JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

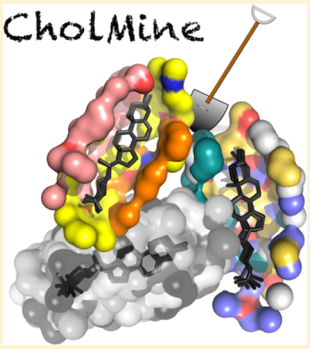
# CholMine: Determinants and Prediction of Cholesterol and Cholate Binding Across Nonhomologous Protein Structures

Nan Liu,[†,‡,§] Jeffrey R. Van Voorst,[‡,§,∥] John B. Johnston,[§] and Leslie A. Kuhn*[,‡,§]

[†]Department of Chemistry, [‡]Department of Computer Science and Engineering, and [§]Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, Michigan 48824-1319, United States

Supporting Information

**ABSTRACT:** Identifying physiological ligands is necessary for annotating new protein structures, yet this presents a significant challenge to biologists and pharmaceutical chemists. Here we develop a predictor of cholesterol and cholate binding that works across diverse protein families, extending beyond sequence motif-based prediction. This approach combines SimSite3D site comparison with the detection of conserved interactions in cholesterol/cholate bound crystal structures to define three-dimensional interaction motifs. The resulting predictor identifies cholesterol sites with an ∼82% unbiased true positive rate in both membrane and soluble proteins, with a very low false positive rate relative to other predictors. The CholMine Web server can analyze users' structures, detect those likely to bind cholesterol/cholate, and predict the binding mode and key interactions. By deciphering the determinants of binding for these important steroids, CholMine may also aid in the design of selective inhibitors and detergents for targets such as G protein coupled receptors and bile acid receptors.




## INTRODUCTION

Membrane proteins are surrounded by a complex mixture of lipids, including phospholipids, cholesterol, and some bile salts (bile acids and alcohols). One of the bile salts, cholate, is often used as a detergent to solubilize membrane proteins.[1,2] Different types of lipids influence biological functions of membrane proteins in direct or indirect ways.[3−5] Conserved binding sites for certain lipids have been characterized on membrane proteins,[4,6,7] and these lipids can play an important role in structural stabilization and biological processes. For example, in bovine heart cytochrome c oxidase (CcO), the tails of two phosphatidylglycerol lipids regulate oxygen transfer to the active site, and phosphatidylethanolamine, cardiolipin, and phosphatidylglycerol are all associated with the dimer interface.[4,6] Detergents can occupy natural lipid sites under different experimental conditions.[7] For example, phosphatidylcholine in bovine CcO and the detergents decyl maltoside in *Rhodobacter sphaeroides* and lauryldimethylamine oxide in *Paracoccus denitrificans* CcO occupy the same crevices of the protein in different crystal structures.[7] Defining the determinants of lipid binding can help scientists understand the structural basis for the specificity of these sites and aid in the design of site-selective ligands and detergents for protein purification and structure determination.

Cholesterol (Figure 1A) plays an important role in the function of many biological systems, including eukaryotic, viral, and prokaryotic proteins. While cholesterol is often considered important because of its role in membrane organization, including lipid rafts,[8] cholesterol also exerts important regulatory effects via direct, specific binding to proteins. Through binding to the nicotinic acetylcholine receptor and many G protein-coupled receptors (GPCRs), cholesterol modifies the receptors' affinity for agonists.[9] Additionally, mutations in the cholesterol-binding sites of virus envelope proteins, such as the HIV protein gp41 and Semliki Forest virus E1 protein, inhibit virus invasion at the fusion and budding stages.[10] In addition, cholesterol binding by podocin and MEC-2, members of the prohibitin domain family, is essential for regulating the activity of their ion channel partners.[11]

A recent proteomic study mapped cholesterol-protein interactions in mammalian cells with photoreactive sterol probes, followed by quantitative mass spectrometry.[12] Their work identified over 250 cholesterol binding proteins, including some known to biosynthesize, transport, and regulate cholesterol, as well as others known to regulate sugars and glycerolipids or participate in vesicular transport and protein glycosylation and degradation.

Cholesterol-binding sequence motifs have been proposed for several protein families. For instance, a cholesterol consensus motif (CCM) has been identified in class A GPCRs as matching the amino acid sequence R/K-$(X)_{1-7}$-I/V/L-$(X)_{1-3}$-W/Y on one transmembrane $\alpha$ helix. The "strict CCM" also contains F/Y on a neighboring helix, based on residue conservation analysis between known cholesterol sites.[13] An expanded version of the CCM includes serine/glycine in one helix that forms an interhelical hydrogen bond with the CCM W/Y residue on an adjacent helix. The additional hydrogen bond is proposed to adjust the orientation of the aromatic side chain to enhance its stacking interactions with the steroid ring system.[14] A similar motif, the cholesterol recognition amino acid consensus or CRAC motif, has been defined in the outer

