Integrative Biology

Protein structure networks provide insight into active site flexibility in esterase/lipases from the carnivorous plant *Drosera capensis*

Journal:	*Integrative Biology*
Manuscript ID	IB-ART-08-2018-000140.R1
Article Type:	Paper
Date Submitted by the Author:	22-Nov-2018
Complete List of Authors:	Duong, Vy; University of California, Irvine, Chemistry; University of California, Irvine, Molecular Biology & Biochemistry Unhelkar, Megha; University of California, Irvine, Chemistry Kelly, John; University of California, Irvine, Chemistry Kim, Suhn; University of California, Irvine, Chemistry Butts, Carter; University of California, Irvine, Sociology, Statistics, and Electrical Engineering & Computer Science Martin, Rachel; University of California, Irvine, Chemistry; University of California, Irvine, Molecular Biology & Biochemistry

Esterase/lipases are enzymes involved in biosynthesis of the surface coatings used by plants to form a robust barrier between their tissues and the environment, a critical function for pathogen resistance and drought tolerance. Carnivorous plants in particular must protect the leaf surface from their own digestive enzymes as well as the bacteria and fungi that may grow in this nutrient-rich environment. These enzymes are potentially useful for producing biodegradable hydrophobic coatings. Here we present a method for using molecular modeling and protein structure network analysis to predict which esterase/lipases catalyze a specific reaction vs. acting as more general-purpose catalysts. We find that most of the esterase/lipases from *Drosera capensis* have relatively constrained active sites, consistent with specific functionalities.

Journal Name

ARTICLE TYPE

Cite this: DOI: 10.1039/xxxxxxxxxx

Protein structure networks provide insight into active site flexibility in esterase/lipases from the carnivorous plant *Drosera capensis*

Vy T. Duong,[1,2] Megha H. Unhelkar,[1] John E. Kelly,[1] Suhn H. Kim,[1] Carter T. Butts,[3*] Rachel W. Martin[1,2*]

Received Date
Accepted Date

DOI: 10.1039/xxxxxxxxxx

www.rsc.org/journalname

In plants, esterase/lipases perform transesterification reactions, playing an important role in the synthesis of useful molecules, such those comprising the waxy coatings of leaf surfaces. Plant genomes and transcriptomes have provided a wealth of data about expression patterns and the circumstances under which these enzymes are upregulated, e.g. pathogen defense and response to drought; however, predicting their functional characteristics from genomic or transcriptome data is challenging due to weak sequence conservation among the diverse members of this group. Although functional sequence blocks mediating enzyme activity have been identified, progress to date has been hampered by the paucity of information on the structural relationships among these regions and how they affect substrate specificity. Here we present methodology for predicting overall protein flexibility and active site flexibility based on molecular modeling and analysis of protein structure networks (PSNs). We define two new types of specialized PSNs: sequence region networks (SRNs) and active site networks (ASNs), which provide parsimonious representations of molecular structure in reference to known features of interest. Our approach, intended as an aid to target selection for poorly characterized enzyme classes, is demonstrated for 26 previously uncharacterized esterase/lipases from the genome of the carnivorous plant *Drosera capensis* and validated using a case/control design. Analysis of the network relationships among functional blocks and among the chemical moieties making up the catalytic triad reveals potentially functionally significant differences that are not apparent from sequence analysis alone.

1 Introduction

In land plants, tissues that are exposed to air are protected by the cuticle, a composite biomaterial comprising a cross-linked polyester scaffold interpenetrated by wax components[1]. The cuticle provides a barrier that minimizes water loss and protects the plant from pathogen infection. The relative quantities of hydrophilic and hydrophobic components must be appropriately balanced and spatially located to adhere to the underlying cell walls while presenting a hydrophobic surface to the air interface[2]. Numerous enzymes are involved in producing the polymer components of this material, including esterases, lipases, and GDSL esterase/lipases. Herein we focus on the GDSL esterase/lipases, characterized by the proximity of the active serine residue to the N-terminus, as well as by its surrounding residues (canonically GDSL)[3]. Esterase/lipases belong to the large α/β hydrolase enzyme superfamily, in which the catalytic triad consists of a nucleophile, an acid, and a stabilizing histidine (in this case Ser-Asp-His). In plants, these enzymes are often localized to the cuticle matrix, where they catalyze the reverse reaction (biosynthesis of polyesters) rather than acting as hydrolases[4]. This biosynthetic activity in the waxy cuticle is consistent with *in vitro* results indicating that esterase/lipases are highly tolerant of hydrophobic environments, where they catalyze the formation of polyesters rather than performing hydrolysis reactions[5].

Esterase/lipases present attractive targets for biotechnology applications because of their potential for producing robust yet ultimately biodegradable polyester materials and hydrophobic surface coatings[6–8]. Several microbial GDSL proteins have been characterized as relatively promiscuous enzymes that serve a variety of purposes (e.g. protease, lysophospholipase, thioesterase, arylesterase)[9,10], and accomodating a wide range of sub-

[1] *Department of Chemistry, UC Irvine*
[2] *Department of Molecular Biology & Biochemistry, UC Irvine*
[3] *Departments of Sociology, Statistics, and Electrical Engineering & Computer Science, UC Irvine*
[*] *To whom correspondence should be addressed; E-mail: rwmartin@uci.edu, buttsc@uci.edu.*
Address: Irvine, CA, 92697 USA

† Electronic Supplementary Information (ESI) available: [details of any supplementary information available should be included here]. See DOI: 10.1039/b000000x/

strates[11]. Microbial cutinases, a subclass of serine esterases found in fungi and bacteria, catalyze esterification and transesterfication and can hydrolyze both hydrophobic and lipid substrates in solution or emulsion[12]. In a chemical biology or biotechnology setting, enzymes with different degrees of specificity may be preferred for different applications; for example, promiscuous enzymes are useful for generalized hydrolysis, while those catalyzing a specific reaction are more useful for biosynthetic reactions. Harnessing the potential of these enzymes, given the enormous number of uncharacterized sequences available, requires methodology for predicting their functional characteristics.

Plant GDSL esterase/lipases may provide a rich source of particular chemical functionalities. Many such enzymes have been discovered from genome and transcriptome data[13,14]; however their specific functions and substrate preferences remain relatively unexplored despite their potential commercial and technological importance. 114 esterase/lipases have been identified from the genome of rice (*Oryza sativa*) alone[15], and a survey of 12 plant proteomes found that each plant has many esterase/lipase isoforms, including multiple unique genes as well as splice variants[16]. In genomic terms, the large number of GDSL esterase lipases found in plants results from several gene duplication events, followed by selection for novel functions and/or neutral drift[17]. Although in many cases their precise catalytic activities are yet unknown, esterase/lipases are associated with developmental processes[18], pollen exine formation[19], salt tolerance[20], and stress responses[21,22]. Many of these functions appear to be related to the biosynthesis and metabolism of cutin and waxes[23,24]. A recent investigation by Zhang et al. demonstrated the first plant GDSL (BS1) to exhibit polysaccharide esterase activity, which is vital for maintaining secondary cell wall acetylation levels and homeostasis[25]. In the oil palm (*Elaeis guineensis*), oil yield correlates with expression of genes for GDSL esterase/lipases and expression of these genes in transgenic *Arabidopsis* plants increases their fatty acid production as well[26].

Much of what is known to date about the specific enzymatic activities of proteins in this family comes from studies of either model systems such as *Arapidopsis thaliana* or crop plants that produce large fruits[27]. For example, in the tomato (*Solanum lycopersicum*), the GDSL1 enzyme is required for cuticle formation; knockdown of expression of the GDSL1 enzyme (also called CD1) using RNAi results in porous fruit cuticles. On the molecular level, both a decrease in the density of cutin monomers and a reduction in ester bond cross-links between the polymer chains were observed[4], consistent with the phenotype of the *cd1* mutant, in which this gene is interrupted by a stop codon. Cutin deficiency caused by the *cd1* mutation reduces the thickness of the cuticle, decreases its mechanical flexibility, and increases its susceptibility to water loss, unlike some other cutin-deficient mutants[28]. GDSL1 (CD1) acts as an acyltransferase, building up the polyester oligomers of the cuticle[29]. This finding highlights the importance of characterizing esterase/lipases in plants; studies in *A. thaliana* have shown that multiple enzymes are required to form a functional cuticle[30], and technological applications will likely also require a series of enzymatic reactions. The esterase/lipases from carnivorous plants have the potential to be particularly useful from a biotechnology standpoint because of the unique challenges faced by their leaf surfaces, which must withstand the harsh chemical environment associated with their digestive fluids for extended time periods.

