JOC The Journal of Organic Chemistry

Article

pubs.acs.org/joc

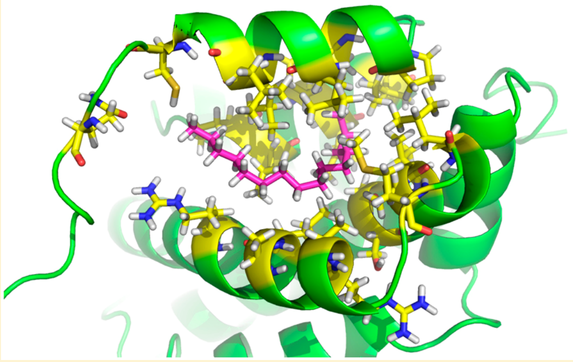
Hydrocarbon Binding by Proteins: Structures of Protein Binding Sites for $\geq C_{10}$ Linear Alkanes or Long-Chain Alkyl and Alkenyl Groups

Jiyong Park,[†] Hung V. Pham,[†] Kristian Mogensen,[‡,§] Theis Ivan Solling,[‡] Martin Vad Bennetzen,[‡] and K. N. Houk*,[†]

[†]Department of Chemistry and Biochemistry, University of California, Los Angeles, California 90095, United States

[‡]Maersk Oil Research and Technology Centre, Education City, Al Gharrafa Street, Al Rayyan, PO Box 22050, Doha, Qatar

ABSTRACT: In order to identify potential de novo enzyme templates for the cleavage of C−C single bonds in long-chain hydrocarbons, we analyzed protein structures that bind substrates containing alkyl and alkenyl functional groups. A survey of ligand-containing protein structures deposited in the Protein Data Bank resulted in 874 entries, consisting of 194 unique ligands that have ≥10 carbons in a linear chain. Fatty acids and phospholipids are the most abundant types of ligands. Hydrophobic amino acids forming α-helical structures frequently line the binding pockets. Occupation of these binding sites was evaluated by calculating both the buried surface area and volume employed by the ligands; these quantities are similar to those computed for drug−protein complexes. Surface complementarity is relatively low due to the nonspecific nature of the interaction between the long-chain hydrocarbons and the hydrophobic amino acids. The selected PDB structures were annotated on the basis of their SCOP and EC identification numbers, which will facilitate design template searches based on structural and functional homologies. Relatively low surface complementarity and ~55% volume occupancy, also observed in synthetic-host, alkane-guest systems, suggest general principles for the recognition of long-chain linear hydrocarbons.



1. INTRODUCTION

Long-chain alkyl or alkenyl groups are commonly found in nature; vital components of living organisms such as fatty acids, lipids, and biological surfactant molecules all contain long hydrocarbon moieties. Thus, the recognition of specific alkyl substrates by proteins is of utmost biological importance. For example, P450 enzymes containing a heme cofactor can catalyze the hydroxylation of long-chain alkanes under aerobic conditions,[1] drawing interest from both science and engineering disciplines due to their potential utility in biofuel production.[2] Intriguing examples of long-chain alkane recognition can also be found in microorganisms residing in deserted geographical regions such as swamps, marine sediment, and deep oil wells, where they have evolved to thrive under these harsh conditions by utilizing long-chain hydrocarbons as their carbon source.[3] More recently, microbial genomic studies suggest the presence of enzymes capable of decomposing long-chain alkanes under anaerobic conditions;[4,5] however, detailed structural information about the conformation of the bound substrate has yet to be determined.

The recognition of linear alkane motifs is of interest to biochemists as well as synthetic chemists. Because C−C bond activation has become an important research topic of synthetic chemistry, there is a growing interest in catalysts that are capable of promoting C−C bond activation with proper regio- and stereoselectivity.[6] We envision de novo-designed enzymes capable of catalyzing the functionalization and cleavage of C−C bonds in long-chain alkanes.[7] As a first step in the design process, scaffolds are sought upon which the catalytic groups required to effect the chemical reaction of interest can be installed. Further understanding of substrate−host interactions is necessary to optimize the substrate recognition capacity of these de novo enzymes, facilitating development of a regio- and stereoselective catalyst. These requirements motivated us to collect and curate the structural information on proteins bound to long-chain alkanes. Specifically, we aimed to answer the following questions: How do proteins recognize and bind long-chain alkyl and alkenyl motifs? What characteristics are shared by the binding pockets of these proteins? Can structural and functional characterization of these proteins lead to valuable insights useful for the development of C−C bond-cleaving enzymes?

In order to answer these questions, we selectively retrieved atomic-resolution protein structures with bound ligands containing long-chain alkyl functional groups (10 carbons or greater) from the Protein Data Bank (PDB).[8] The selection criteria resulted in 874 hits in total, encompassing 194 unique ligands and 737 distinct proteins. We analyzed both the bound substrates and the protein binding sites, generating statistics

Received: November 1, 2014
Published: December 19, 2014

DOI: 10.1021/jo502488e

based on the following data: the type and size distribution of ligands, the binding pocket amino acids and their secondary structures, the solvent-accessible surface area (SASA)[9] buried upon ligand binding, surface complementarity of the ligand−protein interface, and the fraction of the binding pocket volume occupied by the ligand. Finally, we classified select PDB entries according to both the structural classification of proteins (SCOP)[10] and functional categories based on UniProt.[11] We also discuss similarities to synthetic hosts capable of recognizing linear alkanes as guest molecules.

2. RESULTS AND DISCUSSION

2.1. Searching for High-Resolution Protein Structures Containing Long-Chain Alkanes. We searched for known protein structures deposited in the PDB and retrieved entries containing linear alkane motifs (Figure 1). Out of over 87 000 ligand-containing PDB structures, we filtered out any ligands having fewer than 10 carbon atoms in a linear chain or possessing a cyclic moiety. As summarized in Table 1, 874 PDB entries isolated with 194 unique ligands bound to proteins were identified (**set 1**). In the following sections, we present a statistical analysis on the nature of interactions between proteins and long-chain alkyl ligands.

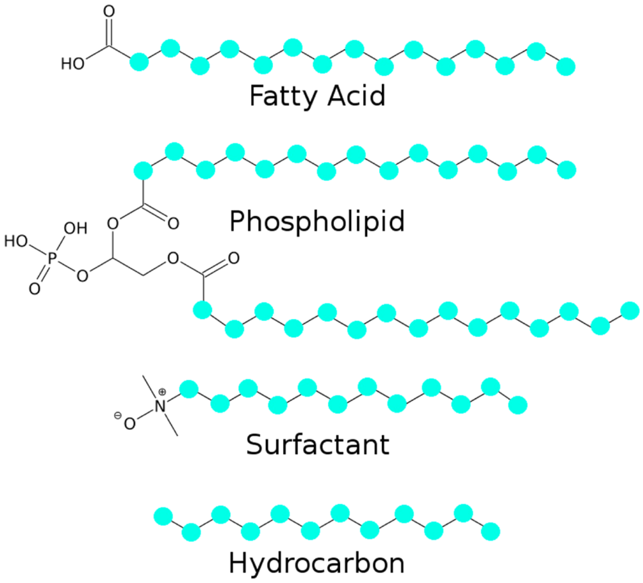


Figure 1. Selection of ligands having linear alkane motifs. Non-amino acid ligands such as fatty acids, phospholipids, surfactants, and hydrocarbons were selectively retrieved from the PDB database. Linear alkane motifs are highlighted in cyan.