**Received:** October 29, 2014
**Published:** March 11, 2015





DOI: 10.1021/ci5006542
J. Chem. Inf. Model. 2015, 55, 747−759

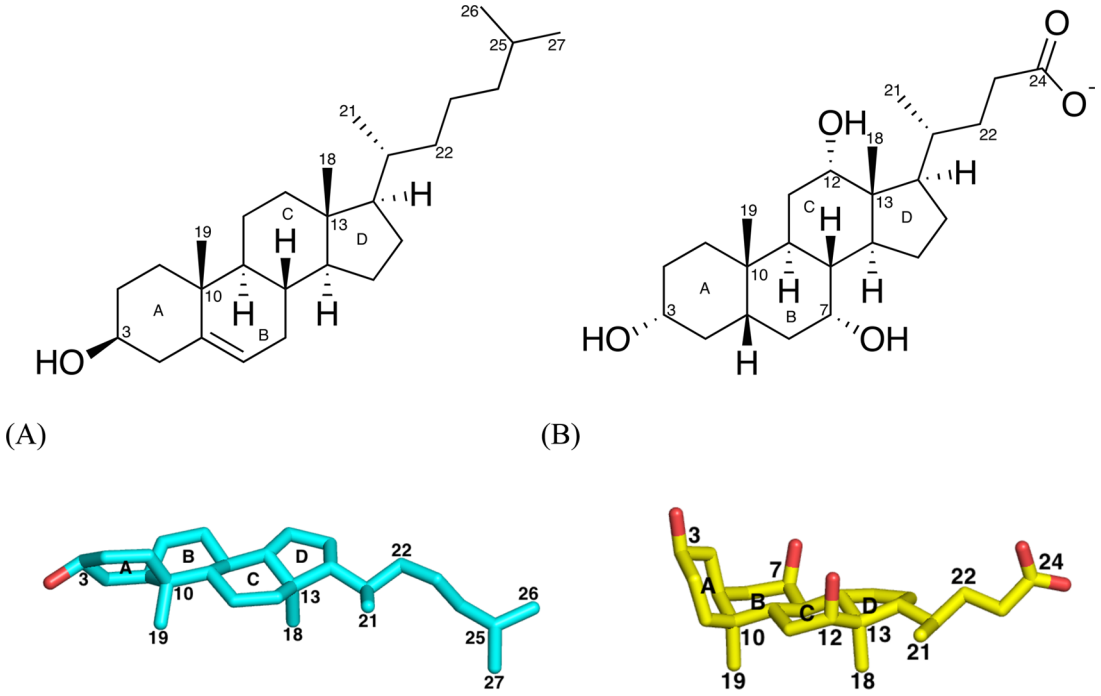



**Figure 1.** 2D and 3D chemical structures of (A) cholesterol (blue) and (B) cholate (yellow), with the flexible tails from C21 to C24/C25 shown in arbitrary favorable conformations.

mitochondrial membrane translocator protein (TSPO; also known as the peripheral benzodiazepine receptor). This consensus motif is L/V-(X)$_{1-5}$-Y-(X)$_{1-5}$-R/K, based on the loss of cholesterol uptake in TSPO Y153 and R156 mutants and alignment of this sequence region with other cholesterol binding proteins.[15,16] Recently, an enhanced version of the CRAC motif, LAF-CRAC, has been shown to be associated with nanomolar affinity for cholesterol in TSPO.[17] CARC, a cholesterol binding motif in the nicotinic acetylcholine receptor,[18] and a tilted peptide cholesterol binding motif have also been described.[19,20]

However, sequence motifs derived from one protein family often do not generalize well to predicting cholesterol-binding sites in other families, and these sequence motifs also match sites that do not bind cholesterol. For instance, analysis of 2 100 proteins in a bacterium that does not contain cholesterol found 5 000 matches to the CRAC motif.[21] Additional cholesterol binding sites are known that do not match any previously known motifs, for instance, the additional cholesterol sites known in some class A GPCRs. A GXXXG motif has been found to be critical for cholesterol binding to the $\beta$-amyloid precursor protein, as characterized by cholesterol titration and mutagenesis.[22] Cholesterol binding to this protein has been proposed to promote amyloidogenesis in Alzheimer's disease.[23] For cytolytic toxin recognition of cholesterol, a simple motif composed of a threonine-leucine pair in loop L1 has been identified by mutation analysis.[24] Thus, cholesterol binding sequence motifs appear to be fairly specific to protein families. Our aim is to uncover general features of cholesterol recognition that are shared by different protein families and which discriminate cholesterol binding sites from other ligand sites. These features can then be tested for their ability to capture a broader range of cholesterol-binding sites via application of the resulting predictor, CholMine.

Prediction of cholate binding sites also attracts our attention for several reasons. Cholate (Figure 1B) is used extensively as a membrane protein solubilizing detergent.[1,2] Crystal structures show cholate occupying binding pockets on membrane proteins, and this molecule shares significant similarity with cholesterol in shape and steroidal chemistry, aside from its dissimilar polar tail. Cholate, a bile acid, functions in some cells as a steroid hormone that binds to nuclear receptors to modulate gene expression.[25] Several soluble nuclear receptors have been reported to bind bile acids, including farnesoid X receptor (FXR), liver X receptor alpha, and cyclopentyladenosine receptor. The resulting complexes stimulate or suppress gene transcription by binding to promoter regions.[25] Cholate is also one of the two major bile acids synthesized from cholesterol and plays an essential role in the absorption of fat and lipidic vitamins, by forming micelles to solubilize fat.[26,27] Cholate has been shown to be an agonist for the human bile acid G protein coupled receptor TGR5, involved in suppression of macrophage function.[28,29] Lastly, a relative of cholate, 3-keto petromyzonol sulfate, acts as a vertebrate pheromone through interaction with two other GPCRs.[30] Thus, understanding the determinants of cholate binding and identifying features that distinguish between cholate and cholesterol sites will be useful for designing site-selective ligands and detergents for stabilizing and purifying membrane proteins and for interpreting ambiguous electron density in crystallography.