Here, we present molecular modeling and functional analyses of 26 esterase/lipases recently discovered from the genome of the Cape sundew (*Drosera capensis*)[31]. The conservation of active site residues, key functional sequence blocks, and overall protein folds suggests that many of the *D. capensis* esterase/lipase sequences form functional enzymes; however the diversity of sequence and structural features indicates a range of potential molecular targets and enzymatic activities. We use sequence analysis, comparative modeling with all-atom refinement followed by *in silico* maturation, and comparison of protein structure networks (PSNs) to identify distinct subgroups of proteins as a first step toward target selection for subsequent expression and biochemical characterization. To enable analysis of structural features with potential functional relevance, we define two novel types of *functionally-targeted protein structure networks* (FT-PSNs) generated using functional information specific to this protein class. In particular, sequence region networks (SRNs) are based on connectivity among previously identified functional sequence blocks, while active site networks (ASNs) are based on interactions among chemical moieties comprising the active site residues. Clustering of SRNs reveals several classes with distinct structural characteristics, providing a parsimonious descriptor of protein structure and a predictor of global flexibility. ASNs are used to construct a measure we hypothesize to correlate with active site flexibility and hence enzyme promiscuity. A case-control comparison with a pair of experimentally characterized esterase-lipases (one promiscuous and one specific) suggests that most of the *D. capensis* esterase/lipases have relatively rigid active sites, consistent with their having specific functionalities. This approach is readily adaptable to other incompletely characterized enzyme classes, providing a potentially useful way of selecting experimental targets based on predicted catalytic specificity.

2 Methods

2.1 Clustering, Sequence Alignment and Prediction of Putative Protein Structures

D. capensis proteins were annotated using the MAKER-P (v2.31.8) pipeline[32,33], a BLAST search against SwissProt, and InterProScan[34], as previously described in[31]. The protein set for this study was chosen starting from all sequences identified as having esterase/lipase functionality, followed by elimination of truncated proteins for which one or more of the active site residues were in the missing regions. Clustering of sequences was performed by first aligning sequences using ClustalOmega[35], with settings for gap open penalty = 10.0 and gap extension penalty = 0.05, hydrophilic residues = GPSNDQERK, and BLOSUM weight matrix, and then computing a complete link hierarchical clustering of the resulting dissimilarity scores (one minus the ClustalOmega sequence similarity divided by 100, yielding in values on the [0,1] interval). Clustering and other data analyses were performed using the R statistical computing platform[36]. For pur-

poses of subsequent alignment and comparison, subclusters were then made by defining a cutoff point at a sequence dissimilarity value of 0.7. The presence and position of potential signal sequences flagging the protein for extracellular transport were assessed using the program SignalP 4.1[37], using the following settings: organism group = eukaryotes, D-value cutoff = default (optimized for correlation), and method = input sequences may include transmembrane regions. Structures were predicted from sequences using a three-stage process, following the *in silico maturation* protocol of[38]. First, an initial model was created for each complete sequence using the Robetta implementation of the Rosetta[39,40] package. These structures were modified in the second stage of the process by removing any residues not present in the mature proteins and by correcting protonation states to reflect their predicted cellular or extracellular environments (with protonation states predicted using PROPKA 3.1[41]). In the third phase, each corrected model structure was equilibrated in explicit solvent; simulations were carried out using NAMD[42] with the CHARMM36 forcefield[43] and the TIP3P water model[44] at 293K under periodic boundary conditions. Solvated models were energy-minimized for 10,000 iterations before being simulated for 500ps, with the final configuration being employed in subsequent analyses. This process was performed for the 26 esterase/lipase sequences from *D. capensis* and several reference sequences from other plants. At least one reference sequence was included per subcluster. These proteins were chosen for purposes of sequence annotation: their active sites and functional regions are relatively well annotated in the UniProt database[45], enabling comparisons to the newly characterized sequences. To the best of our knowledge, no structures have yet been solved for a plant esterase/lipase, therefore we also predicted structures for the annotation reference sequences. The PDB files corresponding to the initial and equilibrated structures for all the proteins discussed in this manuscript are available in the Supplementary Information (Supplementary Tables 1 and 2).

2.2 Network Modeling and Analysis

A protein structure network (PSN) was calculated for each protein from its predicted three-dimensional structure using software tools from[38] (which also make use of VMD[46] and the `statnet` library[47,48] for R[36]). Nodes and edges were defined per[49] (see Figure 1A), in which each node represents a chemical group and two nodes are adjacent if they potentially interact (as determined by a distance criterion). Specifically, two nodes i and j are considered adjacent if i contains at least one atom of any type that is within 4.6Å of at least one atom in j, or if i contains at least one carbon that is within 5.4Å of at least one carbon in j. These structures were then secondarily processed to construct functionally targeted PSNs (FT-PSNs) using the `sna` library[50] within `statnet`. A sequence region network (SRN) was constructed from each PSN by identifying all vertices associated with each conserved sequence block or inter-block region (IBR, region between conserved sequence blocks) and defining two regions to be adjacent in the SRN if and only if there were more than five edges between their respective vertex sets in the corresponding PSN (Figure 1B). Each SRN thus encodes the non-trivial interactions among chemical groups within each functionally significant sequence region. Active site networks (ASNs) were also constructed from each PSN as follows. First, all vertices associated with active site residues were identified, as were all vertices adjacent to these vertices within the PSN. The ASN was then defined as the subgraph of the corresponding PSN induced by this combined vertex set. Thus, each ASN represents the local interactions among chemical groups in the active site and the other groups with which they are in contact, irrespective of where these groups reside within the primary sequence.

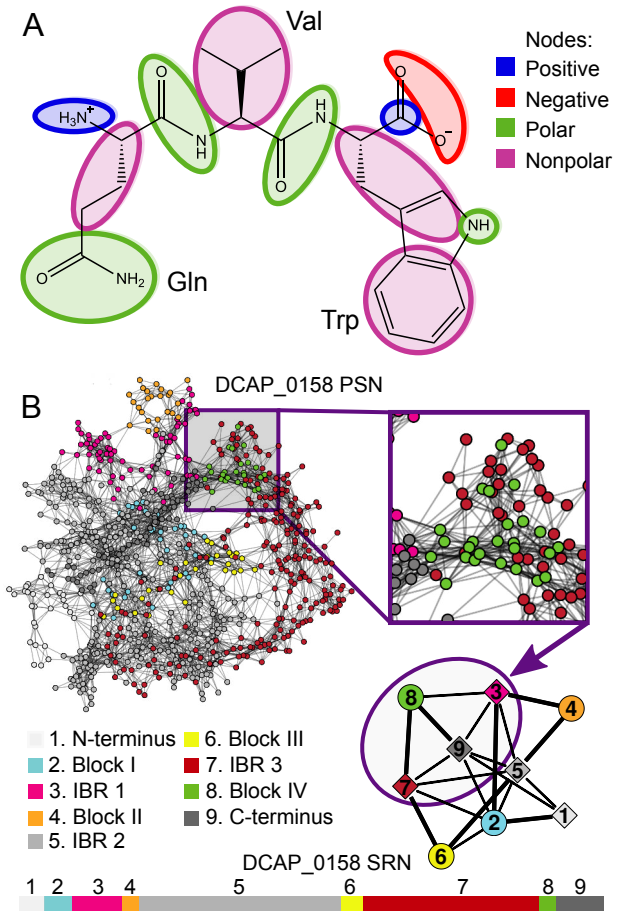

Fig. 1 A. Node definitions for protein structure networks (PSNs). A polypeptide (here illustrated by the tripeptide QVW) is divided into chemical groups using the Benson-Daggett typology (colored ovals), each group becoming a small-moiety node in the PSN. Nodes are adjacent if at least one atom pair is within a critical radius. B. SRNs are formed from PSNs by first grouping all nodes associated with residues in each sequence region, and then defining region pairs to be adjacent if a threshold number of their respective PSN nodes are adjacent (here, > 5). Schematic shows correspondence between local structure involving the Block IV region and its SRN neighbors (IBR1, IBR3, and the C-terminal region). Shaded bar (bottom) shows relative lengths of each sequence region; although longer regions (e.g., IBR3) are often well-connected, short regions (e.g. IBR1) can also be extremely central.