Table 1. Description of the Different Protein−Ligand Complex Datasets

name	selection criteria	no. of protein−ligand complexes	no. of distinct ligands
set 1	Contains ligand with ≥10 linear carbons	874	194
set 2	Subset of **set 1**, containing water-soluble proteins only	428	143
set 3	Subset of **set 2**, containing hydrocarbon ligands	28	9
CSAR[12] (http:www.csardock.org)	A benchmark data set for ligand−protein docking studies	118	116

We also considered two subsets of proteins that are especially significant to the understanding of long-chain alkane recognition in aqueous solutions. Out of the 874 PDB entries with ligands containing linear alkyl groups ≥10 carbons, 428 entries were soluble proteins (**set 2**). Furthermore, 28 of those soluble proteins were bound to pure hydrocarbons (**set 3**), which are listed in Table 2. As our motivation was to identify scaffolds for de novo-designed enzymes in aqueous media, we extended our statistical analysis to include the subsets of water-soluble proteins with ligands containing long-chain alkyl groups (**set 2**) and water-soluble proteins with hydrocarbon ligands (**set 3**).

Table 2. PDB Entries of Soluble Proteins and Their Corresponding Ligands Containing Alkyl and Alkenyl Groups Larger than C_{10}

PDB ID	protein	ligand name	formula
1EVY	Glycerol-3-phosphate dehydrogenase	Pentadecane	$C_{15}H_{32}$
1EVZ	Glycerol-3-phosphate dehydrogenase	Pentadecane	$C_{15}H_{32}$
1GKA	Beta-crustacyanin	Dodecane	$C_{12}H_{26}$
1GZP	T-cell surface glycoprotein CD1b	Dodecane	$C_{12}H_{26}$
1GZP	T-cell surface glycoprotein CD1b	Docosane	$C_{22}H_{46}$
1GZQ	T-cell surface glycoprotein CD1b	Dodecane	$C_{12}H_{26}$
1GZQ	T-cell surface glycoprotein CD1b	Docosane	$C_{22}H_{46}$
1JDJ	Glycerol-3-phosphate dehydrogenase	Pentadecane	$C_{15}H_{32}$
1TI1	Thiol:disulfice interchange protein DsbA	Dodecane	$C_{12}H_{26}$
1Y9L	Lipoprotein MxiM	Undecane	$C_{11}H_{24}$
1Z4A	Ferritin	Eicosane	$C_{20}H_{42}$
1Z5L	T-cell surface glycoprotein CD1d antigen	Hexadecane	$C_{16}H_{34}$
2CME	Protein 9b	Decane	$C_{10}H_{22}$
2EUM	Glycolipid transfer protein	Decane	$C_{10}H_{22}$
2EVS	Glycolipid transfer protein	Decane	$C_{10}H_{22}$
2H4T	Carnitine O-palmitoyltransferase 2	Decane	$C_{10}H_{22}$
2ZYH	Lipase	Hexadecane	$C_{16}H_{34}$
3ARB	Antigen-presenting glycoprotein CD1d1	Dodecane	$C_{12}H_{26}$
3FE6	Pheromone binding protein ASP1	(20*S*)-20-methyldotetracontane	$C_{43}H_{88}$
3FE8	Pheromone binding protein ASP1	(20*S*)-20-methyldotetracontane	$C_{43}H_{88}$
3FE9	Pheromone binding protein ASP1	(20*S*)-20-methyldotetracontane	$C_{43}H_{88}$
3OAX	Rhodopsin	(4*E*,6*E*)-hexadeca-1,4,6-triene	$C_{16}H_{28}$
3OV6	T-cell surface glycoprotein CD1b	Dodecane	$C_{12}H_{26}$
3R9B	Cytochrome P450 164A2	Dodecane	$C_{12}H_{26}$
3TZV	T-cell receptor; glycoprotein CD1d	Dodecane	$C_{12}H_{26}$
3U0P	Antigen-presenting glycoprotein CD1d	Undecane	$C_{11}H_{24}$
4FXZ	Bacterial leucine transporter	Undecane	$C_{11}H_{24}$

2.2. Statistics of Bound Ligands Containing Alkyl Groups ≥C_{10}. Figure 2 contains histograms portraying the abundance of the different types of ligands with alkyl groups

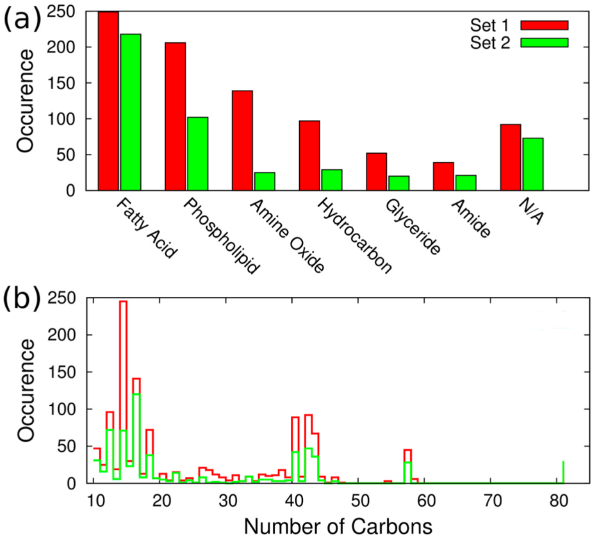


Figure 2. Functional classification and the size distribution of the ligands containing linear alkane motifs. (a) Ligands were classified according to the functional group attached to the linear alkane motif. (b) The distribution of the number of carbon atoms in each ligand. Ligands from all proteins are color-coded in red, whereas those from water-soluble proteins are in green.

longer than C_{10} as well as the number of carbon atoms present in those ligands. As shown in Figure 2a, fatty acids are the most abundant type of ligand in **set 1**, followed by phospholipids, amine oxides, hydrocarbons, glycerides, and amides. Fatty acids remained the most common ligand once we limited our interest to only soluble proteins (**set 2**), whereas the abundance of other ligands significantly decreased; typical surfactants[13] like phospholipids and amine oxides are associated with membrane-bound proteins, significantly reducing their presence in water-soluble proteins.

We then considered the size of the bound ligands in Figure 2b. The number of carbon atoms was used as an index of ligand size. The histogram of **set 1** has its maximum at 14 carbon atoms, composed of lauryl dimethylamine-*N*-oxide, myristic acid, (10*E*,12*Z*)-tetradeca-10,12-dien-1-ol, *S*-[2-(acetylamino)-ethyl](3*R*)-3-hydroxydecanethioate, tetradecane, and (*R*)-3-hydroxytetradecanal (Figure 3). The largest entry found was cardiolipin (81 carbon atoms),[14] a diphosphatidylglycerol molecule having four linear alkyl functional groups. The histogram of long-chain alkanes interacting with water-soluble proteins (**set 2**) was qualitatively similar to that of **set 1**; ligands with 16 carbons were now the most common, whose members include 16-hydroxyhexadecanoic acid, (11*Z*,13*Z*)-hexadeca-11,13-dien-1-ol, (4*E*,6*E*)-hexadeca-1,4,6-triene, (2,2-diphosphonoethyl)(dodecyl)dimethylphosphonium, (10*E*,12*Z*)-hexadeca-10,12-dienal, (10*E*)-hexadec-10-en-12-yn-1-ol, hexadeca-10,12-dien-1-ol, *N*-dodecyl-*N*,*N*-dimethylglycinate, decamethonium ion, 1-decyl-3-trifluoro ethyl-*sn*-glycero-2-phosphomethanol, 1-hexadecanesulfonic acid, 1-iodohexadecane, 10-oxohexadecanoic acid, hexadecan-1-ol, palmitic acid, and hexadecane. Once again, the largest ligand bound was cardiolipin.