What is known about the determinants for protein interaction with lipids, in general? Four important factors can be summarized from the literature. The first is the presence of aromatic residues such as tryptophan (W), tyrosine (Y), and phenylalanine (F). Tryptophan and tyrosine are preferred at membrane interfaces.[31] In the Ballesteros-Weinstein numbering scheme to facilitate comparison of G-protein coupled receptors (GPCRs), residues are labeled by two indices, X.Y, the first indexing the transmembrane helix number in which the residue occurs, and the second indicating the position within the helix. The position number 50 is assigned to the most highly conserved position in each helix, with numbers increasing


DOI: 10.1021/ci5006542
J. Chem. Inf. Model. 2015, 55, 747−759

toward the C-terminus.[32] The Trp residue at position 4.50 in class A GPCRs, involved in cholesterol binding, is highly conserved (94%).[13] Aromatic residues contribute to cholesterol binding through favorable $\pi$ and hydrophobic interactions with the steroid ring system of cholesterol.[13] The second class of residues contributing to lipid binding includes the positively charged residues lysine (K), arginine (R), and histidine (H), which form electrostatic interactions with the polar or negatively charged head groups of lipids.[3,33] Uncharged polar residues such as serine (S), threonine (T), and cysteine (C) also contribute by forming hydrogen bonds with lipids (where cysteine acts as a weak hydrogen-bond acceptor).[3,33] The last class of residues involved in lipid binding includes the moderately bulky hydrophobic residues isoleucine, leucine, and valine (I, L, V), as found in the CCM and CRAC motifs. Position 6.57 in GPCRs is conserved with isoleucine and valine in adenosine receptors.[34] These residues form van der Waals interactions with the hydrophobic part of lipids, participate in stacking interactions, and form hydrophobic grooves for binding.[3,31,34]

Regions of lipid interaction have also been predicted using entire amino acid sequences, rather than motifs, along the lines of the transmembrane protein segment predictors that became popular in the 1980s. However, this type of prediction typically focuses on annotating membrane spanning regions of the protein sequence and does not provide information about pockets comprised of discontiguous parts of the protein that bind lipids tightly, the kind of lipid occupying each pocket or the chemical and spatial determinants of lipid specificity. For example, different categories of lipid-interacting proteins have been predicted, according to lipid degradation, metabolism, synthesis, transport, and other functions, by using amino acid sequence information from the SwissProt database.[35] In addition, residues involved in lipid binding have been predicted based on amino acid sequence and residue conservation using a support vector machine.[36] However, this approach does not provide spatial or lipid-specificity information that extends to new protein classes. Lipid-binding sites in several key cytoskeletal proteins have been predicted using a matrix-based algorithm to identify highly hydrophobic or amphipathic amino acid segments,[37] again predicting transmembrane secondary structure segments rather than pockets where lipids bind tightly and specifically. The goal of the work presented here is to identify the shared chemical determinants of cholesterol and cholate binding across nonhomologous protein sites and develop a sensitive and specific predictor for these sites.

## METHODS

Our identification of the determinants for cholesterol and cholate binding employs SimSite3D to align and quantify the similarity between pairs of binding sites.[38] The predictive accuracy is enhanced by incorporating knowledge of conserved interaction hotspots shared by cholesterol or cholate binding sites. In developing the CholMine predictor, we test the hypothesis that cholesterol (or cholate) binding in different proteins involves a characteristic set of interactions that distinguish cholesterol/cholate binding from other ligands.

**SimSite3D and Site Maps for Aligning and Comparing Protein Sites.** To align pairs of nonhomologous protein sites and find the relative orientation with maximum shape and chemical similarity in the absence of ligand information, we use SimSite3D.[38,41] This method aligns two protein sites based on their similarity in surface shape and chemical features, without requiring underlying sequence or structural similarity. For a given query site, the similarity to another site is measured in standard deviations relative to the query's mean score when aligned to all cases in a set of 140 ligand-binding sites (including one cholesterol site) chosen from proteins with undetectable sequence and structural homology to one another, representing a highly diverse set of ligand sites (Table S1 in the Supporting Information). This *Z*-score measures the statistical significance of a match. An alignment between two sites with a SimSite3D score less than −1.5 (in standard deviation units, where more negative values indicate greater similarity) results in 2 Å RMSD or better site alignment in 80% of cases, based on tests across pterin, adenine, peptide, and xenobotic binding sites from which the ligand has been removed.[38,41] SimSite3D alignment and scoring can also discriminate binding sites with similar chemical features that do not bind the same ligand. By contrast, other ligand site prediction methods either use information for both the ligand and receptor[39] or they only predict binding sites with high sequence similarity within certain protein families such as GPCRs.[40]