Clustering of SRNs was performed by calculating the Hamming distance between SRNs (i.e., the number of edge changes needed

to convert one SRN into another) and computing a complete link hierarchical clustering solution for the resulting distance matrix (all analyses performed using `statnet` and R). Inspection of the dendrogram (Figure 5A) indicated a four-cluster solution, and central graphs were calculated from the networks in each respective cluster. Block image matrices showing the fraction of SRNs having each respective inter-region edge are shown in Figure 6.

Constraint of active site residues within ASNs was assessed as follows. For each vertex associated with a moiety in the active site, three measures were computed: the *degree,* or number of ties to other vertices; the *triangle degree,* or number of triangles (3-cliques) to which the vertex belongs; and *core number,* or number of the highest degree k-core[51] to which the vertex belongs. Physically, these respectively indicate the total number of contacts associated with the chemical group (potentially impeding its motion), the number of truss-like, triangular structures in which the group is embedded (again, restricting mobility), and the extent of local cohesion around the chemical group (found to distinguish "tighter" and "looser" packing regimes[52]). To summarize the impact of each measure over the active site as a whole, values were averaged across active-site vertices. To obtain an additional constraint measure, the number of paths between each pair of active-site vertices through neighboring (i.e., non-active site) vertices was computed, and the log of the minimum of this value over the set of active site vertex pairs was employed as a measure of *site cohesion.* Intuitively, high values of site cohesion indicate that all active site chemical groups are connected by a large number of indirect contacts, while low values suggest that at least one pair of active site moieties has few local pathways holding them together. These four indices (mean active site degree, mean active site triangle degree, mean active site core number, and site cohesion) were used to produce an omnibus index of *site constraint* via principal component analysis (PCA) of the standardized network measures. The PCA solution revealed one primary dimension, with the first principal component accounting for 75% of the total variance among the four measures (ratio of first eigenvalue to second greater than 5), and the scores on this first component scores were hence employed for subsequent analysis as the constraint index.

3 Results and Discussion

3.1 *D. capensis* Esterase/Lipases Cluster Into Distinct Subfamilies Based on Sequence Features

All enzymes from the *D. capensis* genome previously annotated as functional esterase/lipases were clustered by sequence similarity (Figure 2). Several annotation reference sequences from other plants were also included to facilitate identification of the active site residues and functional sequence blocks. The reference sequences (referred to by their UniProt IDs) are from the plants *Carica papaya* (GDL1_CARPA) and *Arabidopsis thaliana* (GLIP6_ARATH, GDL7_ARATH, EXL3_ARATH, APG2_ARATH). Although the active site residues and functional sequence blocks are readily found, plant esterase/lipases are relatively poorly characterized; these reference sequences lack high-resolution structures and in most cases detailed functional information, e.g. experimental data about their substrate preferences. One of the objectives of this work is to provide a starting point for approaching such studies in undercharacterized enzyme classes such as this one.

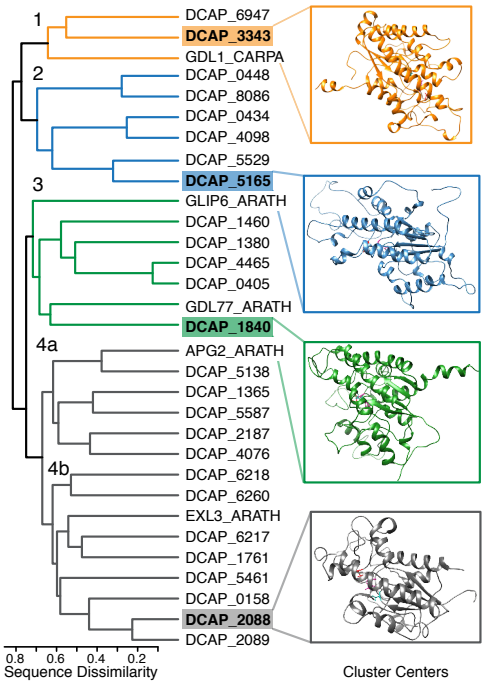

Fig. 2 Protein sequence clustering of esterase/lipase sequences from the *D. capensis* genome, denoted by DCAP, and annotation reference sequences from other plants, which are identified by their UniProt IDs: *Carica papaya* (GDL1_CARPA) and *Arabidopsis thaliana* (GLIP6_ARATH, GDL7_ARATH, EXL3_ARATH, APG2_ARATH). Information about these annotation reference sequences found in UniProt enabled identification of functional sequence features in the novel *D. capensis* proteins via sequence alignment and comparison. Annotation details are shown in Supplementary Figures S1-S5.

In all the sequences examined here, the active site residues are consistent with the catalytic triad of a serine hydrolase, and the functional sequence blocks characterizing the GDSL esterase/lipase family are readily identified by comparison to the work of Akoh et al.[3] and Vujaklija et al.[16]. In most cases, SignalP 4.1 predicts the presence of a signal peptide sequence tagging these esterase/lipases for extracellular secretion. Annotated protein sequence alignments showing functional sequence features can be found in Supplementary Figures S1-S5. The sequence alignments are color-coded to indicate both individual amino acid properties and important sequence regions. Sequence-based clustering yields four major groups with greater than 30% sequence identity among all members. As previously observed for this protein class, each group has significant diversity among its component sequences; only one pair in this set (DCAP_0405 and DCAP_4465) has more than 80% sequence identity. For each cluster, the central sequence (the protein having the minimum average distance in sequence space from all the others) is highlighted. Comparative models for these central sequences are shown to the right of the cluster figure, revealing variations on a common struc-

tural theme.

Cluster 1 contains sequences that have the canonical GDSL motif, as found in the reference sequence GDL1_CARPA, which was isolated from papaya latex[53] and has been proposed as a "naturally immobilized" biocatalyst for performing regioselective esterification and transesterification reactions[54]. The enzymes in cluster 2 instead have GDSN in the first functional block. Clusters 3 and 4 contain the motif GDSX, where X is usually a hydrophobic residue, but is Ser or Thr in some cases. Overall, the presence of the three active site residues in 24 of the 25 *D. capensis* esterase/lipases suggests they are functionally active enzymes.

3.2 Conserved Active Site Residues Suggest Functional Enzymes

In general, esterase/lipases are characterized by four moderately conserved sequence blocks of length 8-13 residues that contain the cataytic triad, the oxyanion hole proton donors, and other functionally relevant residues[55]. These blocks are always found in the same order in sequence space, though the lengths of the intervening sequences can vary substantially[19]. Functional sequence blocks I-IV are highlighted in the sequence alignments (Supplementary Figures S1-S5.) In Figure 3A, these functional blocks are represented as sequence logos, where the size of each residue label correlates with the number of instances at that sequence position within each cluster. The Ser-Asp-His catalytic triad is located within two block regions: block I (Ser) and block IV (Asp-His). The remaining two blocks contain conserved oxyanion hole residues, Gly in block II and Asn in block III[3]. Most of the proteins in this set contain the expected functional residues, as exemplified by the reference sequences GDL1_CARPA, GLIP6_ARATH, and GDL7_ARATH, as well as the functionally characterized GDSL esterase/lipase G1DEX3_SOLLC from the tomato.

Some variation is observed in the oxyanion hole residues: the stabilizing Asn residue in block III is replaced by Ile in DCAP_0434, Ser in APG2_ARATH and DCAP_ 5138, and Asp in EXL3_ARATH. These substitutions are consistent with almost all of the *D. capensis* enzymes following the canonical GDSL mechanism[56]. The two exceptions in this set are DCAP_2088, which is missing the entirety of block III, and DCAP_6260, which has substitutions to the two active site residues located in block IV (Asp to Leu and His to Ser, see Figure S4). DCAP_6260 is the only protein in this set that does not contain all three active site residues, although it retains the canonical GDSX motif in block I and the stabilizing oxyanion residues in block II and III. The potentially catalytically inactive sequences (DCAP_6260 and DCAP_2088) were included because they do contain most of the relevant sequence and structural features; we hypothesize that these proteins may play a binding rather than catalytic role. Alternatively, they may represent pseudogenes. DCAP_4076 has a C-terminal extension not found in the other esterase/lipases, the role of which is currently not known, although it has moderate sequence similarity to transcriptional regulation proteins in *Arabidopsis thaliana* and soybean (Supplementary Figure S8A.).