2.3. Statistics on Amino Acids and Folds of Ligand-Binding Pockets.

Figure 4a portrays the frequency of amino acids defining the ligand-binding pocket. As detailed in the Methods, Rosetta Interface Analyzer[15] was used to identify the binding pocket amino acids surrounding the bound ligand. The analysis of all long-chain alkane-binding proteins (**set 1**) resulted in the hydrophobic residues leucine (14%) and phenylalanine (9%) as two of the most abundant amino acids. These became more abundant in water-soluble proteins (**set 2**) and hydrocarbon-bound soluble proteins (**set 3**). The binding pocket residues surrounding long-chain linear alkyl groups were then compared to those of drug−target proteins deposited in **CSAR**.[12] As drug-like molecules tend to include ring moieties and polar functional groups, this assessment could

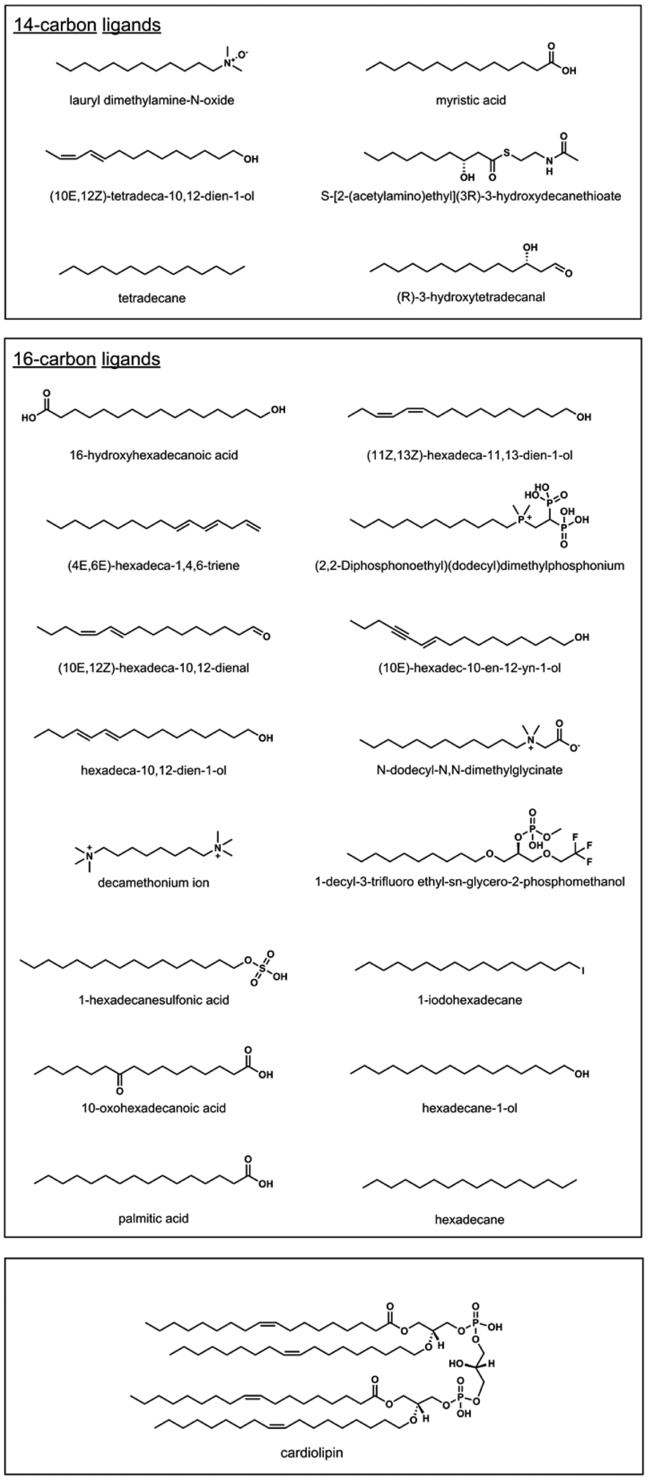

14-carbon ligands
lauryl dimethylamine-N-oxide
myristic acid
(10E,12Z)-tetradeca-10,12-dien-1-ol
S-[2-(acetylamino)ethyl](3R)-3-hydroxydecanethioate
tetradecane
(R)-3-hydroxytetradecanal
16-carbon ligands
16-hydroxyhexadecanoic acid
(11Z,13Z)-hexadeca-11,13-dien-1-ol
(4E,6E)-hexadeca-1,4,6-triene
(2,2-Diphosphonoethyl)(dodecyl)dimethylphosphonium
(10E,12Z)-hexadeca-10,12-dienal
(10E)-hexadec-10-en-12-yn-1-ol
hexadeca-10,12-dien-1-ol
N-dodecyl-N,N-dimethylglycinate
decamethonium ion
1-decyl-3-trifluoro ethyl-sn-glycero-2-phosphomethanol
1-hexadecanesulfonic acid
1-iodohexadecane
10-oxohexadecanoic acid
hexadecane-1-ol
palmitic acid
hexadecane
cardiolipin

Figure 3. Structures of the 14-carbon ligands, 16-carbon ligands, and cardiolipin.

