Protein Science (1993), *2*, 366-382. Cambridge University Press. Printed in the USA.

Relationship between sequence conservation and three-dimensional structure in a large family of esterases, lipases, and related proteins

MIROSLAW CYGLER,[1] JOSEPH D. SCHRAG,[1] JOEL L. SUSSMAN,[2] MICHAL HAREL,[2] ISRAEL SILMAN,[3] MARY K. GENTRY,[4] AND BHUPENDRA P. DOCTOR[4]

[1] Biotechnology Research Institute, National Research Council of Canada, Montréal, Québec H4P 2R2, Canada
[2] Department of Structural Biology and [3] Department of Neurobiology, The Weizmann Institute of Science, Rehovot 76100, Israel
[4] Division of Biochemistry, Walter Reed Army Institute of Research, Washington, D.C. 20307-5100

(RECEIVED August 13, 1992; REVISED MANUSCRIPT RECEIVED October 30, 1992)

Abstract

Based on the recently determined X-ray structures of *Torpedo californica* acetylcholinesterase and *Geotrichum candidum* lipase and on their three-dimensional superposition, an improved alignment of a collection of 32 related amino acid sequences of other esterases, lipases, and related proteins was obtained. On the basis of this alignment, 24 residues are found to be invariant in 29 sequences of hydrolytic enzymes, and an additional 49 are well conserved. The conservation in the three remaining sequences is somewhat lower. The conserved residues include the active site, disulfide bridges, salt bridges, and residues in the core of the proteins. Most invariant residues are located at the edges of secondary structural elements. A clear structural basis for the preservation of many of these residues can be determined from comparison of the two X-ray structures.

Keywords: acetylcholinesterase; conserved residues; esterases; lipases; sequence alignment; three-dimensional structure

Acetylcholinesterase (acetylcholine acyl hydrolase, EC 3.1.1.7) and neutral lipases (triacylglycerol acyl hydrolase, EC 3.1.1.3) act on two very different classes of substrates. The former rapidly hydrolyzes the neurotransmitter acetylcholine and thus plays an important role in synaptic transmission at cholinergic synapses (Quinn, 1987). Its substrate is a small water-soluble molecule, and catalysis takes place in a homogeneous aqueous phase. Lipases preferentially hydrolyze triacylglycerols. They act in situ at the lipid-water interface, and their principal biological role is the breakdown of lipids as an initial event in the utilization of fat as an energy source (Borgström & Brockman, 1984). They form a rather diverse group of enzymes that can be divided into several classes based on amino acid sequence homology (Cygler et al., 1992).

Although these enzymes hydrolyze vastly different substrates, substantial sequence homology was recognized between well-studied members of the acetylcholinesterase family, such as the enzyme from the electric organ of *Torpedo californica* (Schumacher et al., 1986) and *Torpedo marmorata* (Sikorav et al., 1987), and a class of lipases represented by the enzyme from the fungus *Geotrichum candidum* (Shimada et al., 1989; Slabas et al., 1990; Schrag et al., 1991). These proteins contain approximately 550 amino acid residues. A high degree of similarity was detected predominantly in the N-terminal half of the molecule (ca. 60% of the total sequence), especially in the region encompassing the active-site serine. Similarity was less evident in the C-terminal half of the molecule. These regions were thought to be structurally different due to the involvement of the C-terminus of *T. californica* acetylcholinesterase (TcAChE) in dimer formation via an intermolecular disulfide bridge and in attachment of the protein to the membrane via a phosphatidylinositol anchor (Silman & Futerman, 1987), whereas the *G. candidum* lipase (GCL) is a monomeric protein. Recently determined three-dimensional (3D) structures of GCL to 1.8 Å resolution (Schrag et al., 1991; Schrag & Cygler, 1993) and of TcAChE to 2.8 Å resolution (Kinemage 1; Sussman et al., 1991) have, however, revealed a surprising de-

Reprint requests to: Miroslaw Cygler, Biotechnology Research Institute, National Research Council of Canada, Montréal, Québec H4P 2R2, Canada.

gree of structural similarity that extends through the whole length of the polypeptide chain. The root-mean-square deviation between the 399 corresponding C_α atoms after superposition of the two molecules was 1.90 Å (Ollis et al., 1992). The same topological fold, named the α/β hydrolase fold, has been identified in a number of other hydrolases with no sequence similarity to either GCL or TcAChE or to each other (Ollis et al., 1992).

A number of other proteins have been identified that are homologous to TcAChE and GCL, and some of the sequence alignments have already been reported. The most comprehensive studies were carried out by Gentry and Doctor (1991) and Krejci et al. (1991) and comprised 16 and 17 different sequences, respectively. These sequences included, in addition to several vertebrate acetylcholinesterases (AChE), members of the closely related butyrylcholinesterase (BChE) family, as well as insect AChEs, which display properties intermediate between vertebrate AChE and BChE (Gnagey et al., 1987). Other sequences represented various other esterases such as cholesterol and carboxyl esterases and three proteins devoid of known catalytic function: thyroglobulin (Mercken et al., 1985), the protein that is the precursor of the thyroid hormone, and two *Drosophila* adhesion proteins, glutactin (Olson et al., 1990) and neurotactin (de la Escalera et al., 1990). For the discussion that follows, we will refer to this large class of proteins as the lipase/esterase family. It should be stressed, however, that not all lipases nor esterases belong to this family.

In this study we present an improved alignment of sequences that show homology to TcAChE and to GCL, based on the 3D superposition of these two enzymes and inclusion of additional sequences that have recently been reported in the literature. We have identified a total of 32 homologous proteins. Of these, 29 are either lipases or esterases, whereas three others, mentioned above, do not possess a catalytic triad and play other biological roles. Although all the enzymes catalyze hydrolysis of an ester bond, they hydrolyze substrates varying widely in size and complexity. The majority of the enzymes contain a Ser-His-Glu catalytic triad, but some (e.g., cholesterol esterases) have Asp as the acidic member of the triad. Analysis of the conserved positions, identified by the variability index and knowledge of the 3D structures of GCL and TcAChE, provided a basis for understanding their conservation. Furthermore, inspection of the nonconserved regions of GCL and TcAChE provides some information about the parts involved in substrate binding in this array of enzymes, which differ greatly in substrate specificities. Information obtained through such comparisons is in many ways complementary to mutagenesis studies; the sequences represent a natural pool of "mutant" proteins, which preserve catalytic function.

To gain a better understanding of the relationship between sequence and structure and the role of the highly conserved residues in the preservation of the fold, analyses combining sequence alignment with a structural superposition of a few members of the family of proteins have been carried out for other homologous families, notably globins (Lesk & Chothia, 1980), serine proteases (Greer, 1990), and, very recently, subtilisin-like proteins (Siezen et al., 1991). Greer (1990) coined the term "structurally conserved regions," or SCRs, to describe common regions of proteins belonging to the same family, based on the superposition of their 3D structures. Similarly, he used the term "variable regions," or VRs, to describe those regions where the structures differ significantly. This description was adopted by Siezen et al. (1991) in their analysis of the subtilisin-like family, where quantitative analysis was warranted by the availability of the 3D structure of four members of that family. Because for this study there were only two structures available, we have resorted to a qualitative description of the scaffold SCRs.

Results and discussion

The proteins identified in this study are quite diverse in terms of their substrates and their biological roles. They are relatively large (~60 kDa, >550 amino acids) compared to most serine proteases and some other lipases (fungal *Rhizomucor miehei* lipase, 269 amino acids [Boel et al., 1988]; the catalytic domain of human pancreatic lipase contains 335 amino acids [Winkler et al., 1990]). This invites speculation that, in addition to ester bond hydrolysis, they may also perform other functions and/or be allosterically controlled. In the cases of glutactin, neurotactin, and thyroglobulin, the region homologous to esterases forms only one segment of these proteins, and its function is not related to catalytic activity.

The topology diagram representing the 3D structures of TcAChE and GCL is shown in Figure 1, and secondary structure assignments are listed in Table 1. In the following text, when a specific position is mentioned, two sequence numbers are given: the first corresponds to the GCL numbering and the second to the TcAChE numbering, e.g., Ser [217,200]. The convention used to refer to a particular topological feature is described in the legend to Figure 1. The proteins (Table 2) have been selected and their sequences aligned as described in the Methods. The degree of conservation at each position along the aligned protein sequence was analyzed with the use of the variability index (Kabat et al., 1983). The results are shown in graphic form in Figure 2. The variability plot indicates clearly that homology between these proteins is more pronounced in their N-terminal parts than in C-terminal parts, as noted previously for the comparison of TcAChE and GCL (Slabas et al., 1990; Schrag et al., 1991). A close inspection of the backbone traces of GCL and TcAChE indicates that, in fact, the structures can be divided into N- and C-terminal segments (subdomains) near position [350,323], adjacent to the loop containing the active-site glutamate (Fig. 1).

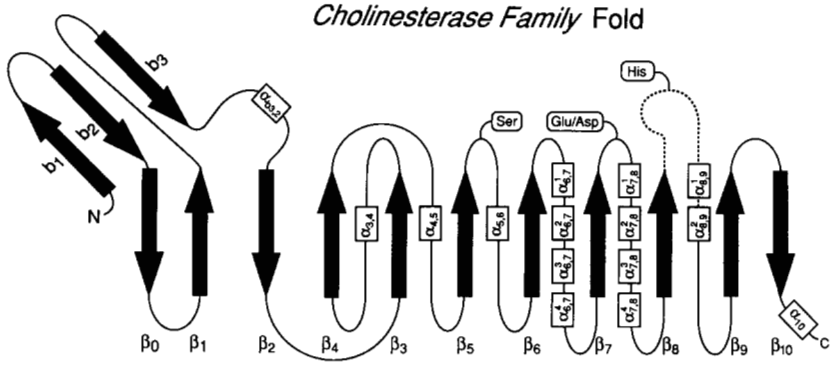

Fig. 1. Topology diagram of the common elements of TcAChE and GCL structures. The strands of the small, three-stranded β-sheet are marked by b_j, where j ranges from 1 to 3. Strands of the large sheets are marked by Greek letter β with a subscript from 0 to 10. Numbering starts at 0 to remain consistent with the nomenclature of the α/β hydrolase fold motif (Ollis et al.,1992). The connection (loop) between strands i and j of the β-sheet is marked as $l_{i,j}$ or $L_{i,j}$ for small or large sheet, respectively. The α-helix is referenced as $\alpha^k_{i,j}$, where subscripts refer to the loop in which it is embedded, and superscript k refers to the sequential number of this helix within the loop. If there is only one helix in the connecting loop the superscript is not used. In some cases the symbol $L^k_{i,j}$ is used, in which case the superscript refers to the sequential order of a part of the $L_{i,j}$ loop contained between two secondary structural elements.