The site map representation used by SimSite3D is a set of chemically labeled points in three-dimensional space derived from residues in a user-defined or known ligand binding site. The site map represents a negative chemical image of the protein, indicating ideal positions for ligand atoms of a given chemistry to interact favorably with the protein. Each site map point can be related back to the corresponding protein atom(s). Hydrophobic site map points are set down discretely in a hemispherical array around hydrophobic protein atoms based on internal protein coordinates, such that two perfectly overlaid identical side chains will have exactly matching hydrophobic points, regardless of their initial Cartesian coordinates. Similarly, polar points are generated according to the favored geometry of hydrogen bonds relative to donor or acceptor groups in the protein (as is done for SLIDE docking templates[42]), with hydrogen-bond donor−acceptor atom interactions in the range of 2.5−3.5 Å, and the angle between the donor, hydrogen, and acceptor atoms falling between 120° and 180°. In SimSite3D, the matches of hydrogen-bonding groups are scaled according to the extent to which their hydrogen bonding vectors point in the same direction, based on the colinearity of (cosine of the angle between) their donor−acceptor vectors. Exact overlap (angle of 0°) yields a weight of 1 for the hydrogen bond match, and an angle of 90° yields a weight of 0. In the CholMine implementation, the boundaries of a site map are determined either by user specification of a set of residues comprising the cleft to be analyzed or by a set of ligand atom coordinates (which can be based on an experimentally determined or hypothesized ligand position that the user would like to assess). The ligand coordinates are then used to define a volume for site map generation, by selecting the set of protein residues containing at least one atom within 4.5 Å of one or more ligand atoms. SimSite3D reads ligand coordinates in Tripos mol2 format for site map generation. Ligand coordinates are converted from PDB format to mol2 format, as needed, by using the molcharge utility in QuacPac v. 1.3.1, utilizing OEChem toolkit v. 1.6.1 (OpenEye Scientific Software, Santa Fe, NM; http://www.eyesopen.com).

**Extraction of an Interaction Motif for Binding the Same Ligand in Nonhomologous Sites.** The goal of this work is to identify a motif that characterizes the binding of cholesterol (or cholate) across nonhomologous proteins. For


DOI: 10.1021/ci5006542
J. Chem. Inf. Model. 2015, 55, 747−759

**Table 1. Cholesterol Binding Proteins in the Training and Test Sets**

| Training Set: Membrane Proteins | | | | | |
|---|---|---|---|---|---|
| PDB code | ligand | source | resolution (Å) | *R*-factor | protein name |
| 2RH1 | cholesterol | *H. sapiens* | 2.4 | 0.198 | $\beta$2-adrenergic G protein-coupled receptor |
| 3AM6 | cholesterol | *A. acetabulum* | 3.2 | 0.290 | proton-pumping rhodopsin II |
| 2ZXE | cholesterol | *S. acanthias* | 2.4 | 0.248 | sodium−potassium pump |
| 3KDP | cholesterol | *S. scrofa* | 3.5 | 0.243 | sodium−potassium pump |
| 4DKL | cholesterol | *M. musculus* | 2.8 | 0.235 | $\mu$-opioid receptor |
| **Test Set: Soluble Proteins** | | | | | |
| PDB code | ligand | source | resolution (Å) | *R*-factor | protein name |
| 1LRI | cholesterol | *P. cryptogea* | 1.45 | 0.161 | beta-elicitin cryptogein |
| 1N83 | cholesterol | *H. sapiens* | 1.63 | 0.202 | retinoic acid-related orphan receptor alpha |
| 1ZHY | cholesterol | *S. cerevisiae* | 1.60 | 0.216 | KES1 protein |
| 3GKI | cholesterol | *H. sapiens* | 1.80 | 0.176 | Niemann-pick c1 protein |
| 3N9Y | cholesterol | *H. sapiens* | 2.10 | 0.207 | cholesterol side-chain cleavage enzyme (Cyp11A1) |

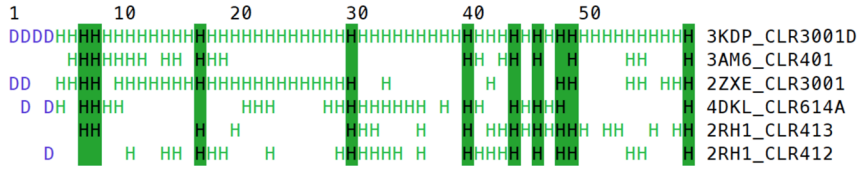



**Figure 2.** Determining conserved site map points. Aligned site map points with matching chemical labels from the training set of cholesterol (CLR) sites are shown following SimSite3D spatial alignment. Hydrophobic (H) or hydrogen-bond donor (D) site map points are shown on lines 2−6 if they fall within 1.5 Å of a site map point of the same chemical type in the query site, 3KDP_CLR3001D, where the number and letter after the CLR residue code indicate its residue number and chain identifier in the PDB file. Hydrogen-bond acceptor (A) and donor and/or acceptor (N) points (e.g., hydroxyl interaction sites) also occur in cholesterol sites but are not found to be conserved between the sites. The 3KDP query site was chosen as the representative query site for cholesterol binding because it has the highest degree of site map point conservation with the other cholesterol sites. Highly conserved points (green backgrounds) comprising the conserved motif for cholesterol interation were identified based on occurring in at least 70% of these training cases aligned to the 3KDP query site.