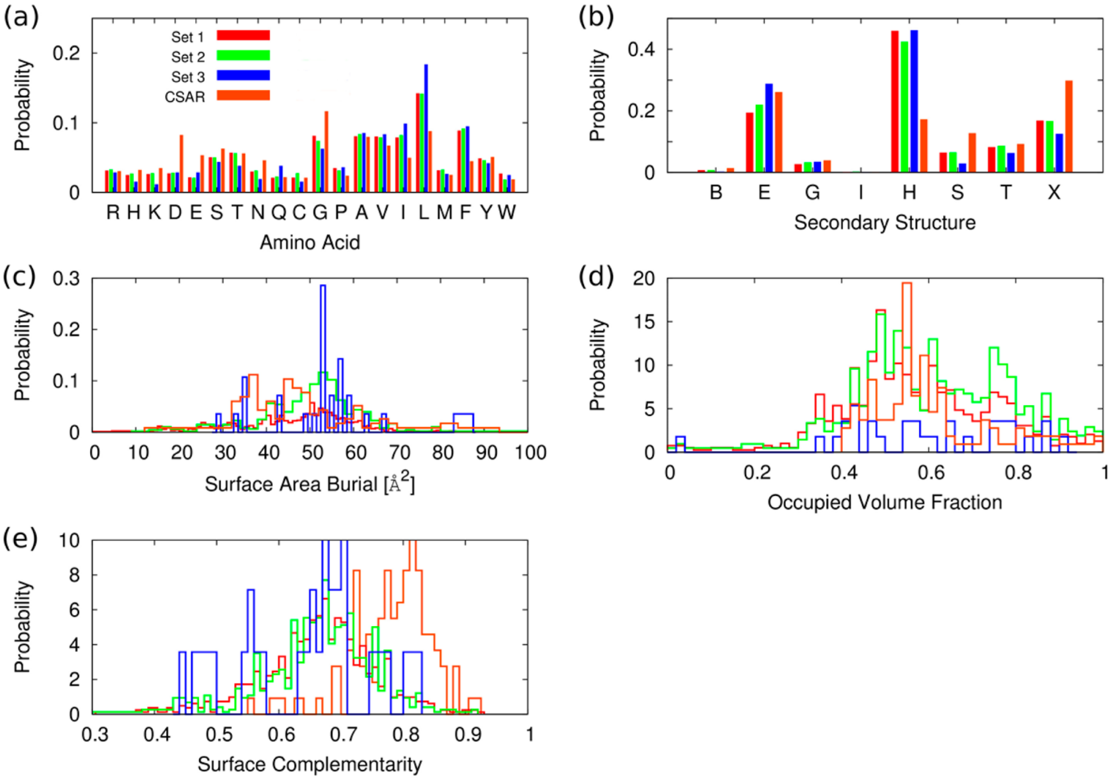


Figure 4. Statistics of the ligand-binding pockets: (a) population of amino acids, (b) backbone secondary structure distribution, (c) surface area burial per ligand-carbon atom, (d) occupied volume fraction of binding pocket by the ligand, and (e) surface complementarity between protein and ligand. Secondary structure abbreviations: B, β-bridge; E, β-sheet; G, turn; I, π-helix; H, α-helix; S, bend; T, hydrogen-bonded turn; and X, unstructured.

be beneficial for isolating unique features of ligand-binding pockets that specifically recognize alkyl groups that are linear, nonpolar, and hydrophobic. The comparison suggests that the populations of hydrophobic residues, including leucine, valine, and phenylalanine, are enriched in hydrocarbon-binding pockets relative to the binding pockets of drug molecules.

The secondary structures of the amino acids forming the protein backbone of the binding pocket were analyzed in Figure 4b using the DSSP software.[16] The α-helices (H, 45%) and β-sheets (E, 20%) are the two most prevalent secondary structures found from the binding pocket residues of **set 1**. The higher frequency of α-helices lining the binding pocket is consistent in **sets 2** and **3**. An examination of the **CSAR** data set (Figure 4b, orange) shows a dramatic decrease in α-helices relative to the other secondary structures. The population of unstructured secondary structures is doubled in reference to **set 3**. These findings strongly indicate that the binding of long-chain linear alkanes occurs at protein surfaces made of α-helices more often than any other secondary structure elements.

In Figure 4c, we plot the results of the computed surface area burial (SAB) per ligand-carbon atom upon formation of the protein−ligand complex. SAB provides a quantitative measure of how tightly a protein captures its ligand. On average, each carbon atom of the ligands in **set 1** buried 47 ± 14 Å^2 of the solvent-accessible surface area (SASA) of the binding pocket. Water-soluble proteins bound to ligands containing long-chain alkanes (**set 2**) and to pure hydrocarbons (**set 3**) resulted in 50 ± 12 and 52 ± 12 Å^2 of the SAB, respectively. Although the average SAB is largest for **set 3**, the difference from **set 1** is within statistical uncertainty. The SAB of drug-binding pockets (49 ± 22 Å^2) is also comparable to that of the long-chain alkane-binding proteins. In short, the SAB per ligand-carbon atom of long-chain alkane-binding proteins is 47−52 Å^2 on average and is similar to that of drug−target proteins.

In addition to surface area burial, we also analyzed the binding pocket volume occupied by the ligand (occupied volume fraction, OVF) in Figure 4d. Mecozzi and Rebek pointed out that 55% is an optimal value for OVF, considering both the favorable enthalpic interactions between ligand and host as well as the entropic penalty associated with the limited conformational degrees of freedom imposed on the bound ligand.[17] We computed the OVF of ligand binding pockets using POVME software.[18] For both **set 1** and **set 2**, the OVF is close to the conjectured optimal value: 57 ± 18 and 61 ± 17%, respectively. The same computation on the **CSAR** data set resulted in a similar observation: the OVF was 59 ± 13%. The average OVF values of naturally occurring proteins bound to long-chain alkanes and of designed drug-like molecules are similar and are close to the optimal OVF value of 55%.

Finally, we computed the surface complementarity[19] (SC) between the binding pocket residues and the bound ligands. SC quantifies the congruency between two interacting molecular surfaces, where the SC of two perfectly complementary surfaces is 1 and that of two adjacent random shapes approaches 0. This quantity has been understood to be one of the fundamental descriptors of the compliance of two interacting molecules.[20] For the long-chain alkane-binding proteins, computed SC values are similar regardless of their solubility profile: 0.66 ± 0.08, 0.66 ± 0.09, and 0.63 ± 0.09 for **set 1**, **set 2**, and **set 3**, respectively. On the other hand, the interfaces of drug-like molecules showed enhanced SC over that of the alkane-binding proteins: SC of the **CSAR** data set is 0.78 ± 0.06. SC is known

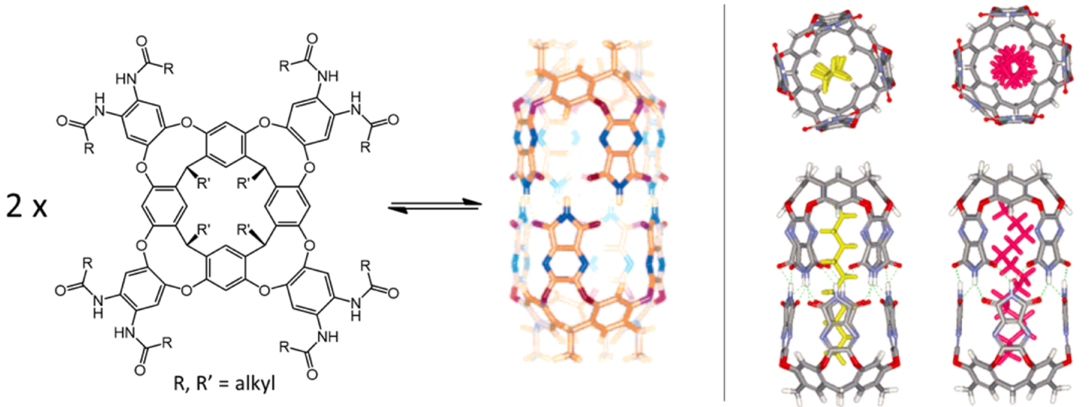


Figure 5. Left: A resorcinarene-based cavitand that can dimerize, creating a molecular capsule. Right: Top and side views of encapsulation of *n*-$C_{10}H_{22}$ in a straight-chain conformation (yellow) and *n*-$C_{14}H_{30}$ (red) in a helical arrangement. Reprinted from ref 23. Copyright 2004 American Chemical Society.