Table 1. *Secondary structure assignments for acetylcholinesterase (AChE) and Geotrichum candidum lipase (GCL)*[a]

AChE	GCL	Type	Name
6–10	4–7	β-Strand	b_1
13–16	11–14	β-Strand	b_2
18–21	16–18	β-Strand	β_0
26–34	21–28	β-Strand	β_1
57–60	51–54	β-Strand	b_3
–	66–77[b]	α-Helix	$\alpha^1_{b3,2}$
79–85[b]	–	α-Helix	$\alpha_{b3,2}$
–	85–95	α-Helix	$\alpha^2_{b3,2}$
96–102	107–113	β-Strand	β_2
109–116	123–128	β-Strand	β_3
132–139	144–152	α-Helix	$\alpha_{3,4}$
142–147	158–162	β-Strand	β_4
–	175–180[b]	α-Helix	$\alpha^1_{4,5}$
168–183	185–200	α-Helix	$\alpha^2_{4,5}$
193–199	210–216	β-Strand	β_5
200–211	218–227	α-Helix	$\alpha_{5,6}$
220–226	244–248	β-Strand	β_6
238–252	266–274	α-Helix	$\alpha^1_{6,7}$
259–268	282–291	α-Helix	$\alpha^2_{6,7}$
271–278	294–308	α-Helix	$\alpha^3_{6,7}$
305–311	332–337	α-Helix	$\alpha^4_{6,7}$
318–324	345–351	β-Strand	β_7
329–335	356–360	α-Helix	$\alpha^1_{7,8}$
349–360	368–378	α-Helix	$\alpha^2_{7,8}$
365–376	384–393	α-Helix	$\alpha^3_{7,8}$
384–411	416–438	α-Helix	$\alpha^4_{7,8}$
417–423	443–449	β-Strand	β_8
443–448	467–471	α-Helix	$\alpha^1_{8,9}$
460–479	478–491	α-Helix	$\alpha^2_{8,9}$
502–505	513–516	β-Strand	β_9
510–514	521–525	β-Strand	β_{10}
518–534	530–538	α-Helix	α_{10}

[a] See Figure 1 for the diagram of the consensus secondary structure diagram of the two structures.
[b] The corresponding helix is absent in the other enzyme.

A pairwise comparison of all sequences, based on the common alignment of Figure 2, is shown in Figure 3A. A similar comparison for the N- and C-terminal subdomains is shown separately in Figure 3B,C. Subfamilies are clearly distinguishable in this representation. The percentage of identical amino acids at equivalent positions for each pair of proteins in this set varies from 16% for the most distantly related to 97% for the most closely related sequences. The average pairwise identity for all 32 proteins is ~29%. The identity figure between N-terminal parts ranges between 22% and 97% with an average of 36%, while that for the C-terminal parts varies between 6% and 98%, with an average of 20%. These numbers represent a comparison based on a multiple-sequence alignment, with restrictions on the positions of insertions imposed by the two known 3D structures, rather than on an alignment optimized for each pair independently, as is done customarily. Therefore, the percentages, especially for C-terminal segments, are in some cases lower than numbers obtained for two unrelated sequences (~15–20% [Doolittle, 1985]). Although the amino acid similarity of the C-terminal parts is weak, there are several positions that are very well conserved.

For three proteins in this set, bovine thyroglobulin (Mercken et al., 1985), *Drosophila* glutactin (Olson et al., 1990), and *Drosophila* neurotactin (de la Escalera et al., 1990), the catalytic triad is not conserved. Thyroglobulin, a 2,750-amino acid residue protein, is a precursor of thyroid hormone. Neurotactin is a large, apparently transmembrane protein, involved in cell–cell adhesion. In both of these cases, it is the C-terminal domain that shows homology to the esterases. In glutactin, a 1,023-amino acid residue glycoprotein located in basement membranes of *Drosophila*, it is the N-terminal segment that is homologous to esterases. Although the sequence identity of these three nonhydrolytic proteins with other proteins in this

Table 2. *Protein sequences included in the comparison*

Protein	Source	Code	Reference
Lipases	*Geotrichum candidum* gene 1	Gclip1	Shimada et al., 1990
	G. candidum gene 2	Gclip2	Shimada et al., 1989
	Candida cylindracea gene 1	Crlip1	Kawaguchi et al., 1989
	C. cylindracea gene 2	Crlip2	Longhi et al., 1992
Acetylcholinesterases	*Torpedo californica*	TcalifAChE	Schumacher et al., 1986
	Torpedo marmorata	TmarmAChE	Sikorav et al., 1987
	Mouse	MouseAChE	Rachinsky et al., 1990
	Fetal bovine serum	FbsAChE	Doctor et al., 1990
	Human	HumanAChE	Soreq et al., 1990
	Drosophila	DrosAChE	Hall and Spierer, 1986
	Anopheles stephensi	AnophAChE	Hall and Malcolm, 1991
Butyrylcholinesterases	Human	HumanBChE	Lockridge et al., 1987
	Mouse	MouseBChE	Rachinsky et al., 1990
	Rabbit	RabbitBChE	Jbilo and Chatonnet, 1990
Carboxylesterases	Rat liver, pI 6.1	Rat61	Robbi et al., 1990
	Human	HumanCE	Munger et al., 1992; Long et al., 1991
	Rabbit liver esterase 1	RabbitCE1	Korza and Ozols, 1988
	Rabbit liver esterase 2	RabbitCE2	Ozols, 1989
	Mouse isoenzyme	MouseCE	Ovnic et al., 1991
	Rat liver esterase E1	RatCE	Long et al., 1988 Takagi et al., 1988
	Drosophila esterase 6	DrosCE6	Oakeshott et al., 1987
	Drosophila esterase P	DrosCEP	Collet et al., 1990
	Heliothis juvenile hormone	HeliotCE	Hanzlik et al., 1989
	Culex esterase B1	CulexCEb1	Mouches et al., 1990
Cholesterol esterases	Rat pancreas	RatCholE	Han et al., 1987; Kissel et al., 1989
	Human	HumanCholE	Hui and Kissel, 1990; Nilsson et al., 1990; Baba et al., 1991
	Bovine	BovCholE	Kyger et al., 1989
Other esterases	*Dictyostelium* D2 esterase	D2	Rubino et al., 1989
	Dictyostelium crystal protein	Dcp	Bomblies et al., 1990
Nonhydrolytic proteins	*Drosophila* neurotactin	Neurotactin	de la Escalera et al., 1990
	Drosophila glutactin	Glutactin	Olson et al., 1990
	Bovine thyroglobulin	Bovthyro	Mercken et al., 1985

family is within 16–28%, there is no evolutionary pressure to maintain the geometry of the active site. One may, therefore, predict that their 3D structures will show more divergence from GCL and TcAChE than other proteins in this group. Consequently, the analysis described below is based primarily on the sequences of 29 esterases and lipases, and the three noncatalytic proteins are mentioned only where appropriate.

Data presented in Figure 3 indicate that proteins with similar substrate specificity share a higher degree of similarity. For example similarity between mammalian AChEs and BChEs (as measured by the percentage of identical amino acids) ranges from ~50% to 93%. Insect AChEs from *Drosophila* and *Anopheles stephensi* show a lower level of identity with their mammalian counterparts, only 35–38%, still well above the average level observed for the whole group. There are two regions where large insertions are observed. One insertion occurs in the $L_{b3,2}$ loop in lipases. This loop blocks access to the active site in GCL, and most likely plays a similar role in other lipases of this family. The second insertion occurs near position [117,106] of insect AChEs. This additional sequence, as deduced from the cDNA sequence, is highly hydrophilic and the mature protein is proteolytically cleaved in this region (Fournier et al., 1988).

A few general conclusions can be drawn from inspection of the variability plots (Fig.1). There are 24 positions (~4.5%) that are absolutely conserved in all 29 enzyme sequences ($V = 1$), 21 of which are in the N-terminal segment (as defined above). Of these, 10 are glycines that

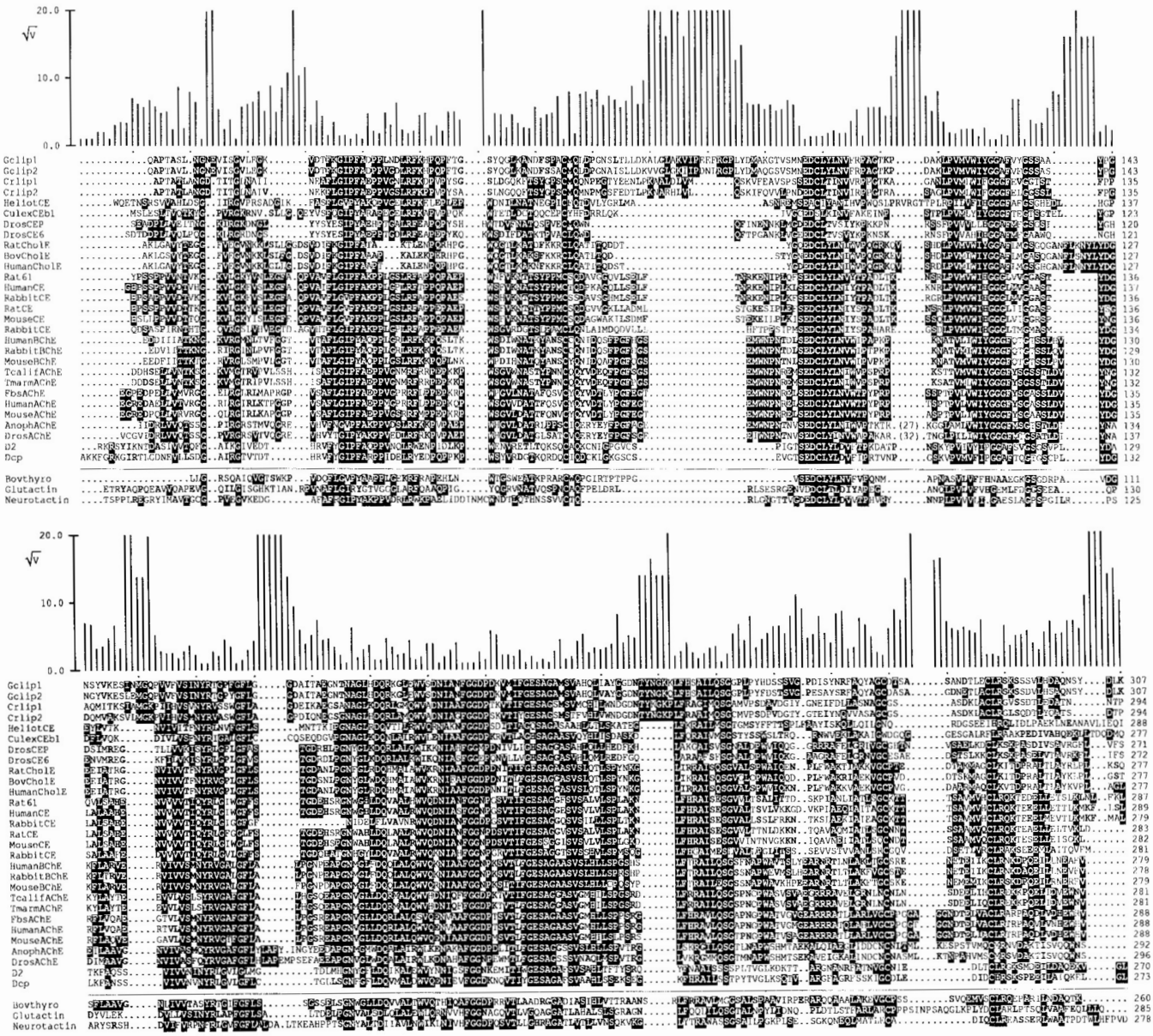

Fig. 2. Amino acid sequence alignment of 32 homologous proteins. Solid line divides hydrolytic enzymes from three other homologous proteins. The most common amino acid type at each position is shown in reversed video mode. Second most abundant amino acid type is shown in bold letters. Bar plot above the aligned sequences represents the square root of the variability index for each position, based on the first 29 sequences. The scale on the vertical axis is from 0 to 20. Values above 20 were truncated to 20 for better readability. (*Figure continues on next page.*)

either have main-chain torsion angles in the range not easily accessible for other residues, or form close contacts through their C_α atoms with other atoms. Eleven of the 24 positions are conserved in thyroglobulin, neurotactin, and glutactin.