moderately to highly polar ligand sites, the SimSite3D score, which calculates the degree of chemical match between two sets of aligned site map points and their degree of molecular surface shape similarity, is usually sufficient to filter out false positive site matches while aligning and detecting most of the true positive sites. However, cholesterol sites are unusually hydrophobic, and the degree of conservation of polar interactions between nonhomologous cholesterol sites is low, particularly because crystal structures show that the cholesterol hydroxyl moiety is often exposed to bulk water rather than interacting directly with protein atoms. As a result, CholMine employs SimSite3D to align and score a pair of site maps and then determines whether this alignment matches a majority of conserved points of hydrophobic interaction identified from known cholesterol binding sites. Table 1 lists protein structures containing the 12 low-homology cholesterol sites, which were divided into two sets: the first set for training to detect conserved points of cholesterol interaction and the second set for unbiased testing of cholesterol site predictions on a series of unrelated proteins. The cholesterol sites from dogfish and pig sodium−potassium pump proteins (PDB entries 2ZXE and 3KDP) were both included in the training set because their cholesterol binding residues were in different conformations. The number of independently determined, well-resolved, nonhomologous cholesterol binding sites in the Protein Data Bank is limited, likely due to the difficulty in handling this ligand, which has extremely low aqueous solubility. However, including several cholesterol sites from the same protein family would bias toward identifying a family specific motif, whereas the goal here is to discover the chemical determinants of cholesterol binding sites in general. Therefore, we tested the extent to which the cholesterol binding motif determined from the training set cases can predict cholesterol sites well in other proteins, including the nonhomologous cholesterol binding sites in the test set, a series of cholesterol-binding class A GPCR structures showing sequence and conformational diversity, a set of noncholesterol steroid binding sites, a set of aliphatic lipid binding sites, a set of 109 bacterial membrane proteins that do not contain cholesterol binding sites, and 139 soluble protein sites known to bind ligands other than cholesterol. Including only membrane protein cholesterol binding sites in the training set and only soluble sites in the training set (and then inverting the sets) allowed us to further test whether cholesterol binding motifs are similar in these different cellular environments.

To determine the conserved cholesterol contacts shared by diverse binding sites, CholMine employs the binary string output of SimSite3D (Figure 2), representing spatially aligned SimSite3D interaction points. Once a set of known cholesterol or cholate training sites has been aligned by SimSite3D based on matching the three-dimensional site map points and the surface shape derived from protein atom coordinates alone, the software determines which site map points overlay in three-dimensional space and have the same chemical interaction type (are conserved between the sites). The most highly conserved interaction points can then serve as a fingerprint, or filter, that aids in recognizing cholesterol sites.

The determination of conserved interaction points can be conceptualized as a matrix of SimSite3D-aligned site map points (Figure 2) indexed relative to the points they match spatially in the representative site, which is the site with the highest degree of interaction point conservation with the other


DOI: 10.1021/ci5006542
J. Chem. Inf. Model. 2015, 55, 747−759

**Table 2. Cholate Binding Proteins in the Training and Test Sets**

| Training Set: Mixture of Membrane and Soluble Proteins | | | | | |
|---|---|---|---|---|---|
| PDB ID[α] | ligand | source | resolution (Å) | *R*-factor | protein name |
| Δ 1EE2 | cholate | *E. caballus* | 1.5 | 0.148 | alcohol dehydrogenase |
| Δ 1S9Q | cholate | *M. musculus* | 2.2 | 0.220 | estrogen-related receptor gamma |
| Δ 2AZY | cholate | *S. scrofa* | 1.9 | 0.167 | phospholipase A2 |
| Δ 2DQY | cholate | *H. sapiens* | 3.0 | 0.226 | liver carboxylesterase 1 |
| ^ 2DYR | cholate | *B. taurus* | 1.8 | 0.202 | cytochrome c oxidase |
| Δ 2HRC | cholate | *H. sapiens* | 1.7 | 0.221 | ferrochelatase |
| Test Set: Soluble Proteins | | | | | |
| PDB ID[α] | ligand | source | resolution (Å) | *R*-factor | protein name |
| Δ 1TW4 | cholate | *G. gallus* | 2.0 | 0.216 | liver bile acid binding protein |
| Δ 2FT9 | cholate | *A. mexicanum* | 2.5 | 0.260 | liver bile acid-binding protein |
| Δ 2QO4 | cholate | *D. rerio* | 1.5 | 0.188 | liver bile acid-binding protein |
| Δ 2RLC | cholate | *C. perfringens* | 1.8 | 0.195 | choloylglycine hydrolase |
| Δ 3ELZ | cholate | *D. rerio* | 2.2 | 0.224 | ileal bile acid-binding protein |
| Δ 3QPS | cholate | *C. jejuni* | 2.4 | 0.204 | CmeR |

[α]Membrane proteins are indicated by ^ and soluble proteins by Δ. In PDB structures 2DYR, 2HRC, 1TW4, 2FT9, and 3ELZ, two or more independent cholate binding sites were included in training or testing.

cholesterol sites. This procedure results in the unbiased detection of a three-dimensional binding motif corresponding to shared interactions in nonhomologous sites binding cholesterol, as indicated by the highlighted vertical green bars showing points of interaction common to 70% or more of the sites (Figure 2).