3.3 Molecular Modeling

The structure of a typical GDSL esterase/lipase has a 4-stranded parallel β-sheet with six α-helices arranged around it (shown for a representative example in Figure 3B). Due to the lack of solved structures for plant esterase/lipases, comparative modeling was used rather than traditional homology modeling. To make a standard homology model, the sequence of interest is threaded onto the known structure of a closely related protein, followed by energy minimization. In comparative modeling, the procedure is similar except that the protein is modeled piecewise using multiple template structures selected by the software (in this case Rosetta[39]) from the Protein Data Bank, followed by global minimization using a simplified force field. This methodology is regularly validated via CAMEO[57], and is the basis of well-known structure prediction systems such as Rosetta (used here) and I-TASSER[58]. All template structures used for a representative example (DCAP_0434) are tabulated in Supplementary Table 3 and the parent structures for each model can be found in the headers for their respective .pdb files (available for download in the SI.)

We used the initial models generated by the Robetta server[40] as a starting point; however as these structures are not calculated in an aqueous environment and do not account for protonation states, we modifed them to produce models that are more representative of the mature enzyme (available for download in the SI.) Signal sequences were removed and protonation states were corrected consistent with their expected functional environments. These structures were then subjected to molecular dynamics simulation in explicit solvent to generate the equilibrated structures (illustrated in Supplementary Figure S6). The equilibrated molecular models of these proteins show that although they all have the expected overall fold, substantial diversity exists in the placement of secondary structure elements, as well as the lengths of the linker regions (Supplementary Figures S7 and S8B). All three active site residues are accessible, in contrast to lipases, where only the serine is exposed due to the hydrophobic "lid" that is characteristic of that enzyme class. The positioning of the catalytic triad residues, which is consistent with catalytic competence, is shown in Figure 3C. The active site residues are located in loop regions, with the occasional exception of the Ser, which is part of an α-helix in some esterase/lipases (e.g. in Cluster 1). The conserved oxyanion hole residues in Block II reside in a loop region, while half of the Block III residues lie in a β-sheet and the other half in an α-helix. This mixture of structural motifs presents a challenge for coarse-grained network analysis, where a common approach is to break up the protein into discrete regions based on secondary structure. In the case of the plant esterase/lipases of this set, that classification does not align with the functional regions identified in previous studies of esterase/lipases; we have therefore used the functional sequence blocks, termini, and inter-block regions rather than secondary structure elements as the basis for constructing the FT-PSN representation of the overall enzyme folds.

3.4 Protein Structure Networks

Contacts between structural regions of the esterase/lipases were analyzed using a network formalism; for each protein, full PSNs

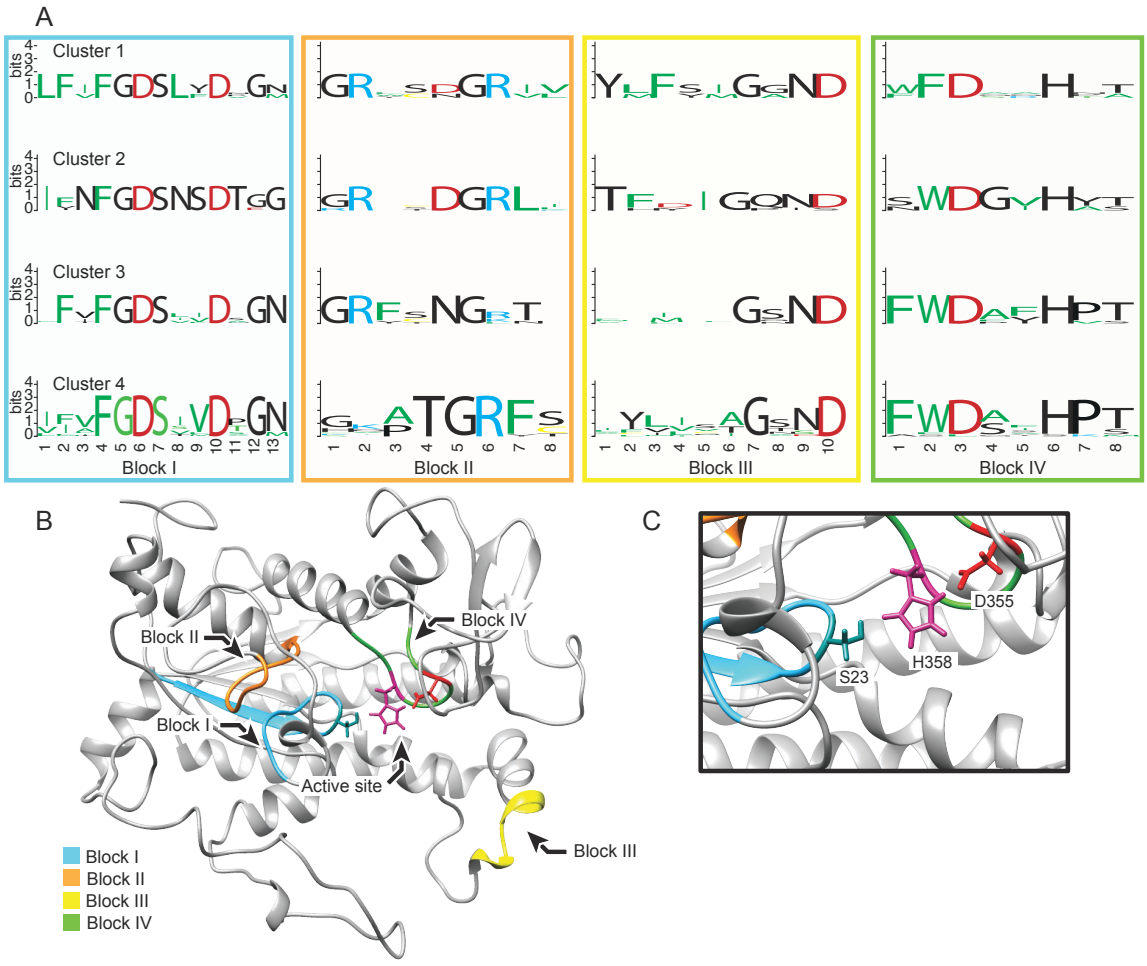

Fig. 3 A. The sequences of the four functional blocks (inside the colored frames) are presented by sequence cluster (arranged from top to bottom as in Figure 2). The sizes of the residue labels correlate with the fraction of sequences in the cluster having that residue in the indicated position. Amino acid properties are color coded as follows: hydrophobic-green, positive-blue, negative-red, cysteine-yellow, other-black. B. A representative molecular model of a *D. capensis* esterase/lipase (DCAP_0434) with the four functional blocks highlighted using the color-coding of the frames in Panel A. C. Expanded view of the active site catalytic triad for a typical esterase/lipase (DCAP_0434), showing that the active site residues are positioned in a manner consistent with catalytic activity.

and two novel types of FT-PSNs were generated. First, full PSNs were calculated for the esterase/lipase molecular models based on the formalism of Benson and Daggett[49], where each amino acid is composed of nodes defined by chemical functionality. Two illustrative visualizations of PSNs from different sequence clusters are shown in Figure 4. Although we refer to the functional blocks themselves by Roman numerals I-IV as defined in the earlier literature for the sake of comparison to prior work, for purposes of generating FT-PSNs we define nine sequence regions comprising the four functional blocks as well as the regions between them (inter-block regions, or IBRs), and the N- and C-termini. These sequence regions are numbered 1-9 in order from the N-terminus to the C-terminus for each protein. In these PSN examples, nodes (chemical moieties) belonging to the termini, functional blocks, and inter-block regions are color coded as indicated in the legend. This representation allows rapid examination of the degree of connectivity between different sequence regions, e.g. it can easily be seen that the nodes of Block II (orange) are more connected to each other in DCAP_0158 (4A) than in DCAP_1380 (4B), while many Block III nodes are connected to those from other sequence regions in both proteins. Although this representation provides a visualization of connectivity between different parts of the protein separate from the three-dimensional structure, the number of nodes and the complexity of the plots makes comparison difficult. Therefore, we define two types of specialized FT-PSNs based on functionally relevant sequence features of these proteins.