In addition to positions having $V = 1$, there are 49 positions with $1 < V < 4$ (42 in the N-terminal part) and another 71 with $4 < V < 9$ (53 in the N-terminal part). In total, 144 positions (~27%) have a variability index less than 9, of which 116 (~33%) are in the N-terminal part and only 28 (~15%) in the C-terminal part. Of the 120 low variability positions ($1 < V < 9$) there are 19 (14 in the N-terminus), where the substitutions are conservative, i.e., of the type Gly/Ala, Val/Ile/Leu, Phe/Tyr/Trp, Ser/Thr/Cys, Asp/Glu, Asn/Gln, or Lys/Arg. Most fall into the first three categories, being either aliphatic (hydrophobic) or aromatic. Low variability positions are not distributed evenly along the sequences but are clustered in specific areas (Table 3).

Analysis of the role of amino acid positions with low variability is of great importance for understanding common characteristics of the proteins under consideration

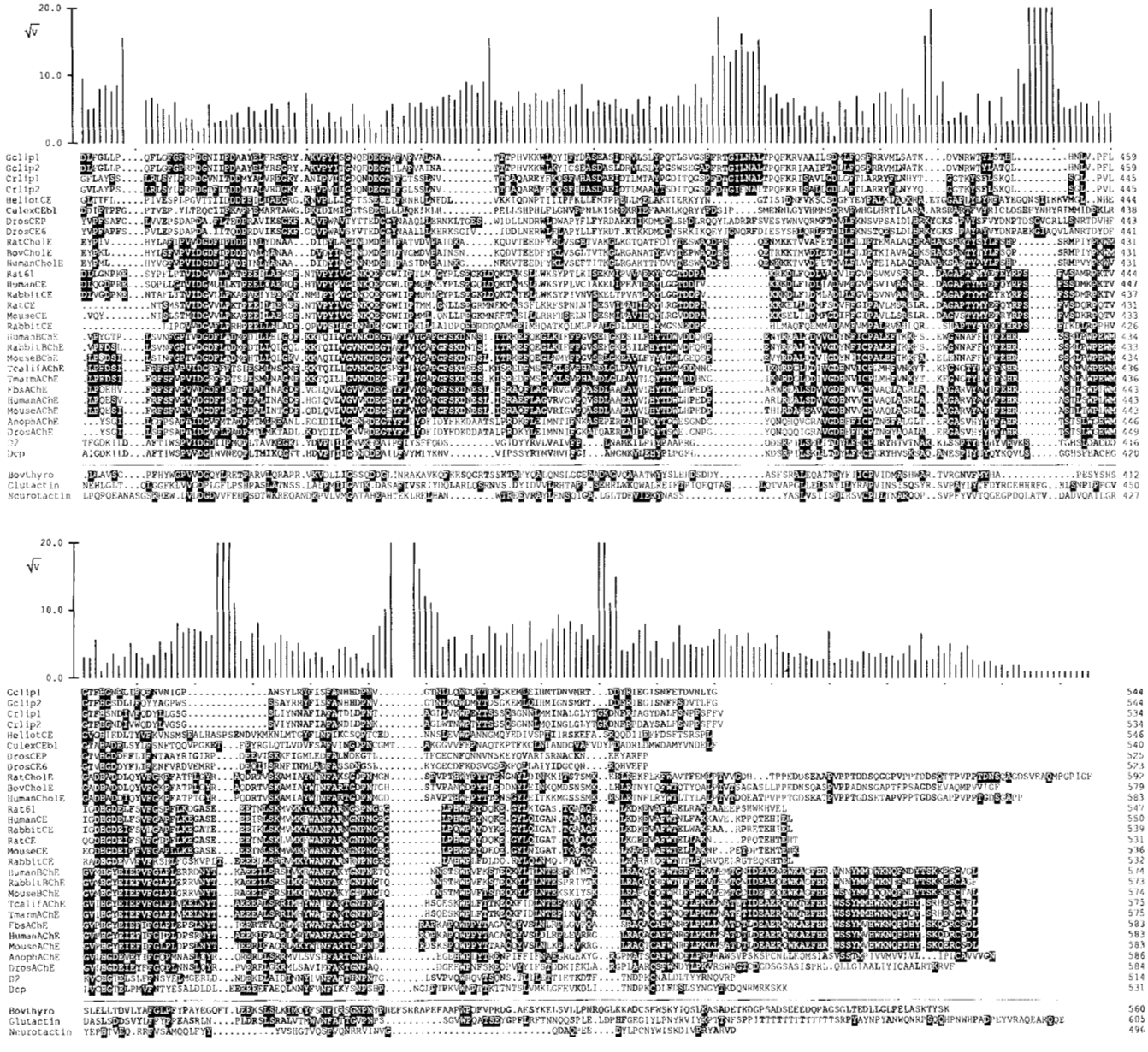

Fig. 2. *Continued.*

in terms of fold and similarity of catalytic mechanisms. Stereodiagrams of the C_α tracings of GCL and TcAChE with these positions color-coded are shown in Figure 4. It is immediately apparent that the invariant residues are located at the edges of the large β-sheet, mainly at the C-terminus of this parallel sheet and in connecting loops. They seem to be placed in key positions to ensure correct folding. The majority of low variability positions are in the core of the protein, within the β-sheet structure. In fact, most are within strands β_1 to β_7 of the large β-sheet or in the small β-sheet (see Fig. 4). Residues forming α-helices are, in general, less well conserved, with the exception of the long helix $\alpha^2_{4,5}$. The latter has an amphipathic character, with one side exposed to the surface and the other packed against the β-sheet. Residues facing the surface show high variability, while those facing the interior of the protein are much more conserved (Fig. 5). Helices identified by Ollis et al. (1992) as part of the α/β hydrolase fold, which pack tightly against the β-sheet, show somewhat lower variability than other helices.

The low variability positions ($V < 9$) can be divided into five groups: (1) catalytic triad, (2) hydrophobic core-forming residues, (3) residues involved in salt bridges, (4) cysteine residues forming disulfide bridges, and (5) residues at the edges of the secondary structural elements (turns and loops).

Table 3. *Low and high variability clusters*

Residue range	Topological position
Low variability regions	
[24,30]-[63,68]	Loop after strand β_1
[102,91]-[113,102]	Strand β_2
[123,110]-[137,124]	Strand β_3 and loop $L^1_{3,4}$
[157,141]-[171,155]	Strand β_4 and loop $L^1_{4,5}$
[183,166]-[228,211]	Helix $\alpha^2_{4,5}$, strand β_5, helix $\alpha_{5,6}$ area around active-site Ser
[241,218]-[250,227]	Strand β_6
[345,318]-[358,331]	Strand β_7
[425,397]-[431,403]	C-terminal half of helix $\alpha^3_{7,8}$
[460,437]-[471,448]	Region around active-site His
[487,475]-[495,483]	Helix $\alpha^2_{8,9}$
High variability regions	
[17,19]-[22,28]	Loop $L_{0,1}$
[60,65]-[92,81]	Loop $L_{1,2}$
[117,106]-[121,108]	Loop $L_{2,3}$
[261,238]-[273,251]	Helix $\alpha^1_{6,7}$
[302,279]-[315,288]	Helix $\alpha^3_{6,7}$
[365,346]-[412,384]	Helices $\alpha^1_{7,8}$ and $\alpha^2_{7,8}$
[433,405]-[456,431]	Strand β_8
[476,454]-[480,468]	Loop between $\alpha^1_{8,9}$ and $\alpha^2_{8,9}$
[496,484]-the end	Strand β_9 to the end

Catalytic triad

Although some of the catalytic residues have been identified previously, only determination of the 3D structures of GCL and TcAChE clearly identified the Ser-His-Glu serine protease-like catalytic triad (Kinemage 2). These residues are conserved with the exception of cholesterol esterases and *Drosophila* carboxylesterases, where Glu is replaced by Asp. The critical role of these residues in catalysis has been confirmed for various enzymes of this family by site-directed mutagenesis. In TcAChE, the mutation of Ser 200 to Cys resulted in diminished activity, while mutation to Val abolished all detectable activity (Gibney et al., 1990). Mutation of the corresponding Ser 217 in GCL to Ala also rendered this enzyme inactive (T. Vernet, pers. comm.). Replacement of His 440 of TcAChE by Glu eliminated activity, whereas the mutation of His 425 reduced activity only slightly (Gibney et al., 1990). Location within the protein sequence of the active-site Ser and His have also been confirmed by site-directed mutagenesis in rat cholesterol esterase (DiPersio et al., 1990, 1991).

The active-site Ser [217,200] is part of a conserved sequence: Gly-Glu(His)-*Ser*-Ala-Gly-Ala/Gly. This sequence, and in particular Gly-Xaa-*Ser*-Xaa-Gly, has been found in many other enzymes containing a catalytic triad (Brenner, 1988). A recent comparison of the structures of five enzymes displaying the α/β hydrolase fold (Ollis et al., 1992) showed that, in all of them, the serine (or the residue with an analogous role) is embedded in a tight turn between a β-strand and an α-helix. This serine is in a strained conformation with backbone torsion angles of 53°,−118° in GCL and 68°,−100° in TcAChE. As a result, the serine hydroxyl group is well exposed and easily accessible to the catalytic histidine and to the substrate. Requirements for glycines 2 residues before and 2 residues after the serine (Gly [215,198] and [219,202]) are due to the close proximity of these two residues in the strand-turn-helix supersecondary structure. Requirement for a small side chain at position [220,203] is also due to steric restrictions. A similar supersecondary structure around the active-site serine has also been found in the structures of two other lipases (Brady et al., 1990; Winkler et al., 1990) and has recently been postulated to occur in some other lipases and esterases (Derewenda & Derewenda, 1991).