**Establishing a Cholate Site Predictor.** Creating a cholate site predictor for the CholMine software followed the same process as for cholesterol prediction. The first step was to set up the training and test databases. Twenty cholate (PDB residue name CHD) binding sites in 12 nonredundant proteins were used to generate SimSite3D site maps representing points of favorable hydrophobic or hydrogen-bond interactions with cholate (Table 2). These 20 cholate binding sites were divided into two data sets of equal size. There were just four nonhomologous membrane protein-bound cholate sites in the PDB, representing limited training power, with the 16 other cholate sites coming from soluble proteins. The training set thus included the 4 membrane protein cholate sites and 6 of the soluble cholate sites. There were no instances of cholate sites repeated (even with low homology) between the training and test sets, to guarantee that the test predictions would be unbiased. Because of the limited availability of unrelated cholate sites in the PDB, four bile acid binding proteins with moderate pairwise sequence identity (~60%) were included in the test set. Inverting the two sets in testing and training then allowed testing whether a more diverse set of cholate sites (the first set, with a mixture of unrelated membrane and soluble sites) or a set of sites with some similarity (from four diverse bile acid binding proteins and two unrelated proteins) provided greater cholate site detection power.

**Summary of the Steps for Establishing a Cholesterol (or Cholate) Site Predictor.** *Step 1: Preparing the Training and Testing Databases.* The binding sites divided into training and test sets were processed by SimSite3D to create site maps. Sets of soluble and membrane proteins containing diverse or lipid ligands (as described in the section above, SimSite3D and Site Maps for Aligning and Comparing Protein Sites and in Bacterial Membrane Proteins for Evaluating False Positive Prediction Rate, below) were also prepared as site maps for alignment and comparison as negative controls, to assess the rate of false positive predictions.

*Step 2: Choosing the Most Representative Cholesterol (or Cholate) Binding Site.* The goal of this step was to select the known site with the best SimSite3D scoring detection and quality of alignment with other cholesterol (or cholate) binding sites (as described for the site from PDB entry 3KDP in Figure 2). For cholesterol sites, the membrane set was initially assigned as the training set, the soluble set as a true positive test set, and the diverse ligand sites as a data set with one true positive buried in many false positive cases. The SimSite3D normalized score threshold was set to 0.0 (keeping the best scoring orientation of any site that aligns favorably with the query site), and each of the 12 cholesterol sites was compared against all the others and to the diverse set of 140 binding sites. The RMSD value representing the closeness of alignment (with 0 Å representing a perfect alignment) between the query site cholesterol atom positions and those in the aligned ligand sites was calculated by using the RMSD function in the OEchem toolkit v.1.6.1 (http://www.eyesopen.com; OpenEye Scientific Software, Santa Fe, NM). Assigning one query site from the training set and a separate query site from the test site allowed the two sets to be inverted for training and testing. The same procedure was followed for cholate sites.

*Step 3: Extracting a Fingerprint of Conserved Interactions from Known Cholesterol (Or Cholate) Sites and Applying It to Predict on the Test Set.* A high false positive rate results when SimSite3D alone is used to align hydrophobic sites with a generous scoring threshold, due to significant hydrophobic contact scores and the absence of directional hydrogen-bonding group matches (which are strong discriminants for polar sites binding the same ligand). This motivated our development of a way to pinpoint additional conserved features of cholesterol or cholate binding sites. Conserved hydrophobic interactions were identified between the cholesterol sites, based on site map points that overlaid in three-dimensional space, as shown in Figure 2, for both the training and test sets. These points represent hydrophobic positions in the cholesterol sites that are ≥70% conserved with respect to the query site for the membrane (3KDP_CLR3001D) or soluble set (1ZHY_CLR1001A). The conserved points and their relative positions in


DOI: 10.1021/ci5006542
J. Chem. Inf. Model. 2015, 55, 747−759

space provide a shared recognition motif or fingerprint for cholesterol interaction that is implemented as a filter (following SimSite3D alignment) in the CholMine predictor. A test site is predicted to bind cholesterol or cholate if, upon three-dimensional site map alignment with the query site, it matches at least 70% of the conserved points. The same procedure was followed for identifying and applying a conserved recognition motif for the cholate training and test sites.

**Bacterial Membrane Proteins for Evaluating False Positive Prediction Rate.** Bacteria contain no cholate or cholesterol and are thus likely to provide a rigorous set of ligand sites to test for the rate of false positive cholesterol predictions because their membrane-exposed surfaces are hydrophobic and interact with other lipids. PDB codes of bacterial membrane proteins were extracted from the Membrane Proteins of Known 3D Structure Database (http://blanco.biomol.uci.edu/mpstruc/) and then entered in the Pisces server[43] (http://dunbrack.fccc.edu/Guoli/PISCES_InputB.php) to select a low-homology set of bacterial membrane proteins using default criteria: crystal structures with ≤25% pairwise sequence identity, ≤ 3.0 Å resolution, *R*-value ≤0.3, and chain length between 40 and 10 000 residues.