Table 3. SCOP Classification of the Selected PDB Entries Containing Linear Alkane Motifs

SCOP fold name	no. of occurrences
Bacterial photosystem II reaction center, L and M subunits	53
Family A G protein-coupled receptor-like	35
Nucleoplasmin-like/VP (viral coat and capsid proteins)	21
Lipocalins	21
Cytochrome c oxidase subunit III-like	14
Serum albumin-like	13
Cytochrome c oxidase subunit I-like	12
Transmembrane beta-barrels	11
Single transmembrane helix	11
Phospholipase A2, PLA2	11
Nuclear receptor ligand-binding domain	11
Cupredoxin-like	11
Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin	11
Thiolase-like	10
PRC-barrel domain	10
Immunoglobulin-like beta-sandwich	10
Ganglioside M2 (gm2) activator	10
alpha/beta-Hydrolases	8
Heme-binding four-helical bundle	7
Cytochrome P450	7
alpha/alpha toroid	7
Ribosomal protein S5 domain 2-like	6
a domain/subunit of cytochrome bc1 complex (Ubiquinol-cytochrome c reductase)	6
6-Phosphogluconate dehydrogenase C-terminal domain-like	6
Voltage-gated potassium channels	4
TIM beta/alpha-barrel	4
S-Adenosyl-L-methionine-dependent methyltransferases	4
Prealbumin-like	4
Cytochrome c	4
SCP-like	3
EF Hand-like	3
DAK1/DegV-like	3
alpha−alpha superhelix	3
Snake toxin-like	2
Rhomboid-like	2
MHC antigen-recognition domain	2
Lysozyme-like	2
Light-harvesting complex subunits	2
GroES-like	2
Glycolipid transfer protein, GLTP	2
Ferredoxin-like	2
beta-hairpin stack	2
Aha1/BPI domain-like	2
Acyl carrier protein-like	2
Thioredoxin fold	1
Thioesterase/thiol ester dehydrase-isomerase	1
TBP-like	1
SH3-like barrel	1
SARS ORF9b-like	1
Saposin-like	1
RuvA C-terminal domain-like	1
RRF/tRNA synthetase additional domain-like	1
Photosystem I subunits PsaA/PsaB	1
(Phosphotyrosine protein) phosphatases II	1
Ntn hydrolase-like	1
NAD(P)-binding Rossmann-fold domains	1
LuxS/MPP-like metallohydrolase	1
Long alpha-hairpin	1
Lipase/lipooxygenase domain (PLAT/LH2 domain)	1
Kringle-like	1
ISP domain	1
Gelsolin-like	1
FAD/NAD(P)-binding domain	1
Double-stranded beta-helix	1
DNA/RNA-binding 3-helical bundle	1
DhaL-like	1
Cystatin-like	1
Clc chloride channel	1
Class II aaRS and biotin synthetases	1
Chlorophyll a-b binding protein	1
Bromodomain-like	1
A DNA-binding domain in eukaryotic transcription factors	1

to be correlated with the specificity of the interaction between a ligand and its binding pocket.[21] The SC of interacting protein surfaces ranges from 0.70 to 0.76,[19] resembling that of the **CSAR** data set. Furthermore, drug molecules are optimized to

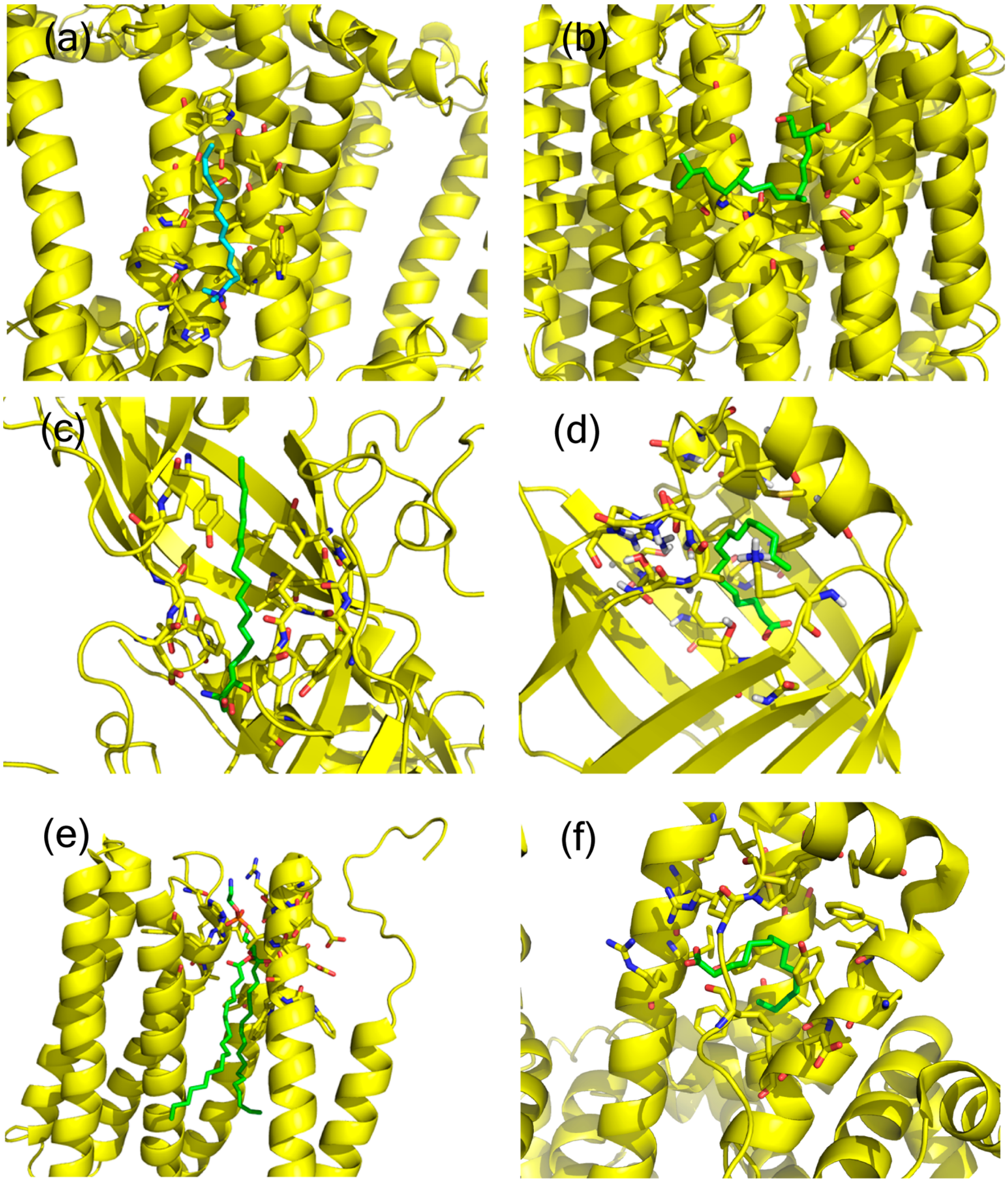


Figure 6. Frequently observed SCOP folds binding ligands with linear alkane motifs: (a) bacterial photosystem II reaction center protein (PDB ID: 1AIJ) bound to lauryl dimethylamine-*N*-oxide, (b) family A G protein-coupled receptor-like protein (PDB ID: 1BRR) bound to 3,7,11,15-tetramethyl-1-hexadecanol, (c) nucleoplasmin-like/VP protein (PDB ID: 1AL2) bound to sphingosine, (d) lipocalin (PDB ID: 1B56) bound to palmitic acid, (e) cytochrome c oxidase, subunit III (PDB ID: 1M56) bound to distearoyl-3-*sn*-phosphatidylethanolamine, and (f) serum albumin-like protein (PDB ID: 1H9Z) bound to myristic acid. PyMOL was used for molecular visualization.[18] Cartoon representations represent backbone arrangements of the protein, and bound ligands and binding pocket residues are shown using stick representations.

achieve enhanced selectivity toward their targets, exploiting specific interactions such as hydrogen bonding and electrostatic interactions. These tendencies are expected to result in higher SC values. In contrast, naturally occurring alkane-binding proteins stabilize their substrates through relatively weaker nonpolar interactions, resulting in smaller SC values than the protein−protein or protein−drug interfaces.