In order to further simplify the graph representations, a block model[51] was constructed for each protein by condensing all nodes within each of these sequence regions to form a coarse-grained FT-PSN whose edges represent contacts between moieties in each pair of sequence regions (each region constituting a node within the block model). These *sequence region networks* (SRNs), provide a direct representation for the structure of contacts among functionally significant components of the protein, which we hypothesize to be related to overall function. To iden-

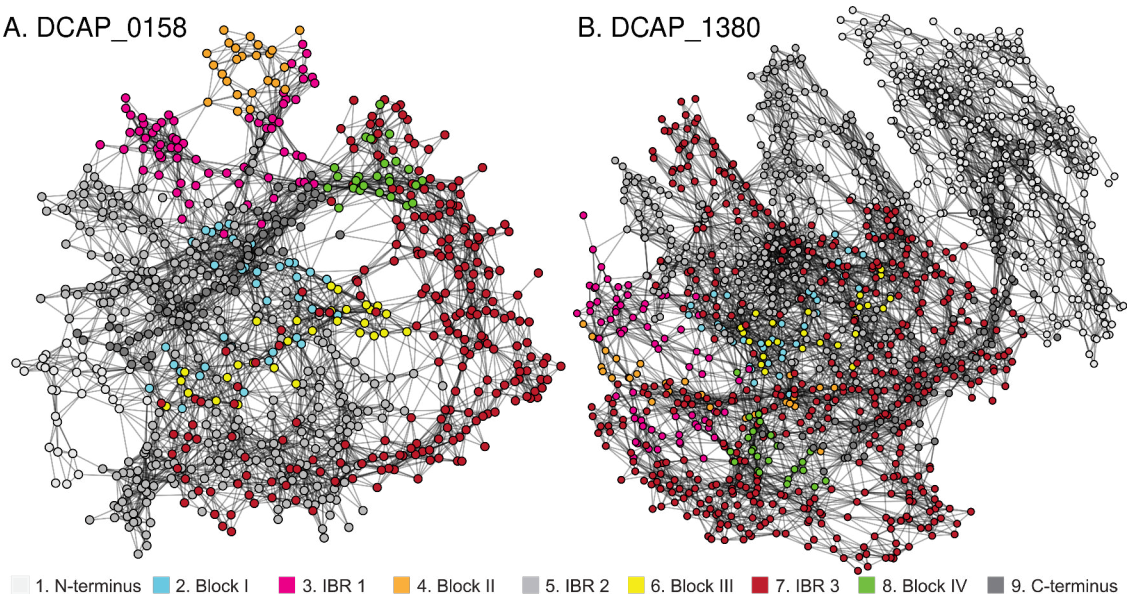

Fig. 4 Protein structure networks of DCAP_0158 (Cluster 4a) and DCAP_1380 (Cluster 3). Each node (closed circles) represents a chemical moiety and is color coded based on its respective sequence position in a functional block, terminus or IBR. Ties (gray lines) indicate physical interactions between a set of nodes. The positioning of the nodes in this representation is optimized to show topology and does not directly correspond to three-dimensional space; proximity within the cutoff distance is solely indicated by the ties.

tify distinct classes of functionally relevant structure within the *D. capensis* esterase/lipase set, we then performed a hierarchical clustering of SRNs by Hamming distance (i.e. the number of adjacency differences among sequence regions between two respective SRNs). Figure 5A shows the dendrogram for the clustered SRNs, along with structural models for the protein structure corresponding to the central graph for each SRN cluster. The central graphs themselves are shown in Figure 5B. Following clustering of SRNs by Hamming distance, clusters were summarized by forming block image matrices[51]. Within each matrix, the i, j cell value corresponds to the fraction of cluster members whose SRN contains an edge between sequence region i and sequence region j. Schematic representations for each cluster, illustrating how the adjacency matrices for these models are constructed, are shown in Figure 6. In addition to showing distinct structural patterns across clusters, Figure 6 shows a fairly high level of consensus within clusters (with most cells having densities close to either 0 or 1). For this reason, we summarize the SRNs within each cluster by their central graph, which is equivalent to dichotomizing the image matrices at 0.5; these networks are shown in Figure 5B.

Clustering of the SRNs reveals important differences among esterase/lipases that are not apparent from the sequence clusters, as well as some common features of potential structural and functional significance. For example, the IBR between Blocks II and III (node 5) is highly central across all structures, being in direct contact with a large number of other sequence regions and frequently bridging regions not otherwise in contact. This suggests a key structural role for this highly variable (i.e. non-conserved) sequence region that may have been overlooked by purely sequence-based analyses. Likewise, Block III has identical neighbors in all clusters, being tied only to its sequence-space neighbors and to Block I (node 2). This highly conserved pattern of both interaction and *non*-interaction is suggestive of functional significance. By contrast, the other interaction partners of Block I vary considerably across clusters, as do e.g. the partners of IBR 1 (node 3). Such variation in interaction among conserved sequence blocks may be indicative of corresponding differences in functional characteristics.

Interestingly, clustering by structural similarity of SRNs yields a pattern that is distinct from clustering by sequence (Figure 2). Although sequence homology is often a good indicator of broad functional similarity at the level of protein classes, structural comparison provides a much more precise tool for functional differentiation among related proteins. As with previous applications of structure networks to study allostery, binding, inter/intramolecular interactions, and other phenomena otherwise difficult to ascertain using only sequence analysis[49,52,59], SRNs such as those introduced here have the potential to complement sequence analytic methods for purposes such as functional prediction and target selection.

The coarse-grained network representations described above provide a useful basis for comparison of overall structural properties among esterase/lipases, but they do not directly address the flexibility and accessibility of the active site itself, which is a potential indicator of enzyme specificity[60]. Most of what is known about the esterase/lipase family to date comes from the microbial esterase/lipases, which are generally regarded as promiscuous enzymes. It has been suggested that this property may generalize to plant esterase/lipases, which have so far not been extensively characterized. However, as discussed above, many plants have numerous esterase/lipase paralogs, possibly indicating that the same diversity of activity is accomplished using multiple enzymes, each with its own functionality, rather than fewer multifunctional enzymes.