The observation that residue [216,199], with its side chain below the catalytic triad, is almost always a glutamate in the lipase/esterase family of proteins, suggests that it may be of importance for catalysis. There is a second acidic side chain in the vicinity (in the interior of the protein), which is also well conserved (Asp/Glu [466,443]). No defined role for these residues has been elucidated so far, but it has been suggested (Schrag et al., 1991; Sussman et al., 1991) that their role is to coordinate water molecules needed for substrate hydrolysis. When Glu [216,199] was changed to Gln in TcAChE, the rate of hydrolysis of acetylcholine was reduced approximately 10-fold (Gibney et al., 1990). However, K_{cat}/K_m was altered approximately 50-fold (P. Taylor, pers. comm.). In a few cases residue [216,199] is in fact Gln (*Heliothis* juvenile hormone esterase) or His (*Drosophila* esterases). Mutation of the tripeptide containing the other acidic residue, Glu [466,443]-Ile-Glu to Gly-Ile-Gln in human BChE reduced the enzymatic activity without affecting its folding (Neville et al., 1992).

The active-site His [463,440], contained in a consensus sequence Gly-Sm-Xaa-*His*-Sm-Xaa-Glu/Asp (Sm, residue with a small side chain), is embedded in a type II β-turn. His is found in the second position in this turn, while the third position is usually occupied by glycine (Gly [464,441], ϕ,ψ of 88°,−13° in GCL; 114°, −1° in TcAChE). This turn is preceded by a type I β-turn in which the second position is often occupied by glycine (Gly [460,437], ϕ,ψ of 62°,−148° in GCL, 68°,−130° in TcAChE). Carboxylesterase (CE) sequences depart somewhat from this pattern and are characterized by the sequence Gly-Asp-*His*-Gly-Asp-Glu, with the first Gly being 1 residue closer to histidine than in other proteins. CE may have a different local structure in the loop leading to histidine. This subfamily also shows differences in the sequence around the active-site acid (see below). Yet another pattern is observed in *Dictyostelium* esterases, where His is preceded by Cys.

The position of the acid member of the active-site triad is very well maintained in the sequence alignment. In most proteins of this family the role of the acid is provided by

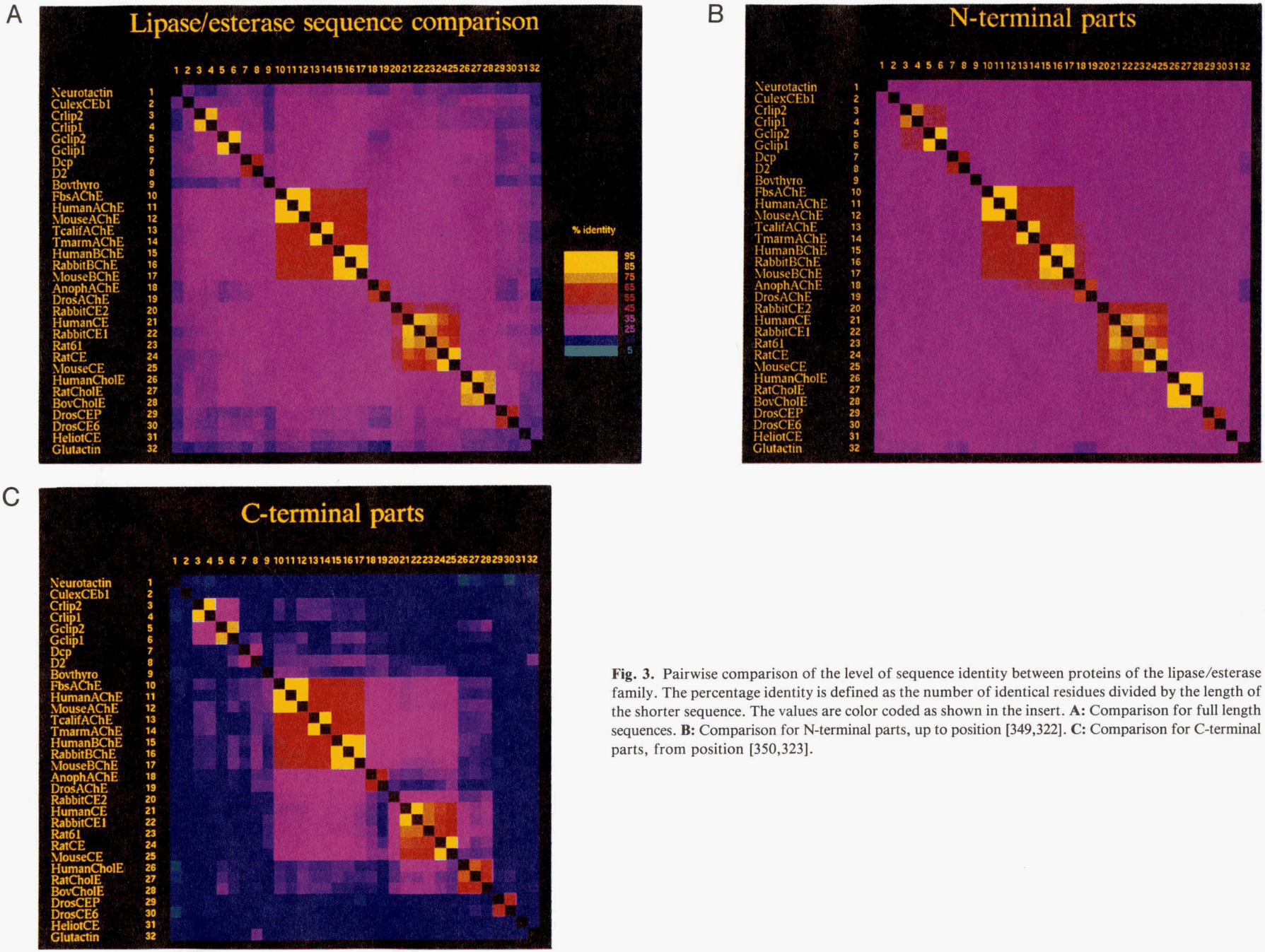

Fig. 3. Pairwise comparison of the level of sequence identity between proteins of the lipase/esterase family. The percentage identity is defined as the number of identical residues divided by the length of the shorter sequence. The values are color coded as shown in the insert. **A:** Comparison for full length sequences. **B:** Comparison for N-terminal parts, up to position [349,322]. **C:** Comparison for C-terminal parts, from position [350,323].

glutamic acid. This is the first family of enzymes containing a catalytic triad, which has Glu in this role. It is important to note, however, that aspartic acid seems to fulfill this role in cholesterol esterases and *Drosophila* esterases 6 and P. The acid takes the second position in a type I turn. The consensus sequence around the acid is Gly-(Xaa)$_4$-*Glu/Asp*-Gly. The first of the two Gly residues is at the end of strand β_7 and assumes an extended conformation. It is occluded by the aromatic ring of a highly conserved residue, Tyr [447,421], embedded in the middle of strand β_8. There is no space for a side chain at this position without some changes in the local structure. The second Gly, which follows the acidic residue, is in the third position of the type I β-turn and packs against the backbone of helix $\alpha^4_{7,8}$ near the conserved Asp [425,397].

The position of the catalytic triad acidic residue in carboxylesterases cannot be predicted unambiguously from the present data. There are two adjacent sequence regions in CEs that align well with the active site acid. The first has the consensus sequence Gly 332-(Xaa)$_4$-*Glu*-Phe/Tyr-Gly (human CE numbers) and has an aromatic side chain, rather than Gly, after the acid (position [355,328]). The glycine is moved one position back, where other sequences have a residue with a small side chain. The second region has the consensus sequence Gly 348-(Xaa)$_4$-*Glu*-Gly (human CE numbers), conforming to the general pattern. The high degree of similarity to other proteins in the [320,293]–[340,313] region, preceding the two aforementioned sequences, suggests that structures in that region are very similar to TcAChE and GCL. If the active-site acid comes from the first of the two possible sequences, the conformation of the backbone around it would have to be somewhat different than the one observed in GCL and TcAChE (see above, the histidine loop conformation) due to the presence of Phe/Tyr side chain in place of Gly [355,328]. If the active-site acid comes from the second sequence, the loop between helix $\alpha^4_{6,7}$ and strand β_7 would have to be longer, whereas loop $L^1_{8,9}$, following the acid, would have to be shorter than in the other enzymes. The alignment shown in Figure 2 presents the first of these two alternatives. Future site-directed mutagenesis experiments may settle this question.

Glu [354,327] is guided into the correct orientation for hydrogen bonding to His [463,440] by two additional hydrogen bonds to the other oxygen atom of its carboxylic group: one from the backbone NH of residue [351,324], three positions before the acidic residue in the sequence, and one from the hydroxyl group of Ser [249,226]. In other enzymes of the α/β hydrolase fold family the corresponding Asp is stabilized through a hydrogen bond from the backbone NH two residues after the acid (Ollis et al., 1992). Ser [249,226] is conserved not only in Glu-containing enzymes but also in those containing Asp in the active site. This suggests that the Ser is also used for hydrogen bonding to Asp, following some small movement of the backbone.

Hydrophobic core

As noted previously for globins (Lesk & Chothia, 1980), the majority of residues with low variability are located in the hydrophobic core of the protein. They tend to cluster in a few regions. One is within strands β_1 to β_7 of the large β-sheet, and, in fact, very few residues of these seven strands have a variability index greater than 9. There is a high concentration of branched aliphatic residues in this cluster in accordance with known preferences and requirements for the packing of β-strands (Levitt, 1978). Well-conserved aromatic residues also tend to cluster together. Their aromatic side chains form a layer above strands β_3 to β_8 on the concave side of the β-sheet.

Although both sides of the large β-sheet have strongly hydrophobic character and are shielded from solvent by numerous α-helices, aliphatic and aromatic residues within the sheet show preferences in their locations. We observe that nearly all branched side chains are found on the convex side of the β-sheet, whereas the aromatic residues strongly favor the concave side of the sheet. Only 1 out of the 7 highly conserved aromatic residues is on the convex side.

Salt bridges

The 3D structures of GCL and TcAChE contain four conserved salt bridges. There are usually two hydrogen bonds between Arg/Lys and Asp/Glu and additional bonds to the protein backbone. These bridges play an important role in stabilizing the 3D fold of the protein by tying neighboring loops together.