2.4. Alkane Binding by Synthetic Hosts. Aside from the naturally occurring biomolecules we surveyed, synthetic supramolecular hosts have also been shown to bind diverse substrates. Cram pioneered the uses of carcerands, cavitands, and other molecular capsules as molecular containers and to catalyze reactions or stabilize reactive intermediates, garnering much interest in the scientific community.[22] Of particular relevance, Rebek investigated the kinetics and thermodynamics of binding medium-chain alkanes with resorcinarene-based cavitands.[23,24] The host molecules dimerize, as shown in Figure 5, forming pill-shaped compartments capable of enclosing *n*-alkanes, from C_9 to C_{14}; shorter alkanes are bound in an extended straight-chain conformation, whereas longer chains adopt a folded and helical arrangement. Although the coiling of the longer alkanes results in unfavorable gauche conformations, the favorable C−H···π interactions with the aromatic walls of the cavitand compensates for the increased torsional strain. The usual cutoff distance for C−H···π interactions is considered to be approximately 3 Å, calculated from the respective van der

Table 4. Subset of the Selected PDB Entries Having Enzymatic Activity

UniProt ID	EC no.	description
O33877	4.2.1.59	3-Hydroxydecanoyl-[acyl-carrier-protein] dehydratase
P0A574	2.3.1.180	3-Oxoacyl-[acyl-carrier-protein] synthase 3
P44783	3.4.21.10	5 Rhomboid protease GlpG
P09391	3.4.21.10	5 Rhomboid protease GlpG
P04058	3.1.1.7	Acetylcholinesterase
P21836	3.1.1.7	Acetylcholinesterase
Q6SLM2	3.1.1.4	Acidic phospholipase A2 1
P0AGG2	3.1.2.-	Acyl-CoA thioesterase 2
Q9NPJ3	3.1.2.-	Acyl-coenzyme A thioesterase 13
Q9I194	3.5.1.97	Acyl-homoserine lactone acylase PvdQ
P11766	1.1.1.1	Alcohol dehydrogenase class-3
O96759	2.5.1.26	Alkyldihydroxyacetonephosphate synthase
P97275	2.5.1.26	Alkyldihydroxyacetonephosphate synthase, peroxisomal
Q7D8I1	2.3.1.-	Alpha-pyrone synthesis polyketide synthase-like Pks18
P21397	1.4.3.4	Amine oxidase [flavin-containing] A
P27338	1.4.3.4	Amine oxidase [flavin-containing] B
P06653	3.5.1.28	Autolysin
B2IZD3	3.6.5.5	Bacterial dynamin-like protein
P59071	3.1.1.4	Basic phospholipase A2 VRV-PL-VIIIa
P14779	1.14.14.1	Bifunctional P-450/NADPH-P450 reductase
P00918	4.2.1.1	Carbonic anhydrase 2
P18886	2.3.1.21	Carnitine *O*-palmitoyltransferase 2, mitochondrial
P07773	1.13.11.1	Catechol 1,2-dioxygenase
P11451	1.13.11.-	Chlorocatechol 1,2-dioxygenase
P00590	3.1.1.74	Cutinase 1
P0C5C2	2.1.1.79	Cyclopropane mycolic acid synthase 1
P0A5P0	2.1.1.79	Cyclopropane mycolic acid synthase 2
Q79FX6	2.1.1.79	Cyclopropane mycolic acid synthase MmaA2
P08067	1.10.2.2	Cytochrome *b*-c1 complex subunit Rieske, mitochondrial
P98005	1.9.3.1	Cytochrome c oxidase polypeptide I+III
P00396	1.9.3.1	Cytochrome c oxidase subunit 1
P33517	1.9.3.1	Cytochrome c oxidase subunit 1
P08306	1.9.3.1	Cytochrome c oxidase subunit 2
P10632	1.14.14.1	Cytochrome P450 2C8
Q9H227	3.2.1.21	Cytosolic beta-glucosidase
Q02127	1.3.5.2	Dihydroorotate dehydrogenase (quinone), mitochondrial
Q08210	1.3.5.2	Dihydroorotate dehydrogenase (quinone), mitochondrial
P45510	2.7.1.29	Dihydroxyacetone kinase
Q9R1E6	3.1.4.39	Ectonucleotide pyrophosphatase/phosphodiesterase family member 2
P0A5Y6	1.3.1.9	Enoyl-[acyl-carrier-protein] reductase [NADH]
P97612	3.5.1.99	Fatty-acid amide hydrolase 1
P03368	3.4.23.16	Gag-Pol polyprotein
P03369	3.4.23.16	Gag-Pol polyprotein
P80035	3.1.1.3	Gastric triacylglycerol lipase
O91734	3.4.22.29	Genome polyprotein
P03300	3.4.22.29	Genome polyprotein
P04936	3.4.22.29	Genome polyprotein
P12915	3.4.22.29	Genome polyprotein
Q66282	3.4.22.29	Genome polyprotein
Q66479	3.4.22.29	Genome polyprotein
Q82122	3.4.22.29	Genome polyprotein
Q12051	2.5.1.-	Geranylgeranyl pyrophosphate synthase
O35000	3.5.99.6	Glucosamine-6-phosphate deaminase 1
P90551	1.1.1.8	Glycerol-3-phosphate dehydrogenase [NAD$^+$], glycosomal
P48449	5.4.99.7	Lanosterol synthase
O59952	3.1.1.3	Lipase
P32947	3.1.1.3	Lipase 3
P41365	3.1.1.3	Lipase B
P37001	2.3.1.-	Lipid A palmitoyltransferase PagP
P23141	3.1.1.1	Liver carboxylesterase
P00698	3.2.1.17	Lysozyme C
Q9I596	3.5.1.23	Neutral ceramidase
Q6UEH2	2.3.1.221	Noranthrone synthase
Q10404	2.3.1.181	Octanoyltransferase
P52708	4.1.2.11	P-(*S*)-Hydroxymandelonitrile lyase
P16233	3.1.1.3	Pancreatic triacylglycerol lipase
P07872	1.3.3.6	Peroxisomal acyl-coenzyme A oxidase 1
P0A921	3.1.1.32	Phospholipase A1
P00593	3.1.1.4	Phospholipase A2
P00592	3.1.1.4	Phospholipase A2, major isoenzyme
P14555	3.1.1.4	Phospholipase A2, membrane associated
P0A405	1.97.1.12	Photosystem I P700 chlorophyll a apoprotein A1
D0VWR8	1.10.3.9	Photosystem II D2 protein
P51765	1.10.3.9	Photosystem Q(B) protein
P50264	1.5.3.17	Polyamine oxidase FMS1
Q05769	1.14.99.1	Prostaglandin G/H synthase 2
P41222	5.3.99.2	Prostaglandin-H2 D-isomerase
P25043	3.4.25.1	Proteasome subunit beta type-2
Q02293	2.5.1.58	Protein farnesyltransferase subunit beta
Q04631	2.5.1.58	Protein farnesyltransferase/geranylgeranyltransferase type-1 subunit alpha
P00735	3.4.21.5	Prothrombin
P0A516	1.14.-.-	Putative cytochrome P450 124
P96416	1.-.-.-	R2-like ligand binding oxidase
P04191	3.6.3.8	Sarcoplasmic/endoplasmic reticulum calcium ATPase 1
P33247	4.2.1.129	Squalene–hopene cyclase
Q5EGY4	2.3.1.-	Synaptobrevin homologue YKT6
P96086	3.4.21.-	Tricorn protease
P00520	2.7.10.2	Tyrosine-protein kinase ABL1
Q06124	3.1.3.48	Tyrosine-protein phosphatase nonreceptor type 11
O67648	3.5.1.-	UDP-3-*O*-[3-hydroxymyristoyl] *N*-acetylglucosamine deacetylase
P0CD76	2.3.1.-	UDP-3-*O*-acylglucosamine *N*-acyltransferase