**CholMine Server.** The overall steps in cholesterol/cholate site prediction by CholMine are summarized in Figure 3. A

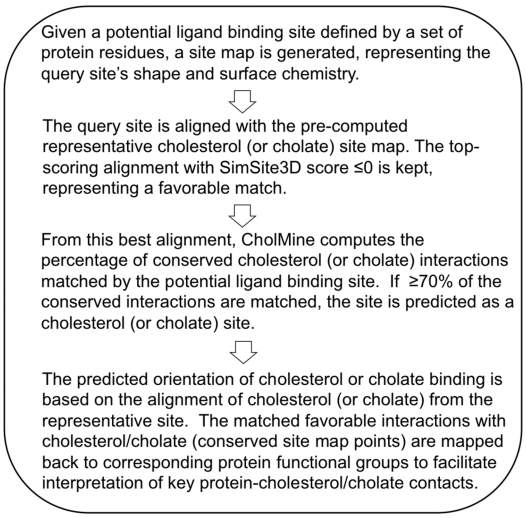



**Figure 3.** Steps in CholMine cholesterol and cholate site prediction.

Web server implementation has been established to support automated prediction of cholesterol and cholate binding sites by users for their own protein structures (http://cholmine.bmb.msu.edu). Given a Protein Data Bank file and a ligand residue number and ligand chain ID for a placemarker ligand in the site, the server will provide the following information: a prediction of whether the site binds cholesterol or cholate; the predicted binding mode of the corresponding steroid; and the residues in the binding site forming conserved interactions with cholesterol or cholate. A prediction summary plus PDB files containing the ligand orientation and essential residues are e-mailed to the user, with an option to also provide a preformatted PyMOL molecular graphics file (Schrödinger, New York, NY; http://pymol.org) showing the predicted interactions. The set of key protein interactions can be used to design experiments that probe ligand binding, for instance by site-directed mutagenesis.

As well as supporting the use of a placeholder ligand (e.g., a crystallographic lipid or user-defined dummy residue) to define the binding site volume to analyze, the server also supports user uploading of a mini PDB file that contains up to 25 residues defining the protein region the user would like to assess for cholesterol or cholate binding. This set of residues is used to define the potential ligand binding site volume as a box bounded by the minimum and maximum *x*, *y*, and *z* coordinates of the residues provided. The volume for site map generation is then refined by placing probes on a 1.0 Å grid in the box and removing any probes within 3.5 Å (van der Waals contact distance) of protein atoms. The site map for CholMine analysis is generated within this volume for comparison to the conserved interaction points characteristic of cholesterol or cholate binding. In the server implementation, 10 000 Å$^3$ was set as the maximum box volume.

## RESULTS

**Cholesterol Binding Site Training and Testing.** Of all the membrane cholesterol sites, 3KDP_CLR3001D gave the lowest average RMSD of alignment against the other membrane sites in the training set when used as the query (Figure 4A), so the site map and positions and chemistry of conserved interactions in this site were used as the basis to align and score the test cases. As shown in Figure 4B, 1ZHY_CLR1001A gave the lowest average RMSD when used as the query for alignment of the set of soluble cholesterol sites. Thus, this site was chosen as the soluble site representative query when the training and test sets were inverted to determine which query had the greatest predictive power and lowest false positive rate.

As shown in Table 3, using the 3KDP_CLR3001D site as the query (where CLR is the residue name for cholesterol and 3001D is the ligand residue number), combined with requiring at least 70% of its conserved interactions to be matched for a site to be predicted as cholesterol binding resulted in prediction of 83% of the membrane protein cholesterol sites (training set) and 80% of the soluble protein cholesterol sites (true positives in the unbiased test set), with a relatively low rate (5%) of false positives in the 140-site diverse data set. Self-prediction of a site (when used as both the query site and as a data set entry) is not included in the calculation of the true positive rate, since self-prediction is guaranteed. In contrast, although the soluble cholesterol site 1ZHY_CLR1001A has a low false positive rate when at least 70% of its conserved interactions are matched, it fails to find any of the membrane protein cholesterol binding sites, while predicting 75% of the soluble sites. These results suggest that the membrane cholesterol sites share a conserved motif that is also part of the soluble site recognition of cholesterol. However, additional shared interactions within the soluble sites are not well-matched by the membrane sites, likely due to the fact that soluble proteins more fully surround and sequester cholesterol. On the basis of its superior performance on soluble as well as membrane cholesterol binding sites, the 3KDP query site and its conserved set of interactions were implemented in the CholMine server for cholesterol site detection.

**Cholate Site Training and Testing.** SimSite3D pairwise comparison of the cholate sites for the two data sets is shown in Figure 5, allowing the identification of the query site within each set that could best detect other cholate sites based on the lowest average RMSD of alignment over the most sites. The membrane protein site representative (2DYR_CHD525C)