The first salt bridge is formed between Arg [38,44] and Glu [103,92]. Glu [103,92] is at the beginning of strand β_2 of the large β-sheet and on the C-terminal side of the loop $L_{b3,2}$, which varies greatly in length among the compared proteins. In TcAChE, part of this loop is positioned near the active-site cleft and is likely to be involved in substrate binding. In GCL this loop covers the active site and presumably undergoes some conformational change upon substrate binding (Schrag et al., 1991). The salt bridge helps to keep the bottom of this loop in place and is located close to the disulfide bridge formed by Cys [61,67] and Cys [105,94] (see below). Glu [103,92] is conserved in all sequences except glutactin, where it is replaced by Asp, that may perform a similar function. Mutation of this Glu to either Gln or Leu in human AChE resulted in loss of activity of the expressed protein and is probably due to improper folding (Bucht et al., 1992). Arg [38,44] is located in a segment $L_{1,b3}$ connecting strand β_1 of the large β-sheet and strand β_3 of the small β-sheet. It is conserved in all sequences except those of cholesterol esterases. ChE sequences have a deletion of three to four residues in this region and must have a different conformation of this loop. All ChEs have, however, a conserved

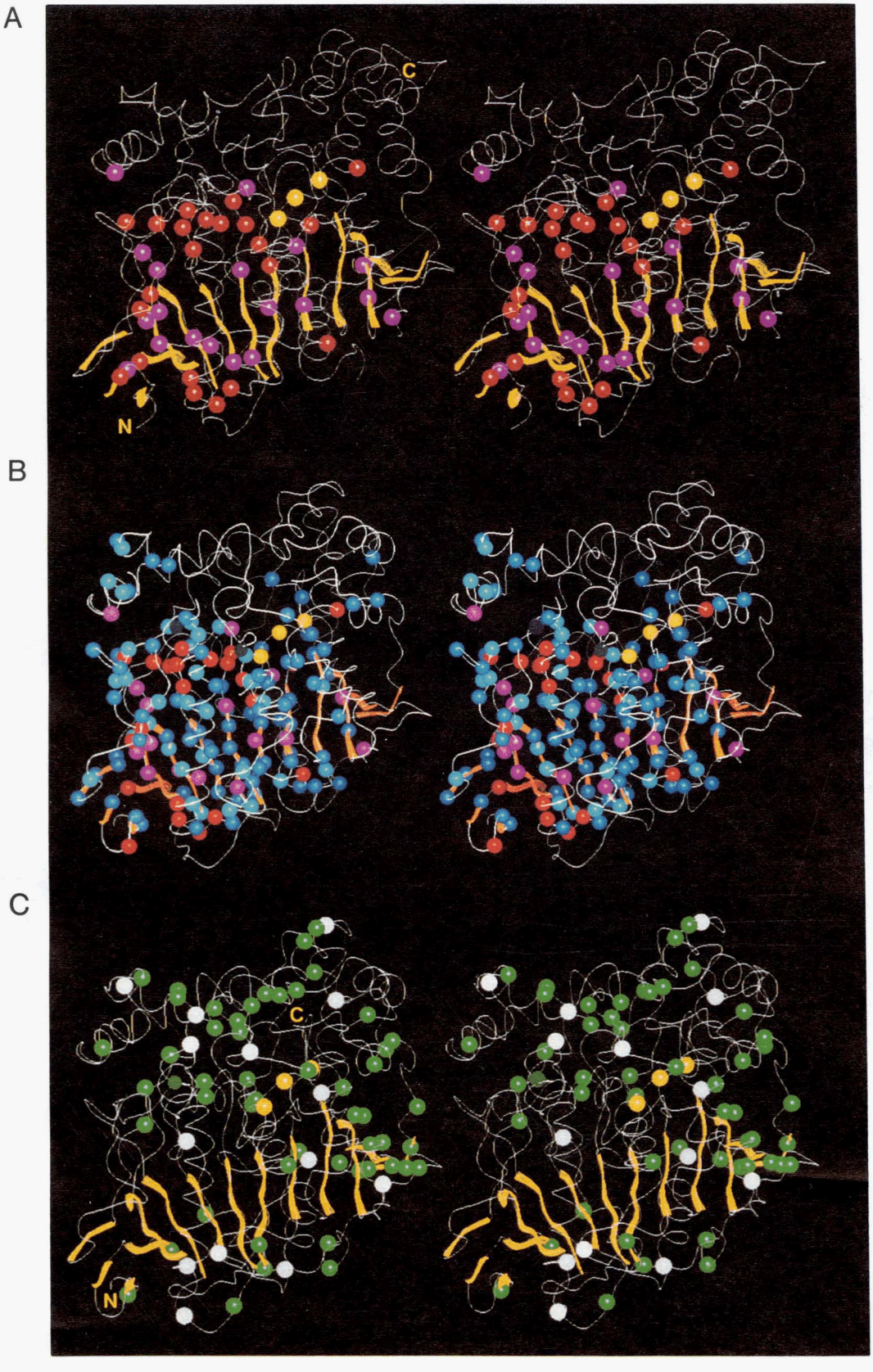

Fig. 4. Ribbon tracing of the molecules showing regions with low and high variability. Yellow ribbons mark the strands of the β-sheets. Comparison of the tracings in A and B gives an idea of the similarity of GCL and TcAChE structures. Red – positions with $V = 1$; magenta – positions conserving the character of the side chain; yellow – active-site triad; blue – positions with $1 < V < 4$; dark blue – positions with $4 < V < 9$; green – positions with $V > 50$; white – positions of insertions or deletions. **A:** TcAChE – positions with $V = 1$ and those where the character of the side chain is conserved. **B:** GCL – positions with $V < 9$. **C:** GCL – positions of deletions and of residues with high variability.

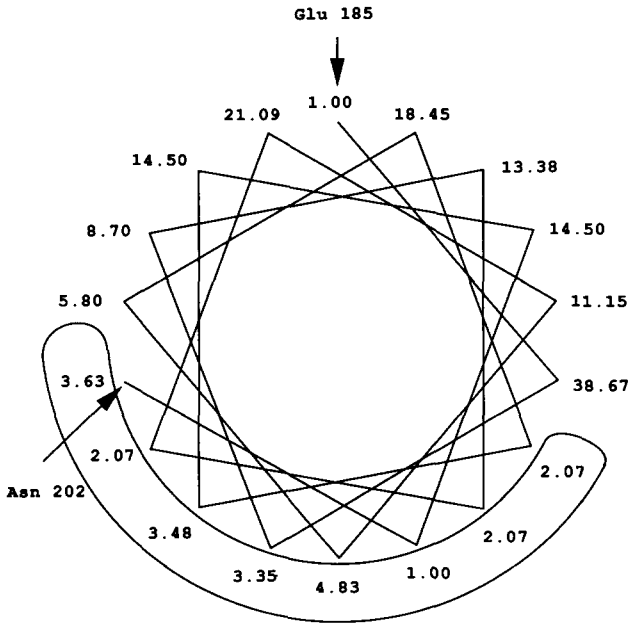

Fig. 5. Helical wheel for helix $\alpha^2_{4,5}$, [185,168]-[200,183] with variability indices, V, at each position.

Lys residue within this stretch and it is likely that this Lys forms analogous hydrogen bonds with Glu [103,92].

The second salt bridge, Arg [165,149]-Asp [189,172], is formed between a highly conserved pair of residues at the bottom of a medium-sized loop, $L^1_{4,5}$, connecting strand β_4 and a long helix, $\alpha_{4,5}$. Arg [165,149] is conserved in all 32 sequences, and Asp [189,172] is conserved in all but one. Rabbit liver esterase 1 (Korza & Ozols, 1988) has a Phe in this position, but there is an Asp-Glu dipeptide 2 residues preceding it. This protein has a deletion of ~9 residues in the surface loop, $L^1_{4,5}$, prior to Glu. This suggests that in rabbit liver esterase 1, the conformation of this loop is somewhat different than its conformation in other proteins and might permit the formation of a salt bridge between Glu [187,170] and Arg [165,149]. The adjacent Tyr [164,148] is also conserved in all sequences. Its side chain not only fits well into the hydrophobic environment, but also makes a hydrogen bond to the main-chain carbonyl group of residue [129,116].

The third salt bridge is between Glu [180,163] and Arg [290,267]. Arg [290,267] is located in a well-conserved short helix, $\alpha^2_{6,7}$, near the tip of a long loop, $L_{6,7}$. It is a highly conserved position, occupied by either Arg or Lys. It is located near Cys [288,265], which is part of the second disulfide bridge. Glu [180,163] is embedded in a surface loop, $L^1_{4,5}$. Although this loop has a different conformation in GCL and TcAChE, this Glu is found in a similar position in both. In GCL, Arg [290,267] forms two hydrogen bonds to the carboxylic group of Glu [180, 163], both side chains being in an extended conformation. In TcAChE, Glu 163 folds back, but there are also two hydrogen bonds from Arg to the Glu, one of them to the backbone carbonyl group (Fig. 6). Although position [180,163] is not as well conserved as position [290,267] and this salt bridge does not exist in all compared proteins, it seems to be important for the lipase and cholinesterase subfamilies. It is also likely to be formed in bovine thyroglobulin and *Drosophila* glutactin.

These three salt bridges are in close spatial proximity to each other. In fact, Arg [38,44] and Arg [290,267] are close neighbors and run antiparallel to one another (Fig. 6). The salt bridges act in a cooperative way to tie together surface loops: the tip of the large $L_{6,7}$ protrusion and the tips of loops $L_{1,b3}$, $L^3_{3,2}$, and $L^1_{4,5}$. In TcAChE, the N-terminal segment of the $L_{6,7}$ protrusion forms part of the "aromatic" gorge, which serves as the entrance to the active site (Sussman et al., 1991). The same segment in GCL is likely to contribute to substrate binding (Schrag et al., 1991). This region may also be involved in substrate recognition and/or binding in other proteins of this family.

The fourth salt bridge is formed between Asp [425,397], located in the middle of a long, kinked helix $\alpha^4_{7,8}$ preceding strand β_8, and Arg/Lys [529,517] at the end of loop L_{10}, just before C-terminal helix α_{10}. Asp is conserved in all sequences. Although it is not in the immediate vicinity of the active site, its site-directed mutagenesis to Asn appears to affect the catalytic activity of the enzyme (Krejci et al., 1991). Additional evidence (Shafferman et al., 1992) suggests that this mutation renders the enzyme inactive by preventing its folding to the native conformation. This salt bridge is well conserved in lipases and in AChEs,

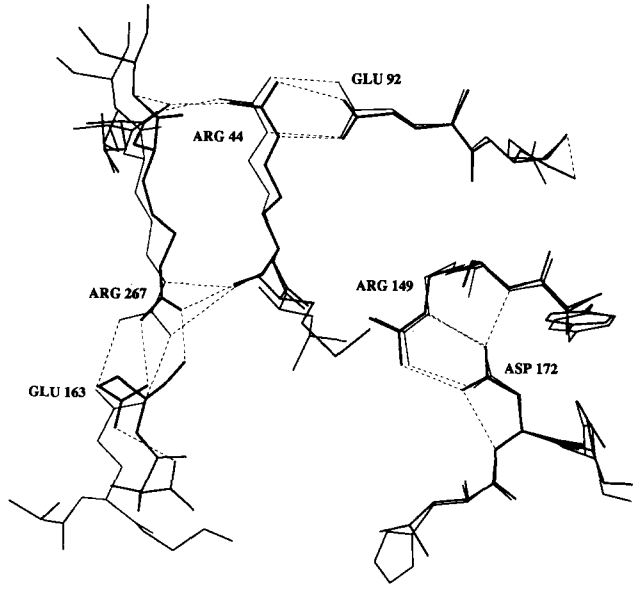

Fig. 6. Salt bridges: Arg [38,44]-Glu [103,92]; Arg [165,149]-Asp 189,172]; Glu [180,163]-Arg [290,267]. Superposition of TcAChE and GCL; TcAChE—thick lines; GCL—thin lines.

BChEs, and CEs. It provides an anchor for the C-terminal helix α_{10}, which might precede, or even guide, the formation of a disulfide bond characteristic for esterases (see below). Although position [529,517] contains a positively charged side chain in only 21 sequences, there are nearby Arg or Lys residues in the other proteins that, judging from the 3D structure of GCL and TcAChE, could participate in a similar interaction.