Waals radii. Both computational[25] and crystallographic data[26] of encapsulated alkyl guests exhibit these interactions.

For guest molecules with narrow, extended conformations, the cavitand host deforms not only to increase C−H···π interactions but also to attain more suitable packing coefficients. Rebek's dimeric host has a calculated volume of 425 Å^3 and can bind guests that occupy about 55 ± 9% of the available volume, similar to the packing efficiency of most organic liquids.[17] Guests that do not sufficiently fill the empty space suffer from the large entropic penalty of complexation; the empty hosts prefer to be filled with solvent when the alkane is too small. In contrast, larger guests cause the binding site to be too crowded, thus experiencing steric repulsion. Other groups have shown that this general 55% parameter also applies to enzyme binding pockets.[27] The lack of observed complexes of the Rebek host cavitands with alkanes smaller than 9 carbons or longer than 14 carbons is a testament to the importance of size and shape complementarity when encapsulating substrates without any functional handles.

Through minor modifications of the cavitand molecules, formation of the dimer was suppressed, and hydrophilic feet were incorporated to create water-soluble supramolecules that were capable of binding medium-chain alcohols.[28] The polar hydroxyl group remains exposed to the aqueous environment, and the hydrophobic alkyl chain coils toward the inner cavity of the host, similar to the alkane conformations mentioned earlier. The alkane size and shape complementarity exhibited by these complexes bears some resemblance to the naturally occurring hydrocarbon-binding sites in proteins.

2.5. Structural and Functional Classification of Long-Chain Alkane-Binding Proteins. We classified each hit from the selected PDB entries based on the structural classification of proteins (SCOP). In general, protein structure determination in the presence of a ligand is more difficult than that in its absence. This implies that high-resolution structures with bound ligands may represent only a subset of proteins having the potential to recognize linear alkanes. Fortunately, structurally similar proteins (homologues) share many functional similarities. Thus, one may establish the structures of proteins interacting with linear hydrocarbon motifs through homologue relationships. As of 2013, only 38 222 PDB entries had SCOP classification IDs, from which we were able to classify 407 out of the 874 hits (Table 3). SCOP classifies proteins into several hierarchical levels, utilizing either their evolutionary relationships or structural similarities: the fold hierarchy of a protein reflects structural relationships with other proteins, whereas both family and superfamily hierarchies are based on evolutionary origin and functional similarity. We focused on the fold classification of proteins in the PDB search hits because we use this classification with the alkane-binding proteins to facilitate identification of protein design scaffolds sharing similar structural features. There are 72 distinct SCOP folds identified out of 407 proteins having SCOP IDs. For multidomain proteins, each domain in contact with the linear hydrocarbon ligand was analyzed separately. The most prevalent fold is the bacterial photosystem II reaction center, L and M subunits. Representative structures of the top six most frequently found SCOP folds are depicted in Figure 6a−f.

Next, we considered functional attributes of the selected protein templates. Our specific interest was to identify alkane-binding proteins that have enzymatic activity. These proteins possess catalytic functional groups and/or bound cofactors that have more easily modifiable characteristics than nonenzymatic proteins. UniProt[11] is the central information repository of genomic sequence and functional information on proteins. Each entry in the PDB has one or more UniProt identification numbers, enabling us to annotate the functional role of each PDB structure containing ligands with a long-chain alkane motif. A subset of the selected PDB templates has enzymatic activity, which is identified by the enzyme commission (EC) number. On the basis of these functional descriptions of each entry from the UniProt database, we classified the 874 selected protein templates into functional categories. First, we identified enzymes having catalytic functionality with specific ligands (Table 4): there are 202 enzymes identified, catalyzing 89 distinct chemical reactions. The most frequently identified enzyme was cytochrome c oxidase (UniProt ID: P00396, 17 entries), and the second was viral protease/RNA transferases (UniProt ID: P03300, 10 entries). There are also enzymes associated with biological reactions involving linear alkyl and alkenyl functional groups, such as 3-oxoacyl-[acyl-carrier-protein] synthase III (UniProt ID: P0A574), phospholipase A2 (UniProt ID: P00592), and cytochrome-P450 monooxygenase (UniProt ID: P14779).

Finally, we questioned whether the binding sites of enzymes are significantly different from those of the nonenzymatic hydrocarbon-binding proteins, as enzyme catalysis usually requires the precise placement of substrates, leading to an enhanced binding specificity. However, statistics such as SC and OVF of the enzymes (0.68 ± 0.08 and 57 ± 13%, respectively) are almost identical to those of long-chain alkane-binding proteins (**set 1**). The findings suggest that enzymatic proteins recognize their substrates based on the same chemical principles governing the binding of long-chain alkanes in nonenzymatic proteins.

3. CONCLUSIONS

We have surveyed proteins capable of recognizing long-chain hydrocarbons and long-chain alkyl groups and have considered various factors that influence this binding. Hydrophobic amino acids forming α-helical secondary structures are frequently a major component of the binding sites. The surface complementarity of the ligand−protein interfaces in alkane-binding proteins is lower than that of drug-binding proteins, which typically have more polar substrates. However, the occupied volume fraction and the surface area burial by the ligand−protein interfaces are both comparable to those of drug-binding sites. The volume fraction occupied by the substrates is close to the ideal value of 55%, suggesting substrate recognition mechanisms similar to those of synthetic host molecules. Moreover, structural and functional classifications of the long-chain alkane-binding proteins will aid future efforts in searching for potential protein scaffolds. The protein structures and the analyzed binding-site characteristics should guide the design of new enzymes that can selectively recognize large alkyl substrates and catalyze their functionalization.

4. METHODS

4.1. PDB Database Search. PDB entries containing one or more ligands with 10 or more carbons were selected using the PDB web-search interface. A Python programming library (OEChem[29]) was used to postprocess the initial hits, ruling out any entry possessing rings. OEChem was also used to identify functional motifs in the identified ligands, leading to the classification of each ligand.

