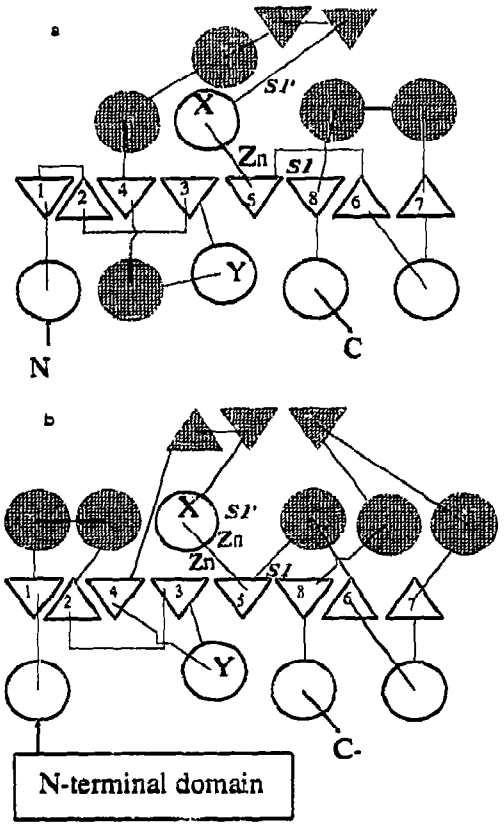

Fig. 2. Topological diagrams [15] of (a) CPA and (b) the C-terminal domain of LAP. Circles represent α helices and triangles represent β strands (apex down indicates the strand is viewed from the C terminus). Open circles and triangles indicate those SSEs that were found by the PROTE search to superpose in 3 dimensions, and shaded ones those that do not. The strands in the main 8-stranded β sheets are numbered according to their order in the sequence. Two of the helices are marked X and Y to facilitate discussion. Positions of the Zn^{2+} ions, which are at the rear of the β sheet in this view, are also indicated, as are the approximate positions of the S1 and S1′ subsites in the two enzymes.

found by PROTE was similar, but the available coordinates (1CPB, [17]) are a preliminary C-α only set and not suitable for further comparison.

4. DISCUSSION

4.1. *Comparison of the carboxypeptidase A and leucine aminopeptidase folds*

An initial superposition of the regions of similarity produced by PROTE was refined by examination of the 1LAP and 5CPA structures on the molecular graphics [14] to give an overall rms deviation over 86 equivalenced C-α atoms in the core regions of the two structures (see Table I) of 1.77 Å. This superposition is shown in Fig. 1 and the superposition matrix is given in Table II. In Fig. 2 the topological similarity between the two folds is illustrated schematically: this shows that the main eight stranded β sheet has an identical organization and order of strands, and furthermore that five helices are also conserved in position, although one of these helices (labelled 'X' in Fig. 2) superposes less well than the others. The areas of 3-D dissimilarity are mainly in the upper part of the diagram in Fig. 2: these are associated with the presence of a double Zn^{2+} site in LAP rather than a single one as in CPA, and with inter-subunit contacts around the 3-fold and 2-fold symmetry axes in LAP which have no counterpart in the monomeric structure of CPA.

4.2. *Comparison of active sites*

For both proteins the zinc-binding sites and the active site are situated at the C-terminal ends of the 4 central strands of the 8-stranded β sheet (see Fig. 2). However, it must be noted that our superposition does not preserve the positions of the zinc ions in 3-D space, which are displaced by almost 10 Å because the loops at the ends of the central β strands are more extended in LAP

Table I

(a) The equivalent helices and strands in LAP and CPA. The first two columns of figures give the SSEs equivalenced in the initial PROTE search, and the final two columns list the equivalent Cαs used to determine the final superposition matrix given in Table II

