Hydrophobic-cluster analysis of plant protein sequences

A domain homology between storage and lipid-transfer proteins

Bernard HENRISSAT,*‡ Yves POPINEAU* and Jean-Claude KADER†

*Laboratoire de Biochimie et Technologie des Protéines, Institut National de la Recherche Agronomique, Rue de la Géraudière, F-44000 Nantes, France, and †Laboratoire de Physiologie Cellulaire (UA 1180), Université Pierre et Marie Curie, Tour 53, E3, 4 Place Jussieu, F-75252 Paris Cedex 05, France

Hydrophobic-cluster analysis was used to characterize a conserved domain located near the C-terminal amino acid sequence of wheat (*Triticum aestivum*) storage proteins. This domain was transformed into a linear template for a global search for similarities in over 5200 protein sequences. In addition to proteins that had already been found to exhibit homology to wheat storage proteins, a previously unreported homology was found with non-specific lipid-transfer proteins from castor bean (*Ricinus communis*) and from spinach (*Spinacia oleracea*) leaf. Hydrophobic-cluster analysis of various members of the present protein group clearly shows a typical domain structure where (i) variable and conserved domains are located along the sequence at precise positions, (ii) the conserved domains probably reflect a common ancestor, and (iii) the unique properties of a given protein (chain cut into subunits, repetitive domains, trypsin-inhibitor active site) are associated with the variable domains.

INTRODUCTION

Hydrophobic-cluster analysis (HCA) has recently proven to be a very powerful and straightforward method for comparing and aligning protein sequences [1,2]. In the case of distantly related proteins, this technique, derived from the theory of Lim [3], helps in finding homologous domains hardly detectable otherwise. We have applied this method to a group of seed storage proteins whose homology is difficult to assess with classical methods because of large variations in their lengths (from less than 100 to more than 600 amino acid residues). Plant seeds are a major source of proteins that are believed to have no other physiological role than the supply of nitrogen during germination [4]. Prolamins are the storage proteins of cereal seeds. They are polymorphic, and can be divided into several groups each containing structurally related proteins that are encoded by a complex multigenic locus [5]. Wheat (*Triticum aestivum*) prolamins can be classified into genetically distinct groups such as α -, γ - and ω -gliadins and high- M_r and low- M_r glutenins [6]. Each of these groups is constituted of homologous proteins that also display inter-group homologies [6]. Particularly, three domains have been identified as being conserved not only among wheat, rye (*Secale cereale*) and barley (*Hordeum vulgare*) prolamins, but also among other seed proteins from some monocotyledonous and dicotyledonous plants [7]. It has thus been suggested that the structural genes for all these proteins evolved from a single ancestral gene probably encoding a proteinase inhibitor [7].

We have applied HCA to characterize a highly conserved domain located near the C-terminus of wheat prolamins. This domain has been transformed into a linear 'template' for a global search for similarities in

over 5200 protein sequences. In the present paper we report the finding that lipid-transfer proteins from plants show significant domain homology with prolamins and other plant storage proteins.

METHODS

Computer analysis of sequences

Sequence data were stored, manipulated and analysed with PC/Gene programs (Genofit, Geneva, Switzerland) running on an IBM-PC compatible microcomputer. Scans through the SWISS-PROT protein sequence data bank [8] were performed with the PESEARCH program, where the sub-sequence to be searched can be defined with all possible ambiguities. Programs FSTPSCAN and SCANSIM were then used to evaluate homologies on the set of sequences found with PESEARCH. FSTPSCAN is based on the method developed by Lipman, Pearson and Wilbur [9,10]. SCANSIM uses the method of Needleman & Wunsch [11] as implemented by Dayhoff [12] and Feng *et al.* [13]. Both programs were used with matrix scoring systems that take into account the structural similarities of amino acids as well as the likelihood of interchange [13], and also the statistics of the accepted point mutation in closely related protein sequences [14].

HCA plot

A simplified two-dimensional sequence representation was done as described in ref. [1] except that here an α -helix having 3.66 amino acid residues per turn was used. In our representation after three turns residues i and $i+11$ have similar positions, whereas in the original HCA plot residues i and $i+18$ have similar positions after five turns [1]. Fig. 1 shows that the simplified HCA

Abbreviation used: HCA, hydrophobic-cluster analysis.

‡ To whom correspondence should be sent, at present address: Centre de Recherches sur les Macromolécules Végétales, C.N.R.S., B.P. 68, F-38402 St. Martin d'Hères, France.

Fig. 1. Comparison between the original HCA plot [1] (a) and the HCA plot used in the present work (b)

plot is almost superimposable on the original HCA plot. The simplified representation was tested with human haemoglobin (α -chain) and lupin (*Lupinus luteus*) leg-haemoglobin and compared with the results obtained with the original HCA plot [1]. Both methods gave identical results in terms of distribution, size, shape and orientation of the clusters (results not shown). The advantage of the present method is that the two-dimensional representation used here can be done simply with a word processor operating from a microcomputer.