Disulfide bridges

There are two disulfide bridges that are conserved in nearly all the 32 sequences, and a third one conserved among the AChEs and BChEs, as well as in thyroglobulin and neurotactin.

The first bridge joins Cys [61,67] and Cys [105,94]. It encompasses a variable length loop, $L_{b3,2}$, that is of importance for substrate binding. Position [61,67] is strictly conserved, whereas position [105,94] contains Cys in all but one sequence. The exception is *Culex* esterase B1 (Mouches et al., 1990), where the Cys residue is three positions prior to [105,94]. Because loop $L_{b3,2}$ in this protein is one of the shortest among all sequences, a different conformation at its base that would still permit formation of a disulfide bridge is possible.

The second disulfide bridge, between Cys [276,254] and Cys [288,268], encompasses a small segment at the tip of a much larger surface protrusion (residues [261,242] to [311,283]). The size of the disulfide loop varies from 4 residues in neurotactin to 17 in glutactin. Even within AChEs this loop varies in length by 4 residues. Only two sequences lack these cysteines: *Culex* esterase B1 (Mouches et al., 1990) and *Heliothis* juvenile hormone esterase (Hanzlik et al., 1989). Although this disulfide bridge is highly conserved, its role and importance are not apparent from the 3D structure alone. Because the large surface loop encompassing this disulfide is near the entrance to the active site, this loop may play a role in substrate binding and/or recognition.

The AChEs and BChEs also contain a third intramolecular disulfide bridge linking positions [430,402] and [533,521]. These two Cys residues are also present in the sequences of bovine thyroglobulin and *Drosophila* neurotactin. This covalent linkage plays an important role in TcAChE. The C-terminus of TcAChE provides contacts for dimer formation (via a four-helix bundle and an intermolecular disulfide bridge) and has an attachment site for anchoring the protein in the membrane (Sussman et al., 1991). The C-terminal helix α_{10} of TcAChE points away from the rest of the protein and makes relatively few contacts with it. The disulfide bridge helps to hold helix α_{10} together with the rest of the protein. Although the backbone conformation around this region in GCL and TcAChE is similar, GCL has Ser and Ile in place of the cysteines. The predominantly hydrophobic character of the residues in these positions (Ile/Val in one and Leu in the other) suggests that they may also come into contact with each other in those proteins that lack the disulfide bridge.

Residues at the edges of secondary structural elements

Most of the strictly conserved residues are located in turns and loops at the edges of the β-strands or α-helices and seem to be important for maintaining the proper fold of the backbone. They are especially frequent on the C-terminal side of the parallel β-strands, where the active site is always located (Brändén & Tooze, 1991). Their environment and possible role are summarized in Table 4. The following paragraphs provide a discussion of those positions that require more elaboration.

Positions Gly [130,117], Gly [131,118], Gly/Ala [132, 119], and Phe/Leu [133,120] form a loop, $L^{1}_{3,4}$, after strand β_3 (Fig. 1). Although this sequence and a few preceding residues in GCL and TcAChE show only one difference (at position [132,119]), the conformation of this loop in these two proteins is quite different. In both enzymes, this loop is in close proximity to the active site and most likely plays a role in the catalytic process. The amino group of Gly 119 in TcAChE has been suggested to be part of the oxyanion hole (Sussman et al., 1991). A similar role has been postulated for the equivalent Ala 132 of GCL (Schrag et al., 1991). The side chain in position [129,116], at the beginning of the $L^{1}_{3,4}$ loop (usually Tyr or His), stacks against the side chain of Tyr [164,148], which is conserved in all sequences except neurotactin. Tyr [164,148] is additionally hydrogen bonded to the carbonyl of residue [129,116], and the latter makes a hydrogen bond to the backbone carbonyl of Tyr [164,148]. Residue [129,116] is tightly constrained by these interactions and may provide a pivot for rotation of loop $L^{1}_{3,4}$ ([130,117]–[137,124]). Such a rotation might be required as part of a conformational change in GCL upon binding to the lipid/water interface and/or binding of substrate. Because no structure of a GCL–substrate analog complex has yet been determined, we do not know if this loop changes its conformation upon substrate binding.

The region of conserved residues [167,151]–[171,155] forms a distorted α-helical turn. Position [167,151], in all sequences except two, is occupied by a glycine at the beginning of the turn. The Gly backbone ϕ,ψ angles are 66°,−155° in GCL and 69°,−167° in TcAChE and correspond to an area of the Ramachandran plot, which, for residues other than glycine, is not accessible without internal strain. Gly [170,154] and Phe [171,155] occur at the end of this turn in all except one sequence. The requirement for Gly most likely originates from the tight packing of this residue against the protein backbone (the carbonyl of conserved Arg [165,149]) and the aromatic ring of Phe [171,155].

Table 4. *Structural context of the conserved (V = 1) residues*[a]

Residue	Position	Context	Possible function
G[9,13]	$l_{1,2}$		
G[15,17]	$L_{b2,0}$		Close packing against F/Y
P[28,34]	$L_{1,b3}$	I/V **P** F/Y **A**	
A[30,36]			
C[61,67]	$L_{b3,0}$	**C** x Q	S-S bridge to C[105,94]
E[103,92]	$L_{b3,2}$	s **E** D C L y	Salt bridge
G[130,117]			
G[131,118]	$L^1_{3,4}$	**G G** G/A F/L x x **G**	Oxyanion hole loop
G[136,123]			
Y[164,148]			Ring stacking
R[165,149]	$L^1_{4,5}$	**Y R** V/L g x x **G** F/L	Salt bridge
G[170,154]			Packing against R[165,149]
N[184,167]	$L^1_{4,5}$	**N** x g l	Anchoring $\alpha^2_{4,5}$
N[200,183]	$\alpha^2_{4,5}$		
F[204,187]	$L^3_{4,5}$	**N** i a x **F G G** d P	Packing
G[205,188]			$(\phi,\psi)_{GCL}$ = 100°,10°
G[206,189]			
G[215,198]	β_5		Close packing
S[217,200]		V/I x I/L f **G** e **S** A **G** A/G	Active site
G[219,202]	$\alpha_{5,6}$		Close packing
S[249,226]	$L^1_{6,7}$	i x x **S** G	H-bond to E[354,327]
C[276,254][b]	$L^2_{6,7}$		S-S bridge to C[288,265]
E[354,327]	$L^1_{7,8}$	d **E/D** g	Active site
D[425,397]	$\alpha^3_{7,8}$	**D** x x F/V	Salt bridge
H[463,440]	$L^1_{8,9}$	g x x **H** G/A x E/D	Active site
F[488,476]	$\alpha^2_{8,9}$	W/F a n **F** A	Packing volume

[a] One letter code is used. In the context column a capital letter identifies a highly conserved residue (bold for invariant residue), X/Y means that only residues X or Y are observed in this position, small letter indicates the type of residue with the highest abundance, and the letter x indicates a variable position.

[b] Not conserved in two CEs: HeliotCE and CulexCEb1.

The helix $\alpha^2_{4,5}$ ([185,168]-[200,183]), which runs nearly parallel to the large β-sheet over strands β_1-β_5, has conserved residues on the side facing the sheet (Fig. 5). Asn [184,167] anchors the N-terminal end of helix $\alpha^2_{4,5}$ through hydrogen bonds to the backbone of residue [322,295]. This Asn is found in a conformation with the backbone torsion angles in the left-handed helical region (ϕ,ψ of 56°,40° in GCL; 25°,29° in TcAChE) usually accessible only to Gly and Asn. Asn [200,183] is just before a highly irregular helical turn formed by residues [201,184]-[204,187], extending helix $\alpha^2_{4,5}$. Its backbone torsion angles, −15°,−122° for GCL and −5°,−143° for TcAChE, fall into a somewhat unfavorable region of the Ramachandran plot, but we do not have a good explanation for its preservation. The side chains of this residue and of residue [199,182] are exposed, and the latter is usually charged. The conserved Phe [204,187], at the end of the helical turn, packs tightly against the interior of the protein (conserved Trp [196,179] and Phe/Tyr [24,30]). Residues forming a turn between helix $\alpha^2_{4,5}$ and strand β_5, which contains at its end the active-site serine, are also strongly conserved in all sequences. Gly [205,188], starting this turn, assumes an unusual conformation with ϕ,ψ of 100°,10° in GCL (125°,0° in TcAChE). The neighboring Gly [206,189] faces the interior and packs closely against the backbone of residue [201,184]. Position [207,190] is occupied by a hydrophilic residue and is exposed, whereas residue [208,191] is nearly always a proline.

Residues [416,388]-[438,410] form a kinked helix, $\alpha^4_{7,8}$, both in TcAChE and in GCL. This helix lies over strands β_6-β_8 and lines one side of the active site. The kink is at position [428,400], usually an aromatic residue, and this side chain is stacked against the ring of His [463,440] and the hydrophobic part of the Glu [354,327]

side chain, members of the catalytic triad. Although the sequence of this helix is not strongly preserved, its hydrophobicity pattern is maintained in all sequences, supporting the notion that an α-helical arrangement also exists in other proteins of this family. In the cholinesterases, one of the cysteines of the third disulfide bridge comes from this helix.

Helix $\alpha^2_{8,9}$ points with its C-terminal end to the N-termini of strands β_6 and β_7 of the β-sheet and is anchored to them by the highly conserved C-terminal aromatic residues Phe/Trp [485,473] and Phe [488,476].

Many of the remaining highly conserved residues form a cluster and pack against each other. This group includes residues [445,419], [447,421], [485,473], [488,476], [494, 482], and [503,492]. Most maintain an aromatic character in all sequences. A buried water molecule was found in this region in both GCL and TcAChE. It forms four hydrogen bonds with nearly ideal tetrahedral coordination: three to the backbone NH and carbonyl groups and the fourth to the NϵH of the indole ring of Trp [503,492]. The presence of this buried water may explain the strong conservation of Trp rather than a more relaxed preservation of an aromatic character of this residue. This Trp [503,492] and Pro [494,482] anchor the bottom of a rather flexible loop that is disordered in TcAChE. Finally, the last 2 residues with low variability are at position [533, 521], which is either Cys (of the third disulfide bridge in AChEs) or Ile/Leu, and at position [536,524], occupied by Trp/Phe. The latter packs against conserved Asp [425, 397] and Tyr [394,375].

High variability positions

In parallel with analysis of the conserved residues, it is important to address the question of spatial distribution of positions with high variability. Most of these positions are also clustered (Table 3). High variability clusters are distributed throughout the sequences and correspond to parts of the polypeptide chain that are on the surface of the molecule in the 3D structures of both GCL and TcAChE (Fig. 4). There is a concentration of high variability residues near the substrate-binding site. Because these enzymes bind different substrates, positions with high variability in this region may be involved in substrate binding. As was noted in other cases, insertions/deletions occur on the protein surface (Delbaere et al., 1975) and are rather evenly distributed over the entire surface.