	PROTE search		C-α superposition	
	LAP	CPA	LAP	CPA
α helix	179–195	14–29	182–189	17–24
β strand 1	197–204	32–37	196–200	31–35
β strand 2	227–236	46–53	227–235	45–53
β strand 3	245–248	60–67	242–250	59–67
α helix Y	274–290	72–90	275–291	74–90
β strand 4	294–301	103–109	295–300	104–109
α helix X	333–349	173–187		
β strand 5	352–359	191–197	351–358	189–196
β strand 6	372–378	200–205	373–377	200–204
α helix	379–394	215–235	383–390	223–230
β strand 7	396–400	238–243	397–399	238–240
β strand 8	444–450	264–271	444–449	265–270
α helix	471–484	288–305	472–473	290–291

Table II

Rotation matrix used to rotate 5CPA into 1LAP

$$X' = \begin{pmatrix} 0.77158 & 0.27094 & -0.57554 \\ -0.61565 & 0.09039 & -0.78281 \\ -0.16007 & 0.95834 & 0.23655 \end{pmatrix} \cdot X_{CPA} + \begin{pmatrix} 58.496 \\ 51.908 \\ -8.057 \end{pmatrix}$$

Where X_{CPA} are column vectors representing the CPA coordinates in the 5CPA frame and X' the same coordinates superposed on the 1LAP structure.

than in CPA. Nevertheless there are considerable *topological* equivalences between the zinc sites in the two enzymes: the single Zn^{2+} in CPA and the first Zn^{2+} in LAP both receive a ligand from strand 3 (His^{69} and Lys^{250}, respectively) and from helix Y (Glu^{72} and Aps^{273}, respectively). However, in CPA the Zn^{2+} is also ligated by His^{196} from strand 5, whereas in LPA the two zinc ions are ligated by Asp^{332} and Glu^{334} from helix X.

Preliminary evidence for the mechanism of LAP has been gleaned from inhibitor-binding studies but this has not yet led to definitive proposals for a mechanism [6], and therefore detailed mechanistic comparisons are not possible at this stage. However, comparisons of subsites S1 and S1′ (lying on the N-terminal and C-terminal sides of the scissile bond, respectively) in CPA [19] with their putative equivalents in LAP [6] shows that the S1 sites in the two proteins lie in roughly equivalent positions between strands 5 and 8 in LAP and above strand 8 in CPA. However, although the S1′ sites both lie above the β sheet in CPA and LAP (Fig. 2), completely different parts of the chains are involved in the interactions. Nevertheless, in very broad terms, the polypeptide substrate appears to run in approximately the same direction with respect to the folds of LAP and CPA, although the detailed interactions are different.

4.3. *Sequence similarity*

We find that even in the regions of structural overlap, the sequence identity is only 7% between the LAP and CPA sequences, which means there is no detectable sequence similarity. This is in agreement with Burley et al. [5,6], who also observed that there was no sequence similarity between carboxypeptidases and leucine aminopeptidase, but did not remark upon the overall 3-D similarity we have described in this paper.

4.4. *Evolutionary implications*

Structural similarities of this kind can have a number of explanations [8]: in terms of convergent evolution towards a common stable fold, in terms of convergent evolution towards a structure dictated by a particular functional requirement, or in terms of arguments based on divergent evolution. Thus one possible explanation for the structural similarity between LAP and CPA is that the same fold occurs in both proteins because it is a particularly stable or simple fold, as is the case, for example, in the α/β barrel proteins [20]. In the present case, such an explanation appears improbable since the common folding motif between LAP and CPA is complex and involves a complicated mixed parallel and antiparallel β sheet as well as several α helices (see Fig. 2), with a topology hitherto only found in carboxypeptidases.

It is theoretically possible that a common function could lead through convergent evolution to similar structural features: this can certainly occur at the level of sidechain orientations, as for example the Asp-His-Ser catalytic triad shared by the trypsin and subtilisin families of serine proteases and the lipases (reviewed in [21]). However, there are no unambiguously established cases of this occurring at the level of the gross tertiary structure of proteins, which is the situation encountered in the present comparison. Nevertheless, this may be a valid explanation for the similarity between LAP and CPA as in this case the two enzymes have related functions: CPA is a Zn peptidase, and the marked structural resemblance is restricted to the Zn-binding C-terminal domain of LAP which is responsible for the enzyme's peptidase activity.