RESULTS AND DISCUSSION

The HCA plots of the C-terminal sequences of various wheat prolamins are shown in Fig. 2. This zone corresponds to the conserved domain identified and named domain C by Kreis *et al.* [7]. Examination of these plots reveals that this domain can be further divided into two sub-domains, one being rich in alanine and/or glutamine whereas the other is characterized by a Leu-Pro-Xaa-Xaa-Cys-Xaa-Haa sequence, where Haa is a hydrophobic residue. It is flanked at its C-terminal side by a 'cluster-breaking group' [1] such as Pro or Gly-Gly found within the first five amino acid residues. Screening the SWISS-PROT sequence data bank (supplemented by a number of recently published sequences) with a Leu-Pro-Xaa-Xaa-Cys-Xaa-Haa 'template' resulted in the selection of a number of proteins. In addition to the

Fig. 2. HCA plots of the C-terminal sequences of wheat storage proteins

(a) α -Gliadin [15-17]; (b) γ -gliadin [18]; (c), high- M_r glutenin [19]; (d) low- M_r glutenin [20]. In each case the asterisk (*) indicates the C-terminal residue. Hydrophobic residues are circled. Proline and cysteine are boxed in black and grey respectively. Vertical lines show the limits of the conserved domain. AQ indicates the Ala/Gln-rich zone.

wheat prolamins that were used to define the template, 28 other sequences have been found out of more than 5200 (Table 1). Since the template for the search was short and had a few ambiguities, a number of possibly unrelated proteins might have been wrongly selected and another selection was devised. Programs FSTPSCAN and SCANSIM, based solely on amino acid similarities and statistics of interchange, failed to discriminate the various proteins further. However, the absence of the Ala/Gln-rich domain and/or the absence of cluster-breaking residues at the extremity of the conserved sequence were successfully used to identify unrelated sequences (Table 1), and this resulted in the elimination of 18 proteins. The sequences that satisfied all the criteria are all from plant origin and they all bear the conserved

Table 1. Protein sequences found to have an L-P-X-X-C-X-(V or I or L or M or F or W) sub-sequence

For conciseness, the wheat protein sequences that had been used for defining the sub-sequence have been omitted. The scan was performed with the program PESEARCH on the SWISS-PROT sequence data bank [6] manually supplemented by the sequences denoted by an asterisk (*). The A + Q entry represents the amount of Ala + Gln found within the 20 residues preceding the sub-sequence. The P + GG entry represents the amount of Pro or Gly-Gly found within the five residues following the sub-sequence. Proteins were selected if $A + Q \geq 4$ and $P + GG \geq 1$.

Protein	Origin	A + Q	P + GG
γ -Secalin* [7]	Rye (<i>Secale cereale</i>) seed	6	1
Hordein B1* [21]	Barley (<i>Hordeum vulgare</i>) seed	7	1
Hordein B3* [21]	Barley (<i>Hordeum vulgare</i>) seed	6	1
Lipid-transfer protein* [22]	Spinach (<i>spinacea oleracea</i>) leaf	5	1
Lipid-transfer protein* [23]	Castor bean (<i>Ricinus communis</i>)	7	1
Napin 1	Rape (<i>Brassica napus</i>) seed	8	1
Napin 2	Rape (<i>Brassica napus</i>) seed	8	1
Glutelin	Maize (<i>Zea mays</i>) seed	12	1
Storage protein	Brazil nut (<i>Bertholletia excelsa</i>)	4	1
Storage protein	Castor bean (<i>Ricinus communis</i>)	9	1
Acetolactate synthase III	<i>Escherichia coli</i>	5	0
13000- M_r protein	<i>Escherichia coli</i>	3	0
E6 protein	Rabbit papillomavirus	3	0
Lutropin β -chain	Human	3	0
Trypsin inhibitor	Maize (<i>Zea mays</i>)	2	0
Trypsinogen	Spiny dogfish (<i>Squalus acanthias</i>)	2	0
Keratin B2A	Sheep	1	2
Oligopeptide permease	<i>Salmonella typhimurium</i>	1	2
Keratin B2B	Sheep	1	1
Keratin B2C	Sheep	1	1
Non-structural proteins	Sindbis virus	1	1
Acid phosphatase	<i>Saccharomyces cerevisiae</i>	1	0
2A protein	Cucumber-mosaic virus	1	0
Keratin 48000- M_r type 1	Sheep	0	1
Von Willebrand factor	Human	0	1
E1B 55000- M_r protein	Adenovirus 2	0	0
E1B 55000- M_r protein	Adenovirus 2	0	0
Fumarate reductase	<i>Escherichia coli</i>	0	0

domain at their C-termini. The selected sequences comprise a series of storage proteins whose homology had already been established [7,24,25], but quite unexpected was the finding, among these sequences, of two lipid-transfer proteins from spinach (*Spinacea oleracea*) leaf [22] and from castor bean [23].