4.2. Analysis of Ligand Binding Pockets. The amino acids located in the binding pockets were identified using the Interface Analyzer module in the Rosetta software package.[15] The surface area burial upon binding of the ligand was computed using the same package. The DSSP program was used to define the backbone secondary structure of the binding pocket amino acids.[16] We used POVME software to calculate the binding pocket volume and the occupied volume fraction.[18] For each statistic provided here, standard deviations were used as a measure of statistical uncertainty.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: houk@chem.ucla.edu.

Present Address

§(K.M.) ENI S.p.A. − Exploration & Production Division, fifth Off. Building, Via Emilia 1, 20097 San Donato Milanese (Milan), Italy.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Financial support of this research came from the Maersk Oil Research and Technology Center, Doha, Qatar. J.P., H.V.P., and K.N.H. are grateful to Maersk for financial support of this research.

■ REFERENCES

(1) Timmis, K. N. *Handbook of Hydrocarbon and Lipid Microbiology*; Springer-Verlag: Berlin, 2010.
(2) Yang, Y.; Liu, J.; Li, Z. *Angew. Chem., Int. Ed.* **2014**, *53*, 3120−3124.
(3) Singh, S. N. *Microbial Degradation of Xenobiotics*; Springer: Berlin, 2012.
(4) Callaghan, A. V. *Front. Microbiol.* **2013**, *4*, 89.
(5) (a) Feng, L.; Wang, W.; Cheng, J.; Ren, Y.; Zhao, G.; Gao, C.; Tang, Y.; Liu, X.; Han, W.; Peng, X.; Liu, R.; Wang, L. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 5602−5607. (b) Li, L.; Liu, X.; Yang, W.; Xu, F.; Wang, W.; Feng, L.; Bartlam, M.; Wang, L.; Rao, Z. *J. Mol. Biol.* **2008**, *376*, 453−465.
(6) Cramer, N.; Dermenci, A.; Dong, G.; Douglas, C. J.; Dreis, A. M.; Fu, X.-F.; Gao, Y.; Jones, W. D.; Jun, C.-H.; Kingsbury, J. S.; Moebius, D. C.; Nakao, Y.; Park, J.-W.; Parker, E.; Rendina, V. L.; Souillart, L.; Xu, T.; Yu, Z.-X. *C−C Bond Activation*; Springer: Berlin, 2014.
(7) Kiss, G.; Çelebi-Ölçüm, N.; Moretti, R.; Baker, D.; Houk, K. N. *Angew. Chem., Int. Ed.* **2013**, *52*, 5700−5725.
(8) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(9) Lee, B.; Richards, F. M. *J. Mol. Biol.* **1971**, *55*, 379−400.
(10) (a) Andreeva, A.; Howorth, D.; Brenner, S. E.; Hubbard, T. J.; Chothia, C.; Murzin, A. G. *Nucleic Acids Res.* **2004**, *32*, D226−D229. (b) Hubbard, T. J.; Murzin, A. G.; Brenner, S. E.; Chothia, C. *Nucleic Acids Res.* **1997**, *25*, 236−239.
(11) UniProt Consortium. *Nucleic Acids Res.* **2014**, *42*, D191−D198.
(12) Dunbar, J. B.; Smith, R. D.; Damm-Ganamet, K. L.; Ahmed, A.; Esposito, E. X.; Delproposto, J.; Chinnaswamy, K.; Kang, Y. N.; Kubish, G.; Gestwicki, J. E.; Stuckey, J. A.; Carlson, H. A. *J. Chem. Inf. Model.* **2013**, *53*, 1842−1852.
(13) Sanderson, H.; Tibazarwa, C.; Greggs, W.; Versteeg, D. J.; Kasai, Y.; Stanton, K.; Sedlak, R. I. *Risk Anal.* **2009**, *29*, 857−867.
(14) Paradies, G.; Paradies, V.; De Benedictis, V.; Ruggiero, F. M.; Petrosillo, G. *Biochim. Biophys. Acta* **2014**, *1837*, 408−417.
(15) Lewis, S. M.; Kuhlman, B. A. *PLoS One* **2011**, *6*, e20872.
(16) Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577−2637.
(17) Mecozzi, S.; Rebek, J. J. *Chem.—Eur. J.* **1998**, *4*, 1016−1022.
(18) Durrant, J. D.; de Oliveira, C. A.; McCammon, J. A. *J. Mol. Graphics Modell.* **2011**, *29*, 773−776.
(19) Lawrence, M. C.; Colman, P. M. *J. Mol. Biol.* **1993**, *234*, 946−950.
(20) Shoichet, B. K.; Kuntz, I. D. *J. Mol. Biol.* **1991**, *221*, 327−346.
(21) Tinberg, C. E.; Khare, S. D.; Dou, J.; Doyle, L.; Nelson, J. W.; Schena, A.; Jankowski, W.; Kalodimos, C. G.; Johnsson, K.; Stoddard, B. L.; Baker, D. *Nature* **2013**, *501*, 212−216.
(22) (a) Liu, F.; Wang, H.; Houk, K. N. *Curr. Org. Chem.* **2013**, *17*, 1470−1480. (b) Warmuth, R.; Yoon, J. *Acc. Chem. Res.* **2001**, *34*, 95−105. (c) Stoddart, J. F. *Annu. Rep. Prog. Chem., Sect. B: Org. Chem.* **1988**, *85*, 353−386. (d) Cram, D. J.; Cram, J. M. *Container Molecules and Their Guests*; Royal Society of Chemistry: Cambridge, 1994.
(23) Scarso, A.; Trembleau, L.; Rebek, J. *J. Am. Chem. Soc.* **2004**, *126*, 13512−13518.
(24) Jiang, W.; Ajami, D.; Rebek, J. *J. Am. Chem. Soc.* **2012**, *134*, 8070−8073.
(25) Ruan, Y.; Peterson, P. W.; Hadad, C. M.; Badjić, J. D. *Chem. Commun.* **2014**, *50*, 9086−9089.
(26) Nishio, M.; Umezawa, Y.; Honda, K.; Tsuboyama, S.; Suezawa, H. *CrystEngComm* **2009**, *11*, 1757−1788.
(27) (a) Zürcher, M.; Gottschalk, T.; Meyer, S.; Bur, D.; Diederich, F. *ChemMedChem* **2008**, *3*, 237−240. (b) Kawasaki, Y.; Chufan, E. E.; Lafont, V.; Hidaka, K.; Kiso, Y.; Mario Amzel, L.; Freire, E. *Chem. Biol. Drug Des.* **2010**, *75*, 143−151. (c) Zürcher, M.; Diederich, F. *J. Org. Chem.* **2008**, *73*, 4345−4361. (d) Morellato-Castillo, L.; Acharya, P.; Combes, O.; Michiels, J.; Descours, A.; Ramos, O. H.; Yang, Y.; Vanham, G.; Ariën, K. K.; Kwong, P. D.; Martin, L.; Kessler, P. *J. Med. Chem.* **2013**, *56*, 5033−5047.
(28) Zhang, K. D.; Ajami, D.; Gavette, J. V.; Rebek, J. *J. Am. Chem. Soc.* **2014**, *136*, 5264−5266.
(29) Marcou, G.; Rognan, D. *J. Chem. Inf. Model.* **2007**, *47*, 195−207.