There does, however, remain one other intriguing possibility: although there is no significant sequence similarity that can be cited as irrefutable evidence of divergent evolution from a common ancestor, it has been noted by Matthews and Rossmann [22] that during the evolution of proteins, tertiary structure is conserved more than amino acid sequence. On this view it is therefore conceivable that the striking structural similarity we have found may be indicative of a very remote, previously unsuspected, divergent evolutionary relationship between the families of aminopeptidases and carboxypeptidases.

Acknowledgements: This work was supported by the Science and Engineering Research Council, The Wellcome Trust, Tripos Associates Inc. and the Department of Education and Science. The Krebs Institute is a designated centre of the SERC Molecular Recognition Initiative. We acknowledge the use of the SERC SEQNET facility. P.J.A. is a Royal Society University Research Fellow and D.W.R. is a Lister Institute Research Fellow.

REFERENCES

[1] Lipscomb, W.N. (1980) Proc. Natl. Acad. Sci. USA 77, 3875-3878.

[2] Rees, D.C., Lewis, M. and Lipscomb, W.N. (1983) J. Mol. Biol. 168, 367-387.

[3] Rees, D.C. and Lipscomb, W.N. (1981) Proc. Natl. Acad. Sci. USA 78, 5455–5459.
[4] Hanson, H. and Frohne, M. (1976) Methods Enzymol. 45, 504–521.
[5] Burley, S.K., David, P.R., Taylor, A. and Lipscomb, W.N. (1990) Proc. Natl. Acad. Sci. USA 87, 6878–6882.
[6] Burley, S.K., David, P.R. and Lipscomb, W.N. (1991) Proc. Natl. Acad. Sci. USA 88, 6916–6920.
[7] Mitchell, E.M., Artymiuk, P.J., Rice, D.W. and Willett, P. (1990) J. Mol. Biol. 212, 151–166.
[8] Artymiuk, P.J., Rice, D.W., Mitchell, E.M. and Willett, P. (1990) Protein Engineering 4, 39–43.
[9] Artymiuk, P.J., Grindley, H.M., Rice, D.W., Ujah, E.C. and Willett, P. (1991) in: Proceedings of the 1991 Chemical Information Conference (H. Collier, Ed.), Infonortics Ltd., Calne, pp. 91–106.
[10] Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer Jr., E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, M. and Tasumi, M. (1977) J. Mol. Biol. 112, 535–542.
[11] Abola, E.E., Bernstein, F.C., Bryant, S.H., Koetzle, T.F. and Weng, J. (1987) in: Crystallographic Databases – Information Content, Software Systems, Scientific Applications (F.H. Allen, G. Bergerhoff and R. Sievers, Eds.), Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester, pp. 107–132.
[12] Bron, C. and Kerbosch, J. (1973) Communications of the A.C.M. 16, 575–577.
[13] Kabsch, W. and Sander, C. (1983) Biopolymers 22, 2577–2637.
[14] Jones, T.A. (1985) Methods Enzymol. 115, 157–171.
[15] Sternberg, M.J.E. and Thornton, J.M. (1975) J. Mol. Biol. 105, 367–382.
[16] Titani, K., Ericsson, L.H., Walsh, K.A. and Neurath, H. (1975) Proc. Natl. Acad. Sci. USA 72, 1666–1670.
[17] Schmid, M.F. and Herriott, J.R. (1976) J. Mol. Biol. 103, 175–190.
[18] Kraulis, P.J. (1991) J. Appl. Cryst. 24, 946–950.
[19] Christianson, D.W., David, P.R. and Lipscomb, W.N. (1987) Proc. Natl. Acad. Sci. USA 84, 1512–1515.
[20] Chothia, C. (1988) Nature 333, 598–599.
[21] Blow, D.M. (1990) Nature 343, 694–695
[22] Matthews, B.W. and Rossmann, M.G. (1985) Methods Enzymol. 115, 397–420.