HCA plots of the two lipid-transfer proteins are presented in Figs. 3(a) and 3(b). They were divided into domains numbered from I to IX. The strong homology of these proteins is obvious and probably reflects a similar folding. This is most probably related to the functional property shared by these two proteins, which promote lipid interchange between membranes [26]. This unusual function involves a reversible binding of lipids in the protein; this has been already observed with lipid-transfer proteins from castor bean, spinach and maize [27–29]. Comparison with the HCA plot of the seed storage protein from castor bean (Fig. 3c) reveals that the homology is not limited to the C-terminal domain, but spans throughout the whole sequence. Domains II, IV, VI and VIII (with the exception of its C-terminal extremity) are well conserved and alternate with domains that display variability (domains I, III, VII and IX) or that are absent (domain V). The same conservation/divergence pattern is also observed with the other proteins of this family, as illustrated by a high- M_r glutenin (Fig. 3d) and a trypsin inhibitor from maize

(Fig. 3e). However, in this last case the polypeptide chain seems to be interrupted within variable domain VII, since domain VIII could not be detected. This explains why no trypsin inhibitor was found in the above data-bank screening, since the template that was used corresponded to domain VIII. It is worth mentioning that a significant domain homology was also found between the above proteins and a soya-bean (*Glycine max*) hydrophobic protein (Fig. 3f). Similarly to the trypsin inhibitor, this soya-bean protein sequence seems to be interrupted within domain VII, since domain VIII is absent. This illustrates the capabilities of HCA as compared with classical methods, which failed to detect sequences significantly homologous to the soya-bean hydrophobic protein, which was therefore reported to represent a hitherto unknown protein family [34].

The present protein family appears to be characterized by a peculiar domain structure where variable and conserved domains are located along the sequences at precise positions. The variable domains seem to give rise to the unique properties of a given protein. For example, the active part of the maize trypsin inhibitor (Arg-36) is found within variable domain III. In the case of proteins with subunits derived from a single polypeptide precursor, chain cleavage giving rise to the subunits occurs in the variable domains III or VII. In a similar fashion, repetitive domains are found within domain VII (high-

Fig. 3. HCA plots of various plant protein sequences

(a) Lipid-transfer protein (spinach leaf) [22]; (b) lipid-transfer protein (castor bean) [23]; (c) storage protein (castor bean) [30,31]; (d) high- M_r glutenin (wheat) [32]; (e) trypsin inhibitor (maize) [33]; (f) hydrophobic protein (soya-bean) [34]. Hydrophobic residues are circled. Proline and cysteine residues are represented by black and grey boxes respectively. Vertical lines indicate the proposed correspondences between the various sequences. The area marked with stars (★) denotes a repeat domain of approx. 540 residues.

M_r glutenins) or within domains I and VII (gliadins, low- M_r glutenins, maize glutelin). On the other hand, apart from reflecting a probable common ancestor, the role of well-conserved domains among these proteins of different origins, size and properties remains unclear.

In the case of the lipid-transfer proteins from plants, nothing is known on the location of the lipid-binding site. However, it is reasonable to assume that such a site should be conserved in the two lipid-transfer proteins examined here, and, conversely, should be absent from the other proteins, which lack the lipid-transfer function. Our observations on the different domains of spinach and castor-bean proteins led us to consider that a possible binding site for lipids could correspond to domain VII or to the C-terminal extremity of domain VIII or to a combination of both.

CONCLUSION

A drawback that had been identified in the HCA is the manual handling of the plots, which prevents direct data-bank screening [1]. This was overcome here by the linearization of the information found within a given domain, followed by a classical search in a data bank. The present work shows that HCA is well adapted to the comparison of proteins displaying homologous domains separated by variable domains of any length.

The two lipid-transfer proteins identified here not only arise from different plants (castor bean and spinach) but also from different plant organs (seed and leaf). Their good overall homology indicates that these lipid-transfer proteins are members of a family of structurally related proteins, as are cytochromes, lysozymes etc. Since these proteins also show domain homology with cereal prolamins, they also belong to the superfamily defined by Kreis *et al.* [7]. It is therefore tempting to conclude that certain seed storage proteins and the lipid-transfer proteins examined here derive from a common ancestor. Whether the sequence homology between these proteins can be related to a folding and/or a functional similarity remains, however, to be demonstrated.

We thank Dr. J.-P. Mornon for stimulating discussions throughout this work.