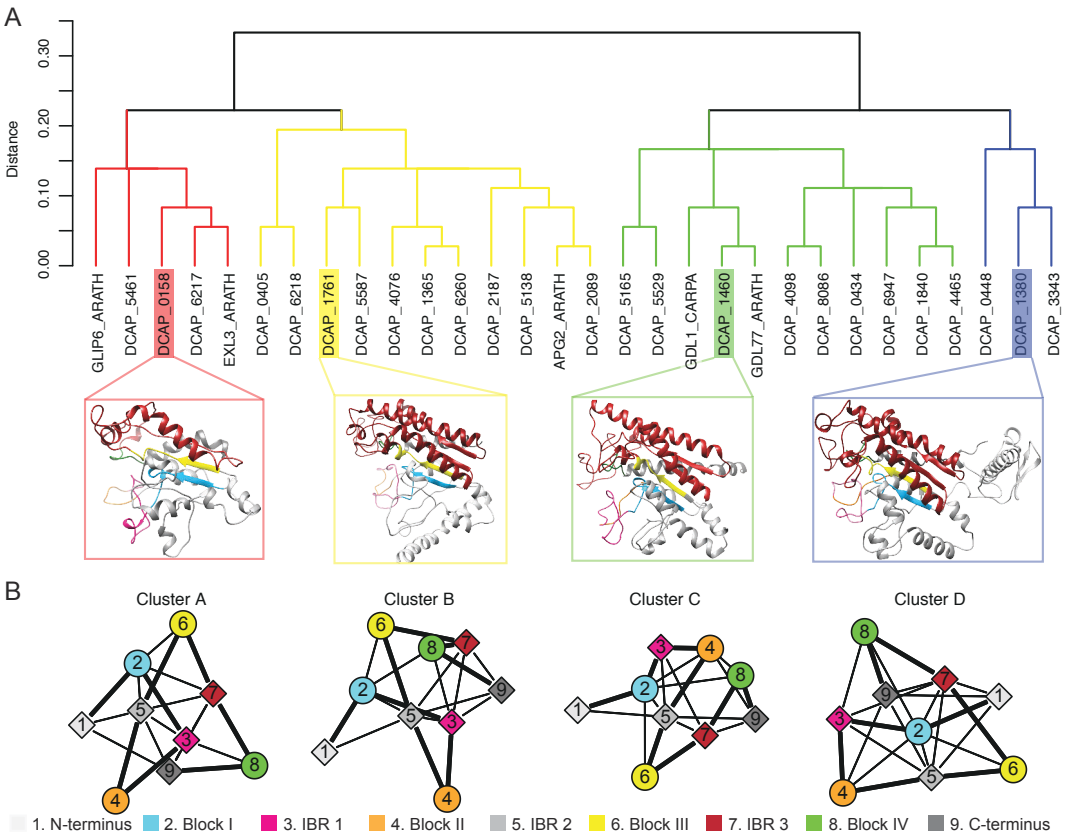

Fig. 5 A. Clustering of sequence region networks (SRNs) for modeled esterase/lipase structures from the *D. capensis* genome and reference sequences from other plants. Inset structures depict the most central member of each cluster. B. Central graphs for the SRNs in each cluster. Colors for nodes corresponding to conserved (circular) and non-conserved (diamond-shaped) sequence regions correspond to residue colors in panel A; thick lines indicate connections along the protein backbone.

Because enzyme promiscuity is strongly correlated with active site flexibility[60], we used a similar analysis of network structure to investigate the ties among nodes in the active site regions of the *D. capensis* esterase/lipases. As before, we began by constructing moiety-level PSNs using the Benson-Daggett representation. We then formed *active site networks* (ASNs) by taking the subgraph of each PSN induced by the nodes corresponding to active site moieties together with the union of their respective network neighborhoods. Each ASN thus represents the pattern of connectivity among moieties topologically local to the active site. Structural constraints on the active site were measured using several common network properties: mean degree (the average number of ties each node has to other nodes), mean triangle degree (the number of memberships in 3-cliques or triangles), mean k-core number (where the kth core of a graph is the maximum set of nodes such that every member of the set is adjacent to at least k other nodes), and inter-node connectivity (counts of paths connecting active-site nodes via other nodes in the ASN). These properties were computed for all nodes corresponding to active site moieties, and are plotted in Supplementary Figure S9. They were then composited by taking their first principal component, yielding a single measure of active site constraint for each network.

Figure 7 shows the active site constraint measure for each enzyme in our set, as well as two enyzmes for which more detailed activity data is available. The latter two, well-characterized enzymes were selected as a "case/control" validation for the functional significance of the constraint measure: the tomato cutinase (G1DEX3_SOLLC), which is known to catalyze a specific reaction (high-specificity "case"); and *E. coli* TesA, (TESA_ECOLI), which is known to accept a variety of substrates (low-specificity "control"). Consistent with the hypothesis that the large number of esterase/lipases in typical plant genomes corresponds with a higher level of substrate specificity, we observe only two plant enzymes with a level of constraint lower than the promiscuous TesA (red-shaded area); of the remainder, roughly half showed constraint levels between TesA and tomato cutinase (yellow-shaded area) and half showed higher constraint levels (blue-shaded area). Our analysis suggests that the majority of esterase/lipases in *D. capensis* are likely to be highly specific, with the prominent exception of DCAP_3343. This enzyme, and GDL77_ARATH from *Arabidopsis*, show extremely low levels of active site constraint implying a very high level of local flexibility. We hypothesize that these enzymes will accept a wider range of substrates than the others examined here, and that they occupy a distinct functional role (perhaps more similar to the role of microbial esterase/lipases).

Figure 8 shows structural models of the *D. capensis* esterase/lipases with the least (red) and most (blue) constrained active sites, as determined by the ASN flexibility metric plotted in

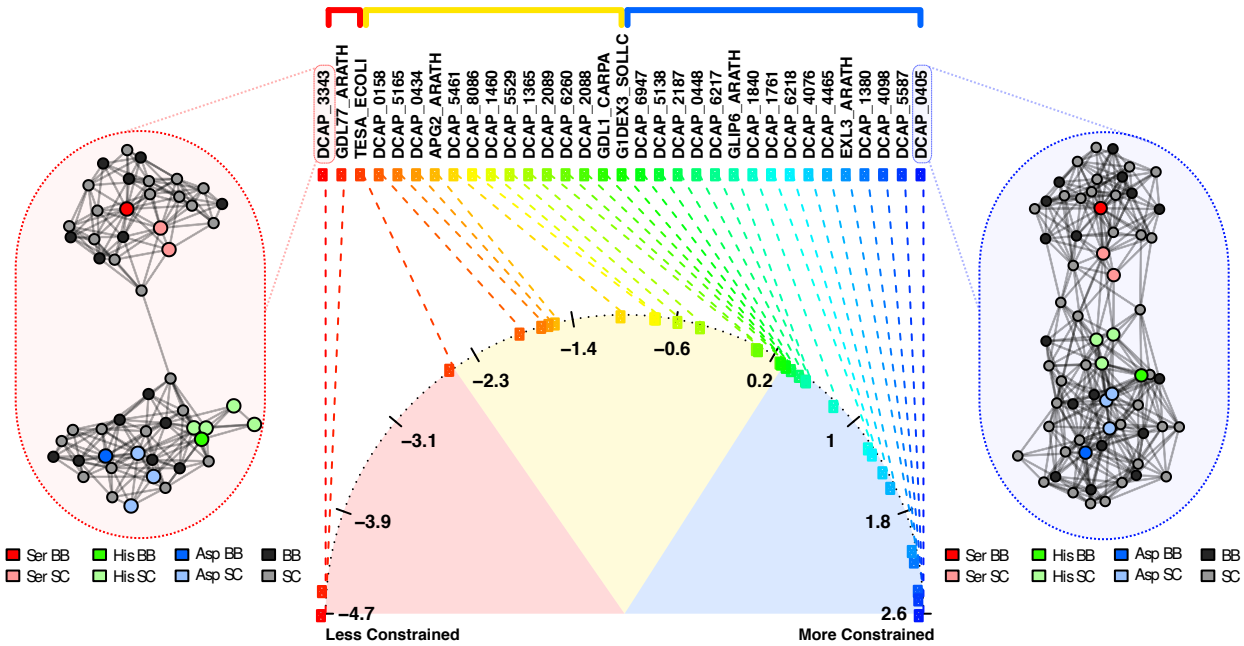

Fig. 7 Main panel: constraint level of active site moieties within protein structure networks. Red-shaded region indicates lower constraint levels than the bacterial enzyme TesA; yellow and blue shaded regions respectively indicate levels of constraint between TesA and tomato cutinase and levels of constraint greater than tomato cutinase. Nearly all plant enzymes studied here show more active site constraint than TesA, with tomato cutinase falling near the median of these. Side panels: ASN visualizations for DCAP_3343 (left) and DCAP_0405 (right) show respective examples of low and high levels of active site constraint. Nodes correspond to moieties, with backbone (BB) and side chain (SC) moieties for the three active site residues indicated by color. Highly cohesive ASNs imply numerous constraints on the motion of active site residues, potentially leading to higher levels of substrate specificity.

Figure 7. Somewhat counterintuitively, the protein with the less flexible active site (DCAP_3343) has a better-defined secondary structure. Based on the DSSP secondary structure definitions[61], DCAP_0405 has 29.3 % α-helix, 2.9 % β-strand, and 67.8% turn/coil, while (DCAP_3343) has 43.6% α-helix, 5.3% β-strand, and 51.1% turn/coil. Although DCAP_3343 has more α-helical and β-strand secondary structure elements, the structure around the active site itself is looser and less densely connected than that of DCAP_0405, where loops and random coil regions interact to hold the active site residues more rigidly in place. Although unstructured regions are often regarded as highly flexible regions, this depends on their context in the overall structure; recent NMR dynamics measurements and MD simulations reveal that loops undergo dynamics over a wide range of timescales[62] and their motions are frequently involved in allosteric regulation[63]. Longer loops, which are more able to become mutually entangled with other structural elements are more likely to be rigid[64], which is consistent with the predicted structure of DCAP_0405.