Conclusions

The degree of amino acid similarity in the proteins identified here is somewhat higher than that found in the globins. Lesk and Chothia (1980) compared the 3D structures of nine globins and identified 5 (~3.3%) residues as absolutely conserved, 2 involved in heme and oxygen binding and the other 3 in packing of α-helices. They also found that residues in the protein core tend to be more conserved than the residues on the surface. The study of Siezen et al. (1991) on 35 subtilisin-like proteases revealed 11 absolutely conserved residues (~3.3%), 3 corresponding to the catalytic triad and another 5 in the substrate-binding region. They identified 36 positions (~12%) as highly conserved (roughly corresponding to $V < 4$). Nine of these highly conserved residues were Gly, many of which have main-chain torsion angles in the range not easily accessible to other residues. The corresponding numbers for the proteins discussed here (4.5% for absolutely conserved and 14% for highly conserved residues with 14 glycines) are somewhat higher than in the other two classes of proteins mentioned above and possibly reflect the higher β-sheet content in esterases.

N-terminal subdomains display a higher percentage of conserved residues than C-terminal subdomains and show a strong conservation pattern, indicating the likelihood that these parts of the proteins will have very similar 3D structures. Although the structures of GCL and TcAChE also show striking similarities in their C-terminal parts, it is likely that changes occur in the C-terminal parts of some of the other proteins. These changes may affect the mutual orientation of the N- and C-terminal subdomains. The observed conservation of some contact residues between these subdomains suggests, however, that a reasonable similarity of packing should be expected. Comparison of TcAChE and GCL showed in fact some rotation of the C-terminal part around strand β_7, relative to the N-terminal part. A similar effect has been observed in the α/β hydrolase fold enzymes, where a different degree of bending of the β-sheet was observed between strands β_1–β_7 and strands β_8–β_9 (Ollis et al., 1992).

Salt bridges play an important role in the preservation of the esterase fold by holding together loops in the vicinity of the active site. The disulfide bridges in the N-terminal part of the structures are most likely important for maintenance of the conformation of two loops, one of which ([64,70]–[101,90]) is clearly involved in substrate binding (M. Harel et al., in prep.). The level of conservation in the C-terminal part of the sequences, where less of the protein is involved in forming the scaffold and more is in the loops, is lower. The pairwise homology of the C-terminal parts is in many cases not very significant, just as that found in distantly related globins. The identity between the C-terminal parts of TcAChE and GCL is only 13%, yet the 3D structures of these subdomains are very similar.

The picture emerging for proteins of this family is, in a way, analogous to that of immunoglobulin variable domains, where the conserved scaffold of the protein is formed by a β-barrel from which extend the loops forming the hapten-binding site (Davies & Metzger, 1983). The scaffold of the proteins discussed in this paper, formed by a large β-sheet and crossover helices, is well conserved

and provides a stable environment for the catalytic triad, whereas substrate specificity is determined by the loops covering the scaffold and surrounding the active site. As discussed above, a rational explanation for conservation can be offered for highly conserved residues in terms of the structural or functional requirements of these proteins. Positions conserved in only a subset of these enzymes are often associated with extended loops or other less rigid elements involved in providing substrate specificity.

Methods

The sequences included in this comparison are listed in Table 2. They were selected through a literature search and through searches of the SWISSPROT data base using the profile analysis method (Gribskov et al., 1987, 1990). The sequence alignment proceeded in steps. First, the sequences of TcAChE and GCL were aligned based on the superposition of their 3D structures. They were used, together with *T. marmorata* AChE and *Candida rugosa* lipase sequences, to construct a profile for searches of the data base (program PROFILEMAKE, GCG package [Devereux et al., 1984]). Sequences selected this way (program PROFILESEARCH, GCG package) were then aligned with the multiple sequence alignment program PILEUP. In a number of places (e.g., region 368–400 in GCL) automatic alignment of sequences was at variance with the structural alignment. These discrepancies, with respect to regions of the 3D structures of TcAChE and GCL displaying very low homology, suggested a few additional modifications to the alignment. Insertions were kept to a minimum and were not allowed in the middle of secondary structural elements common to GCL and AChE, unless suggested by a strong sequence similarity. The final alignment is shown in Figure 2.

The variability index at each position of the aligned set of sequences was calculated following a procedure developed earlier for the analysis of immunoglobulin sequences (Kabat et al., 1983). The variability, V, at any given position is defined as $V = n/p$, where n is the number of different amino acids occurring at this position, and p is the fraction of sequences with the amino acid of highest frequency of occurrence. This parameter varies from 1, for an absolutely conserved residue, to a value of 400 for a position where there is an equal probability of finding any one of the 20 amino acids. The square root of V may be interpreted as a (weighted) number of the different amino acids to be expected at this position. The positions at which some sequences have a deletion (gap) required special treatment. For such a position a deletion in any sequence was treated as a new amino acid type. Effectively, that amounted to a penalty of 1 added to n for each sequence with a deletion, whereas p was calculated as a fraction of all sequences. For such a position the variability index could become greater than 400.

Acknowledgments

We thank Mr. M. Desrochers for help in all computational aspects of this work and in preparation of figures. This is NRCC publication no. 33711.

References

Baba, T., Downs, D., Jackson, K.W., Tang, J., & Wang, C.-S. (1991). Structure of human milk bile salt activated lipase. *Biochemistry 30*, 500–510.

Boel, E., Huge-Jensen, B., Christensen, M., Thim, L., & Fiil, N.P. (1988). *Rhizomucor miehei* triglyceride lipase is synthesized as a precursor. *Lipids 23*, 701–706.

Bomblies, L., Biegelmann, E., Döring, V., Gerisch, G., Krafft-Czepa, H., Noegel, A.A., Schleicher, M., & Humbel, B.M. (1990). Membrane-enclosed crystals in *Dictyostelium discoideum* cells, consisting of developmentally regulated proteins with sequence similarities to known esterases. *J. Cell Biol. 110*, 669–679.

Borgström, B. & Brockman, H.L., Eds. (1984). *Lipases*. Elsevier, Amsterdam.

Brady, L., Brzozowski, A.M., Derewenda, Z.S., Dodson, E., Dodson, G., Tolley, S., Turkenburg, J.P., Christiansen, L., Huge-Jensen, B., Norskov, L., Thim, L., & Menge, U. (1990). A serine protease triad forms the catalytic centre of a triacylglycerol lipase. *Nature 343*, 767–770.

Brändén, C. & Tooze, J. (1991). *Introduction to Protein Structure*. Garland Publishing, New York.

Brenner, S. (1988). The molecular evolution of genes and proteins: A tale of two serines. *Nature 334*, 528–530.

Bucht, G., Artursson, E., Häggström, B., Osterman, A., & Hjalmarsson, K. (1992). Structurally important residues in the region Ser91 to Asn98 of *Torpedo* acetylcholinesterase. 36th Oholo Conference, April 6–10, Eilat, Israel.

Collet, C., Nielsen, K.M., Russell, R.J., Karl, M., Oakeshott, J.G., & Richmond, R.C. (1990). Molecular analysis of duplicated esterase genes in *Drosophila melanogaster*. *Mol. Biol. Evol. 7*, 9–28.

Cygler, M., Schrag, J.D., & Ergan, F. (1992). Advances in structural understanding of lipases. *Biotechnol. Genet. Rev. 10*, 141–181.

Davies, D.R. & Metzger, H.A. (1983). Structural basis of antibody function. *Annu. Rev. Immunol. 1*, 87–117.

de la Escalera, S., Bockamp, E.-O., Moya, F., Piovant, M., & Jiménez, F. (1990). Characterization and gene cloning of neurotactin, a *Drosophila* transmembrane protein related to cholinesterases. *EMBO J. 9*, 3593–3601.

Delbaere, L.T.J., Hutcheon, W.L.B., James, M.N.G., & Thiessen, W.E. (1975). Tertiary structural differences between microbial serine proteases and pancreatic serine enzymes. *Nature 257*, 758–743.

Derewenda, Z.S. & Derewenda, U. (1991). Relationships among serine hydrolases: Evidence for a common structural motif in triacylglyceride lipases and esterases. *Biochem. Cell Biol. 69*, 842–851.

Devereux, J., Haeberli, P., & Smithies, O. (1984). A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res. 12*, 387–395.

DiPersio, L.P., Fontaine, R.N., & Hui, D.Y. (1990). Identification of the active site serine in pancreatic cholesterol esterase by chemical modification and site-specific mutagenesis. *J. Biol. Chem. 265*, 16801–16806.

DiPersio, L.P., Fontaine, R.N., & Hui, D.Y. (1991). Site-specific mutagenesis of an essential histidine residue in pancreatic cholesterol esterase. *J. Biol. Chem. 266*, 4033–4036.

Doctor, B.P., Chapman, T.C., Christner, C.E., Deal, C.D., de la Hoz, D.M., Gentry, M.K., Ogert, R.A., Rush, R.S., Smyth, K.K., & Wolfe, A.D. (1990). Complete amino acid sequence of fetal bovine serum acetylcholinesterase and its comparison in various regions with other cholinesterases. *FEBS Lett. 266*, 123–127.

Doolittle, R.F. (1985). Proteins. *Sci. Am. 253*(4), 88–99.

Fournier, D., Bride, J.-M., Karch, F., & Bergé, J.-B. (1988). Acetylcholinesterase from *Drosophila melanogaster*. Identification of two subunits encoded by the same gene. *FEBS Lett. 238*, 333–337.

Gentry, M.K. & Doctor, B.P. (1991). Alignment of amino acid sequences of acetylcholinesterases and butyrylcholinesterases. In *Cholinesterases: Structure, Function, Mechanism, Genetics and Cell Biology*

(Massoulié, J., Bacou, F., Barnard, E., Chatonnet, A., Doctor, B.P., & Quinn, D.M., Eds.), pp. 394–398. American Chemical Society, Washington, D.C.

Gibney, G., Camp, S., Dionne, M., MacPhee-Quigley, K., & Taylor, P. (1990). Mutagenesis of essential functional residues in acetylcholinesterase. *Proc. Natl. Acad. Sci. USA 87*, 7546–7550.

Gnagey, A.L., Forte, M., & Rosenberry, T.L. (1987). Isolation and characterization of acetylcholinesterase from *Drosophila*. *J. Biol. Chem. 262*, 1140–1145.

Greer, J. (1990). Comparative modeling methods: Application to the family of the mammalian serine proteases. *Proteins Struct. Funct. Genet. 7*, 317–334.

Gribskov, M., Luthy, R., & Eisenberg, D. (1990). Profile analysis. *Methods Enzymol. 183*, 146–159.

Gribskov, M., McLachlan, A.D., & Eisenberg, D. (1987). Profile analysis: Detection of distantly related proteins. *Proc. Natl. Acad. Sci. USA 84*, 4355–4358.

Hall, L.M.C. & Malcolm, C.A. (1991). The acetylcholinesterase gene of *Anopheles stephensi*. *Cell. Mol. Neurobiol. 11*, 131–141.

Hall, L.M.C. & Spierer, P. (1986). The *Ace* locus of *Drosophila melanogaster*: Structural gene for acetylcholinesterase with an unusual 5′ leader. *EMBO J. 5*, 2949–2954.