REFERENCES

- Gaboriaud, C., Bissery, V., Benchetrit, T. & Mornon, J. P. (1987) *FEBS Lett.* **224**, 149–155
- Benchetrit, T., Bissery, V., Mornon, J. P., Devault, A., Crine, P. & Roques, B. P. (1988) *Biochemistry* **27**, 592–596
- Lim, V. I. (1974) *J. Mol. Biol.* **88**, 857–872
- Spencer, D. (1984) *Philos. Trans. R. Soc. London B* **304**, 275–285
- Shewry, P. R., Mifflin, B. J. & Kasarda, D. D. (1984) *Philos. Trans. R. Soc. London B* **304**, 297–308
- Shewry, P. R., Tatham, A. S., Forde, J., Kreis, M. & Mifflin, B. J. (1986) *J. Cereal Sci.* **4**, 97–106
- Kreis, M., Forde, B. G., Rahman, S., Mifflin, B. J. & Shewry, P. R. (1985) *J. Mol. Biol.* **183**, 499–502
- Bairoch, A. (1987) *SWISS-PROT Protein Sequence Data Bank*, Release 5.0, Medical Biochemistry Department, University of Geneva, Geneva
- Lipman, D. J. & Pearson, W. R. (1985) *Science* **227**, 1435–1441
- Wilbur, W. J. & Lipman, D. J. (1983) *Proc. Natl. Acad. Sci. U.S.A.* **80**, 726–730
- Needleman, S. B. & Wunsh, C. D. (1970) *J. Mol. Biol.* **48**, 443–453
- Dayhoff, M. O. (1978) in *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., ed.), vol. 5, suppl. 3, pp. 1–8, National Biomedical Research Foundation, Washington
- Feng, D. F., Johnson, M. S. & Doolittle, R. F. (1985) *J. Mol. Evol.* **21**, 112–125
- Schwartz, R. M. & Dayhoff, M. O. (1978) in *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., ed.), vol. 5, suppl. 3, pp. 353–358, National Biomedical Research Foundation, Washington
- Rafalski, J. A., Scheets, K., Metzler, M., Peterson, D. M., Hedgcock, C. & Söll, D. G. (1984) *EMBO J.* **3**, 1409–1415
- Sumner-Smith, M., Rafalski, J. A., Sugiyama, T., Stoll, M. & Soell, D. (1985) *Nucleic Acids Res.* **13**, 3905–3916
- Anderson, O. D., Litts, J. C., Gautier, M.-F. & Greene, F. C. (1984) *Nucleic Acids Res.* **12**, 8129–8144
- Sugiyama, T., Rafalski, A., Söll, D. (1986) *Plant Sci.* **44**, 205–209
- Forde, J., Forde, B. J., Fry, R. P., Kreis, M., Shewry, P. R. & Mifflin, B. J. (1983) *FEBS Lett.* **162**, 360–366
- Okita, T. W., Cheesbrough, V. & Reeves, C. D. (1985) *J. Biol. Chem.* **260**, 8203–8213
- Forde, B. J., Kreis, M., Williamson, M. S., Fry, R. P., Pywell, J., Shewry, P. R., Bunce, N. & Mifflin, B. J. (1985) *EMBO J.* **4**, 9–15
- Bouillon, P., Drishel, C., Vergnolle, C., Duranton, H. & Kader, J. C. (1987) *Eur. J. Biochem.* **166**, 387–391
- Takishima, K., Watanabe, S., Yamada, M. & Mamiya, G. (1986) *Biochim. Biophys. Acta* **870**, 248–255
- Reeck, G. R. & Hedgcock, C. (1985) *FEBS Lett.* **180**, 291–294
- Prat, S., Cortadas, J., Puigdomenech, P. & Palau, J. (1985) *Nucleic Acids Res.* **13**, 1493–1504
- Kader, J. C. (1985) *Chem. Phys. Lipids* **38**, 51–62
- Kader, J. C., Julienne, M. & Vergnolle, C. (1984) *Eur. J. Biochem.* **139**, 411–416
- Douady, D., Grosbois, M., Guerbette, F. & Kader, J. C. (1982) *Biochim. Biophys. Acta* **710**, 143–153
- Rickers, J., Spener, F. & Kader, J. C. (1985) *FEBS Lett.* **180**, 29–32
- Sharief, F. S. & Li, S. S.-L. (1982) *J. Biol. Chem.* **257**, 14753–14759
- Odani, S., Koide, T., Ono, T. & Ohnishi, K. (1983) *Biochem. J.* **213**, 543–545
- Thompson, R. D., Bartels, D. & Harberd, N. P. (1985) *Nucleic Acids Res.* **13**, 6833–6846
- Mahoney, W. C., Hermodson, M. A., Jones, B., Powers, D. D., Corfman, R. S. & Reeck, G. R. (1984) *J. Biol. Chem.* **259**, 8412–8416
- Odani, S., Koide, T., Ono, T., Seto, Y. & Tanaka, T. (1987) *Eur. J. Biochem.* **162**, 485–491