4 Conclusion

In summary, molecular modeling and protein structure network analysis of 26 esterase/lipases identified from the genomic DNA of *Drosera capensis* suggest that—with the exception of one protein, DCAP_3343—the active site regions of these enzymes are less flexible than those of related microbial proteins. We hypothesize that these enzymes act (like tomato cutinase) to catalyze specific reactions, with the outlying protein behaving more like microbial esterase/lipases. Two new types of protein structure networks, seqence region networks (SRNs) and active site networks (ASNs) were defined in order to characterize overall protein flexibility and that of the active sites. Principal component analysis of active site constraint measures generated from PSNs enabled us to sort the esterase/lipases from decreasing to increasing active site rigidity; case/control validation using a pair of well-characterized enzymes suggests that our index is related to substrate specificity. Clustering by SRN shows structural differences between enzymes with respect to functionally significant sequence blocks, as well as an apparently conserved structural role for a highly sequence-variable and previously unnoted inter-block region. These results may serve to guide target selection for subsequent structural or functional studies, and the analytical strategy employed may be fruitfully adapted to other protein classes.

Conflicts of interest

There are no conflicts to declare.

Acknowledgments

This research was supported by NSF awards DMS-1361425 to C.T.B and R.W.M and ARO award W911NF-14-1-0552 to C.T.B. V.T.D was supported by the Mathematical, Computational and Systems Biology Predoctoral Training Grant T32 EB009418-08. S.H.K. and R.W.M. acknowledge the California State Summer School for Math & Science (COSMOS) and NSF grant CHE-1308231 to R.W.M. This work was also made possible, in part,

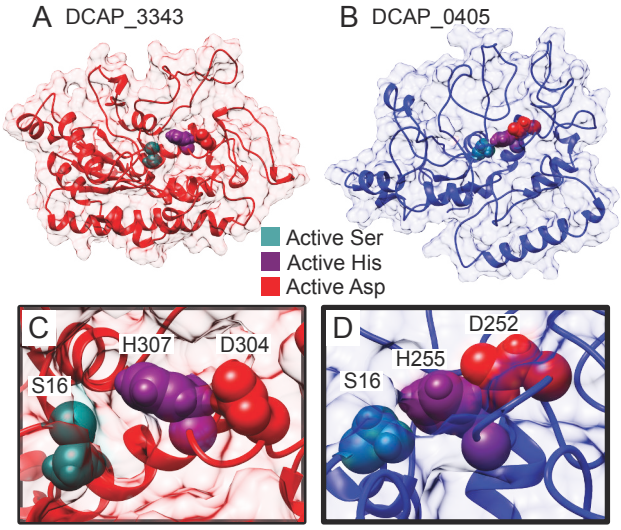

Fig. 8 Structural models of the least and most constrained enzymes based on the ASN analysis shown in Fig 7. A. Surface and ribbon representations of DCAP_3343, which is the only *D. capensis* esterase/lipase with a less constrained active site than that of TesA from *E. coli*. B. Surface and ribbons for DCAP_0405, the most constrained enzyme in this set. C. and D. Expanded views of the active ites of these enzymes show the differences in active site constraint, which are not obvious from examination of the overall structural model. The active site residue side-chains of DCAP_3343 (C) are oriented out and away from each other, while those of DCAP_0405 are tightly held in a closely packed conformation.

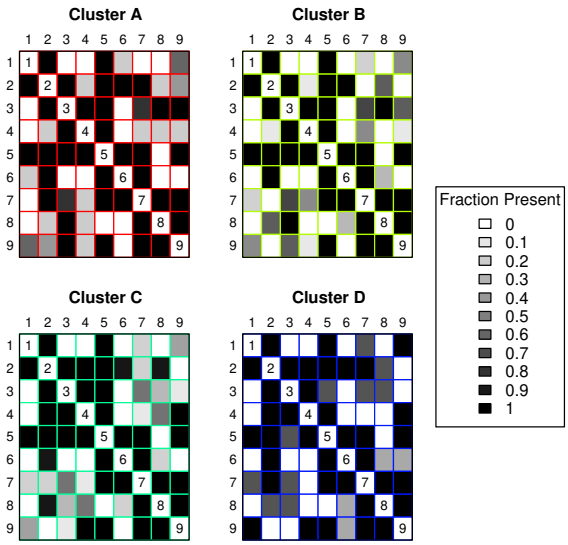

Fig. 6 Block image matrices for the clustered sequence region networks. The i, j cell value for each matrix indicates the fraction of cluster members whose SRN contains a tie from region i to region j. Node numbers correspond to sequence regions numbered from the N-to the C-terminus as defined in the text.

through access to the Genomic High Throughput Facility Shared Resource of the Cancer Center Support Grant (CA-62203) at the University of California, Irvine and NIH shared instrumentation grants 1S10RR025496-01 and 1S10OD010794-01.

Author Contributions

R.W.M., V.T.D., M.H.U., and J.E.K. chose the protein set, determined the functional regions of interest, generated the predicted structures, and analyzed sequence and structure data. C.T.B. performed the cluster analysis, molecular dynamics simulations, and network visualization and analysis. V.T.D., M.H.U., J.E.K., S.H.K and R.W.M performed sequence annotation and comparisons. R.W.M. and C.T.B. designed the study. V.T.D., M.H.U., S.H.K, C.T.B. and R.W.M. wrote the manuscript.

Notes and references

1 O. Serra, S. Chatterjee, W. Huang and R. E. Stark, *Plant Science*, 2012, **195**, 120–124.
2 S. Chatterjee, A. J. Matas, T. Isaacson, C. Kehlet, J. K. Rose and R. E. Stark, *Biomacromolecules*, 2016, **17**, 215–224.
3 C. C. Akoh, G.-C. Lee, Y.-C. Liaw, T.-H. Huang and J.-F. Shaw, *Progress in Lipid Research*, 2004, **43**, 534–552.
4 A.-L. Girard, F. Mounet, M. Lemaire-Chamley, C. Gaillard, K. Elmorjani, J. Vivancos, J.-L. Runavot, B. Quemener, J. Petit, V. Germain, C. Rothan, D. Marion and B. Bakana, *Plant Cell*, 2012, **24**, 3119–3134.
5 H. Ebata, K. Toshima and S. Matsumura, *Macromolecular Bioscience*, 2007, **7**, 798–803.
6 A. Mahapatro, A. Kumar and R. Gross, *Biomacromolecules*, 2004, **5**, 62–68.
7 S. Kobayashi and A. Makino, *Chemical Reviews*, 2009, **109**, 5288–5353.
8 C. Vilela, A. F. Sousa, A. C. Fonseca, A. C. Serra, J. F. Coelho, C. S. R. Freire and A. J. Silvestre, *Polymer Chemistry*, 2014, **5**, 3119–3141.
9 Y.-C. Lo, S.-C. Lin, J.-F. Shaw and Y.-C. Liaw, *Journal of Molecular Biology*, 2003, **330**, 539–551.
10 I. Mathews, M. Soltis, M. Saldajeno, G. Ganshaw, R. Sala, W. Weyler, M. A. Cervin, G. Whited and R. Bott, *Biochemistry*, 2007, **46**, 8969–8979.
11 Y.-L. Lee, J. C. Chen and J.-F. Shaw, *Biochemical and Biophysical Research Communications*, 1997, **231**, 452–456.

12 R. Sharma, Y. Chisti and U. C. Banerjee, *Biotechnology Advances*, 2001, **19**, 627–662.

13 K. Clauß, A. Baumert, M. Nimtz, C. Milkowski and D. Strack, *The Plant Journal*, 2008, **53**, 802–813.

14 Y. Kikuta, H. Ueda, M. Takahashi, T. Mitsumori, G. Yamada, K. Sakamori, K. Takeda, S. Furutani, K. Nakayama and Y. Katsuda, *The Plant Journal*, 2012, **71**, 183–193.

15 H. Chepyshko, C.-P. Lai, L.-M. Huang, J.-H. Liu and J.-F. Shaw, *BMC Genomics*, 2012, **13**, 309.

16 I. Vujaklija, A. Bielen, T. Paradžik, S. Bidin, P. Goldstein and D. Vujaklija, *BMC Bioinformatics*, 2016, **17**, 91.