Han, J.H., Stratowa, C., & Rutter, W.J. (1987). Isolation of full-length putative rat lysophospholipase cDNA using improved methods for mRNA isolation and cDNA cloning. *Biochemistry 26*, 1617–1625.

Hanzlik, T.N., Abdel-Aal, Y.A.I., Harshman, L.G., & Hammock, B.D. (1989). Isolation and sequencing of cDNA clones coding for juvenile hormone esterase from *Heliothis virescens*. *J. Biol. Chem. 264*, 12419–12425.

Hui, D.Y. & Kissel, J.A. (1990). Sequence identity between human pancreatic cholesterol esterase and bile salt-stimulated milk lipase. *FEBS Lett. 26*, 131–134.

Jbilo, O. & Chatonnet, A. (1990). Complete sequence of rabbit butyrylcholinesterase. *Nucleic Acids Res. 18*, 3990.

Kabat, E.A., Wu, T.T., Bilofsky, H., Reid-Miller, M., & Perry, H. (1983). *Sequences of Proteins of Immunological Interest*. National Institutes of Health, Bethesda, Maryland.

Kawaguchi, K., Honda, H., Taniguchi-Morimura, J., & Iwasaki, S. (1989). The codon CUG is read as serine in an asporogenic yeast *Candida cylindracea*. *Nature 341*, 164–166.

Kissel, J.A., Fontaine, R.N., Turck, C.W., Brockman, H.L., & Huik, D.Y. (1989). Molecular cloning and expression of cDNA for rat pancreatic cholesterol esterase. *Biochim. Biophys. Acta 1006*, 227–236.

Korza, G. & Ozols, J. (1988). Complete covalent structure of esterase isolated from 2,3,7,8-tetrachlorodibenzo-*p*-dioxin-induced rabbit liver microsomes. *J. Biol. Chem. 263*, 3486–3495.

Krejci, E., Duval, N., Chatonnet, A., Vincens, P., & Massoulié, J. (1991). Cholinesterase-like domains in enzymes and structural proteins: Functional and evolutionary relationships and identification of a catalytically essential aspartic acid. *Proc. Natl. Acad. Sci. USA 88*, 6647–6651.

Kyger, E.M., Wiegand, R., & Lange, L.G. (1989). Cloning of the bovine pancreatic cholesterol esterase/lysophospholipase. *Biochem. Biophys. Res. Commun. 164*, 1302–1309.

Lesk, A.M. & Chothia, C. (1980). How different amino acid sequences determine similar protein structures: The structure and evolutionary dynamics of the globins. *J. Mol. Biol. 136*, 225–270.

Levitt, M. (1978). Conformational preferences of amino acids in globular proteins. *Biochemistry 17*, 4277–4285.

Lockridge, O., Bartels, C.F., Vaughan, T.A., Wong, C.K., Norton, S.E., & Johnson, L.L. (1987). Complete amino acid sequence of human serum cholinesterase. *J. Biol. Chem. 262*, 549–557.

Long, R.M., Calabrese, M.R., Martin, B.M., & Pohl, L.R. (1991). Cloning and sequencing of a human liver carboxylesterase isoenzyme. *Life Sci. 48*, 43–49.

Long, R.M., Satoh, H., Martin, M., Kimura, S., Gonzalez, F.J., & Pohl, L.R. (1988). Rat liver carboxylesterase: cDNA cloning, sequencing, and evidence for a multigene family. *Biochem. Biophys. Res. Commun. 156*, 866–873.

Longhi, S., Fusetti, F., Grandori, R., Lotti, M., Vanoni, M., & Alberghina, L. (1992). Cloning and nucleotide sequences of two lipase genes from *Candida cylindracea*. *Biochim. Biophys. Acta 1131*, 227–231.

Mercken, L., Simmons, M.-J., Swillens, S., Massaer, M., & Vassart, G. (1985). Primary structure of bovine thyroglobulin deduced from the sequence of its 8,431-base complementary DNA. *Nature 316*, 647–651.

Mouches, C., Pauplin, Y., Agarwal, M., Lemieux, L., Herzog, M., Abadon, M., Beyssat-Arnaouty, V., Hyrien, O., de St. Vincent, B.R., Georghiou, G.P., & Pasteur, N. (1990). Characterization of amplification core and esterase B gene responsible for insecticide resistance in *Culex*. *Proc. Natl. Acad. Sci. USA 87*, 2574–2578.

Munger, J.S., Shi, G.-P., Mark, E.A., Chin, D.T., Gerard, C., & Chapman, H.A. (1991). A serine esterase released by human alveolar macrophages is closely related to liver microsomal carboxylesterases. *J. Biol. Chem. 266*, 18832–18838.

Neville, L.F., Gnatt, A., Loewenstein, Y., Seidman, S., Ehrlich, G., & Soreq, H. (1992). Intra-molecular relationships in cholinesterases revealed by oocyte expression of site-directed and natural variants of human BChE. *EMBO J. 11*, 1641–1649.

Nilsson, J., Bläckberg, L., Carlsson, P., Enerbäck, S., Hernell, O., & Bjursell, G. (1990). cDNA colning of human-milk bile-salt-stimulated lipase and evidence for its identity to pancreatic carboxylic ester hydrolase. *Eur. J. Biochem. 192*, 543–550.

Oakeshott, J.G., Collet, C., Phillis, R.W., Nielsen, K.M., Russell, R.J., Chambers, G.K., Ross, V., & Richmond, R.C. (1987). Molecular cloning and characterization of esterase-6, a serine hydrolase of *Drosophila*. *Proc. Natl. Acad. Sci. USA 84*, 3359–3363.

Ollis, D.L., Cheah, E., Cygler, M., Dijkstra, B., Frolow, F., Franken, S.M., Harel, M., Remington, S.J., Silman, I., Schrag, J.D., Sussman, J.L., Verschueren, K.H.G., & Goldman, A. (1992). The α/β hydrolase fold. *Protein Eng. 5*, 197–211.

Olson, P.F., Fessler, L.I., Nelson, R.E., Sterne, R.E., Campbell, A.G., & Fessler, J.H. (1990). Glutactin, a novel *Drosophila* basement membrane-related glycoprotein with sequence similarity to serine esterases. *EMBO J. 9*, 1219–1227.

Ovnic, M., Tepperman, K., Medda, S., Elliott, R.W., Stephenson, D.A., Grant, S.G., & Ganschow, R.E. (1991). Characterization of a murine cDNA encoding a member of the carboxylesterase multigene family. *Genomics 9*, 344–354.

Ozols, J. (1989). Isolation, properties, and the complete amino acid sequence of a second form of 60-kDa glycoprotein esterase. *J. Biol. Chem. 264*, 12533–12545.

Quinn, D.M. (1987). Acetylcholinesterase: Enzyme, structure, reaction dynamics and virtual transition states. *Chem. Rev. 87*, 955–979.

Rachinsky, T.L., Camp, S., Li, Y., Elström, T.J., Newton, M., & Taylor, P. (1990). Molecular cloning of mouse acetylcholinesterase: Tissue distribution of alternatively spliced mRNA species. *Neuron 5*, 317–327.

Robbi, M., Beaufay, H., & Octave, J.-N. (1990). Nucleotide sequence of cDNA coding for rat liver pI 6.1 esterase (ES-10), a carboxylesterase located in the lumen of the endoplasmic reticulum. *Biochem. J. 269*, 451–458.

Rubino, S., Mann, S.K.O., Hori, R.T., Pinko, C., & Firtel, R.A. (1989). Molecular analysis of a developmentally regulated gene required for *Dictyostelium* aggregation. *Dev. Biol. 131*, 27–36.

Schrag, J.D. & Cygler, M. (1993). The 1.8 Å refined structure of lipase from *Geotrichum candidum*. *J. Mol. Biol. 230*.

Schrag, J.D., Li, Y., Wu, S., & Cygler, M. (1991). Ser-His-Glu trian forms the catalytic site of the lipase from *Geotrichum candidum*. *Nature 351*, 761–765.

Schumacher, M., Camp, S., Maulet, Y., Newton, M., MacPhee-Quigley, K., Taylor, S.S., Friedmann, T., & Taylor, P. (1986). Primary structure of *Torpedo californica* acetylcholinesterase deduced from its cDNA sequence. *Nature 319*, 407–409.

Shafferman, A., Kronman, C., Flashner, Y., Leitner, M., Grosfeld, H., Ordentlich, A., Gozes, Y, Cohen, S., Ariel, N. Barak, D., Harel, M., Silman, I., Sussman, J.L., & Velan, B. (1992). Mutagenesis of human acetylcholinesterase. Identification of residues involved in catalytic activity and in polypeptide folding. *J. Biol. Chem. 267*, 17640–17648.

Shimada, Y., Sugihara, A., Iizumi, T., & Tominaga, Y. (1990). cDNA cloning and characterization of *Geotrichum candidum* lipase II. *J. Biochem. 107*, 703–707.

Shimada, Y., Sugihara, A., Tominaga, Y., Iizumi, T., & Tsunasawa, S. (1989). cDNA molecular cloning of *Geotrichum candidum* lipase. *J. Biochem. 106*, 383–388.

Siezen, R.J., de Vos, W.M., Leunissen, J.A.M., & Dijkstra, B.W. (1991).

Homology modelling and protein engineering strategy of subtilases, the family of subtilisin-like serine proteinases. *Protein Eng. 4*, 719–737.

Sikorav, J.-L., Krejci, E., & Massoulié, J. (1987). cDNA sequences of *Torpedo marmorata* acetylcholinesterase: Primary structure of the precursor of a catalytic subunit; existence of multiple 5′-untranslated regions. *EMBO J. 6*, 1865–1873.

Silman, I. & Futerman, A.H. (1987). Modes of attachment of acetylcholinesterase to the surface membrane. *Eur. J. Biochem. 170*, 11–22.

Slabas, A.R., Windust, J., & Sidebottom, C.M. (1990). Does sequence similarity of human choline esterase, *Torpedo* acetylcholine esterase and *Geotrichum candidum* lipase reveal the active site serine residue? *Biochem. J. 269*, 279–280.

Soreq, H., Ben-Aziz, B., Prody, C.A., Seidman, S., Gnatt, A., Neville, L., Lieman-Hurwitz, J., Lev-Lehman, E., Ginzberg, D., Lapidot-Lifson, Y., & Zakut, H. (1990). Molecular cloning and construction of the coding region for human acetylcholinesterase reveals a G+C-rich attenuating structure. *Proc. Natl. Acad. Sci. USA 87*, 9688–9692.

Sussman, J.S., Harel, M., Frolov, F., Oefner, C., Goldman, A., Toker, L., & Silman, I. (1991). Atomic structure of acetylcholinesterase from *Torpedo californica*: A prototypic acetylcholine-binding protein. *Science 253*, 872–879.

Takagi, Y., Morohashi, K., Kawabata, S., G, M., & Omura, T. (1988). Molecular cloning and nucleotide sequence of cDNA of microsomal carboxyesterase E1 of rat liver. *J. Biochem. 104*, 801–806.

Winkler, F.K., D'Arcy, A., & Hunziker, W. (1990). Structure of human pancreatic lipase. *Nature 343*, 771–774.