17 M. Volokita, T. Rosilio-Brami, N. Rivkin and M. Zik, *Molecular Biology and Evolution*, 2011, **28**, 551–565.

18 D. Cao, H. Cheng, W. Wu, H. Soo and J. Peng, *Plant Physiology*, 2006, **142**, 509–525.

19 Dong, Xiangshu and Yi, Hankuil and Han, Ching Tack and Nou, Ill Sup and Hur, Yoonkang, *Molecular Genetics and Genomics*, 2016, **291**, 531–542.

20 M. Naranjo, J. Forment, M. Roldan, R. Serrano and O. Vicente, *Plant, Cell & Environment*, 2006, **29**, 1890–1900.

21 J. Hong, H. Choi, I. Hwang, D. Kim, N. Kim, d. Choi, Y. Kim and B. Hwang, *Planta*, 2008, **227**, 539–558.

22 C. P. Lai, L. M. Huang, L. F. O. Chen, M. T. Chan and J. F. Shaw, *Plant Molecular Biology*, 2017, **95**, 181–197.

23 D. Panikashvili, J. Shi, S. Bocobza, R. Franke, L. Schreiber and A. Aharoni, *Molecular Plant*, 2010, **3**, 563–575.

24 K. Takahashi, T. Shimada, M. Kondo, A. Tamai, M. Mori, M. Nishimura and I. Hara-Nishimura, *Plant Cell Physiology*, 2010, **5**, 123–131.

25 B. Zhang, L. Zhang, F. Li, D. Zhang, X. Liu, H. Wang, Z. Xu, C. Chu and Y. Zhou, *Nature Plants*, 2017, **3**, 17017.

26 Y. Zhang, B. Bai, M. Lee, Y. Alfiko, A. Suwanto and G. H. Yue, *Scientific Reports*, 2018, **8**, 11406.

27 B. Bakan and D. Marion, *Plants*, 2017, **6**, doi:10.3390/plants6040057.

28 T. Isaacson, D. K. Kosma, A. J. Matas, G. J. Buda, Y. He, B. Yu, A. Pravitasari, J. D. Batteas, R. E. Stark, M. A. Jenks and J. K. C. Rose, *The Plant Journal*, 2009, **60**, 363–377.

29 T. H. Yeats, L. B. B. Martin, H. M. Viart, T. Isaacson, Y. He, L. Zhao, A. J. Matas, G. J. Buda, D. S. Domozych, M. H. Clausen and J. K. C. Rose, *Nature Chemical Biology*, 2012, **8**, 609–611.

30 M. Pollard, B. F., Y. Li and J. Ohlrogge, *Trends in Plant Science*, 2008, **13**, 236–246.

31 C. T. Butts, J. C. Bierma and R. W. Martin, *Proteins: Structure, Function, and Bioinformatics*, 2016, **84**, 1517–1533.

32 M. Campbell, M. Law, C. Holt, J. Stein, G. Moghe, D. Hufnagel, J. Lei, R. Achawanantakun, D. Jiao, C. J. Lawrence, D. Ware, S. H. Shiu, K. L. Childs, Y. Sun, N. Jiang, and M. Yandell, *Plant Physiology*, 2013, **164**, 513–524.

33 M. S. Campbell, C. Holt, B. Moore and M. Yandell, *Current Protocols in Bioinformatics*, 2014, **48**, 4.11.1–4.11.39.

34 E. Quevillon, V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler and R. Lopez, *Nucleic Acids Research*, 2005, **33**, W116–W120.

35 F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J. D. Thompson and D. G. Higgins, *Molecular Systems Biology*, 2011, **7**, 539–539.

36 R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2018.

37 T. Petersen, S. Brunak, G. von Heijne and H. Henrik Nielsen, *Nature Methods*, 2011, **8**, 785–786.

38 C. T. Butts, X. Zhang, J. E. Kelly, K. W. Roskamp, M. H. Unhelkar, J. A. Freites, S. Tahir and R. W. Martin, *Computational and Structural Biotechnology Journal*, 2016, **14**, 271–282.

39 D. Kim, D. Chivian and D. Baker, *Nucleic Acids Research*, 2004, **32**, W526–31.

40 S. Raman, R. Vernon, J. Thompson, M. Tyka, R. Sadreyev, J. Pei, D. Kim, E. Kellogg, F. DiMaio, O. Lange, L. Kinch, W. Sheffler, B.-H. Kim, R. Das, N. V. Grishin and D. Baker, *Proteins*, 2009, **77**, 89–99.

41 M. H. Olsson, C. R. Sondergaard, M. Rostkowski and J. H. Jensen, *Journal of Chemical Theory and Computation*, 2011, **7**, 525–537.

42 J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé and K. Schulten, *Journal of Computational Chemistry*, 2005, **26**, 1781–1802.

43 R. B. Best, X. Zhu, J. Shim, P. E. M. Lopes, J. Mittal, M. Feig and A. D. Mackerell, Jr, *J Chem Theory Comput*, 2012, **8**, 3257–3273.

44 W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *The Journal of Chemical Physics*, 1983, **79**, 926–935.

45 The UniProt Consortium, *Nucleic Acids Research*, 2017, **45**, D158–D169.

46 W. Humphrey, A. Dalke and K. Schulten, *Journal of Molecular Graphics*, 1996, **14**, 33–38, 27–28.

47 M. S. Handcock, D. R. Hunter, C. T. Butts, S. M. Goodreau and M. Morris, *Journal of Statistical Software*, 2008, **24**, 1–11.

48 C. T. Butts, *Journal of Statistical Software*, 2008, **24**, 1–36.

49 N. C. Benson and V. Daggett, *Journal of Bioinformatics and Computational Biology*, 2012, **10**, 1250008.

50 C. T. Butts, *Journal of Statistical Software*, 2008, **24**, 1–51.

51 S. Wasserman and K. Faust, *Social network analysis: Methods and applications*, Cambridge University Press, 1994, vol. 8.

52 M. H. Unhelkar, V. T. Duong, K. N. Enendu, J. E. Kelly, S. Tahir, C. T. Butts and R. W. Martin, *Biochimica et Biophysica Acta*, 2017, **1861**, 636–643.

53 S. Abdelkafi, H. Ogata, N. Barouh, B. Fouquet, R. Lebrun, M. Pina, F. Scheirlinckx, P. Villeneuve and F. Carrière, *Biochimica et Biophysica Acta*, 2009, **1791**, 1048–1056.

54 A. El Moussaoui, M. Nijs, R. Paul, C.and Wintjens, J. Vincentelli, M. Azarkan and Y. Looze, *Cellular and Molecular Life Sciences*, 2001, **58**, 556–570.

55 C. Upton and J. Buckley, *Trends in Biochemical Science*, 1995, **20**, 178–179.

56 A. Rauwerdink and R. J. Kazlauskas, *ACS Catalysis*, 2015, **5**, 6153–6176.

57 J. Haas, S. Roth, K. Arnold, F. Kiefer, T. Schmidt, L. Bordoli and T. Schwede, *The Protein Model Portal - a comprehensive resource for protein structure and model information*, Database (PMID: 23624946), 2013.

58 Y. Zhang, *BMC Bioinformatics*, 2008, **9**, 40.

59 A. Sethi, J. Eargle, A. A. Black and Z. Luthey-Schulten, *Proceedings of the National Academy of Sciences of the United States of America*, 2009, **106**, 6620–6625.

60 A. Babtie, N. Tokuriki and F. Hollfelder, *Current Opinion in Chemical Biology*, 2010, **14**, 200–207.

61 W. Kabsch and C. Sander, *Biopolymers*, 1983, **22**, 2577–2637.

62 E. Papaleo, G. Saladino, M. Lambrughi, K. Lindorff-Larsen, F. L. Gervasio and R. Nussinov, *Chemical Reviews*, 2016, **116**, 6391–6423.

63 K. A. Henzler-Wildman, M. Lei, V. Thai, S. J. Kerns, M. Karplus and D. A. Kern, *Nature*, 2007, **450**, 913–916.

64 Y. Gu, D.-W. Li and R. Brüschweiler, *Journal of Chemical Theory and Computation*, 2015, **11**, 1308–1314.