DOI: 10.1021/ci5006542
J. Chem. Inf. Model. 2015, 55, 747−759

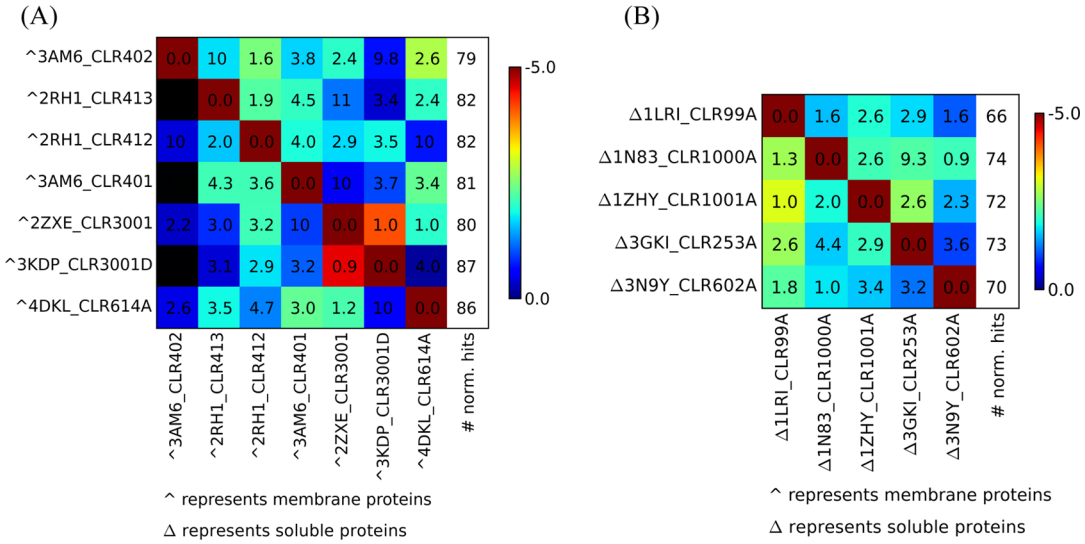



**Figure 4.** Pairwise alignment and similarity scoring. (A) All-against-all SimSite3D comparison for membrane protein cholesterol binding sites. (B) All-against-all comparison for soluble protein cholesterol binding sites. For the top-scoring alignment of each site pair, the SimSite3D similarity score values are colored from red (most similar) to dark blue (marginally similar) with corresponding score values ranging from −5 to 0 (in standard deviations above the mean score when the same query site is compared to the set of 140 diverse ligand binding sites, where more negative is more significant). Black indicates failure to meet the normalized score threshold of 0. Numbers reported in the grid are the RMSD values (Å) between cholesterol rings following SimSite3D site alignment. Lower RMSD indicates better alignment between sites. The "# norm. hits" column on the right side of each matrix reports the number of sites meeting the scoring threshold for similarity to the query site (labeled to the left in each row) when searching against the 140 sites in the diverse data set (Table S1 in the Supporting Information), which includes one true positive cholesterol site. The high number of false positives is based on SimSite3D alignment score only, before the conserved interaction points for cholesterol sites have been considered.

**Table 3. Prediction Results for Using Cholesterol Sites in 3KDP_CLR3001D (a membrane protein) and 1ZHY_CLR1001A (a soluble protein) for Detecting Cholesterol Sites in Other Proteins, Plus Assessment of False Positives in a Set of 139 Non-Cholesterol Ligand Sites[a]**

| query ID | true positive rate for training data set | unbiased true positive rate for test data set | false positive rate for diverse data set |
|---|---|---|---|
| 3KDP_CLR3001D | 5/6 (83%) | 4/5(80%) | 7/139 (5%) |
| 1ZHY_CLR1001A | 3/4 (75%) | 0 | 2/139 (1.4%) |

[a]When 1ZHY_CLR1001A was used as the query in the results above, the training and test sets were inverted relative to those listed in Table 1. Query self-matches were excluded from the statistics.

provided better predictive ability overall (Table 4). Predicting cholate sites as those matching at least 70% of the conserved interactions in this query site gave a true positive rate of 67% for cholate sites in the training set, a true positive rate of 70% for cholates in the unbiased test set, and a false positive rate of 12% on the set of 140 diverse ligand binding sites. 2QO4_CHD130A was identified as the best representative of the second, entirely soluble cholate site data set. When this site was used as the query to find cholate sites matching its conserved interactions, a true positive rate of 67% was observed in the entirely soluble cholate site set, a true positive rate of only 10% in the mixed membrane/soluble protein set, and a false positive rate of 1.4% when applied to the set of 140 diverse cholate sites. The decreased generalization of the soluble site query and conserved points for predicting other cholate sites was expected, since a substantial number of sites in this set came from two sites in diverse members of the $\beta$-clamshell bile acid binding protein family. Similarly, by being a more family specific motif, this query's lower false positive rate was expected on the diverse set of 140 noncholate binding sites. The membrane cholate site query performed better as a cholate site predictor that generalizes across protein families, with 3 times as many true positive predictions as the soluble site query (Table 4). Therefore, cholate site prediction in CholMine uses 2DYR_CHD525C as the query, combined with conserved interactions derived from the first data set of mixed membrane and soluble protein cholate sites.

**Evaluating the Statistical Significance of the Cholesterol and Cholate Site Predictors.** The lift value is a common way to evaluate models in data mining, reflecting the enhancement in predictivity relative to random selection.[44] Suppose the predictor rule is that A implies B (e.g., a positive prediction by CholMine implies that the site binds cholesterol). The lift value for CholMine predictions can be calculated as

$$\mathrm{Lift}(A \Rightarrow B) = \frac{P(B|A)}{P(B)} = \frac{P(A \cap B)}{P(A)P(B)}$$

$\mathrm{Lift}(A \Rightarrow B) > 1$ means A and B have a positive relationship, and the numeric value reflects the $n$-fold enhancement of the predictive rate (how many times higher?) relative to random prediction. $\mathrm{Lift}(A \Rightarrow B) = 1$ indicates that A and B are independent, and $\mathrm{Lift}(A \Rightarrow B) < 1$ means A and B have an inverse relationship. The chi-squared test can also be used to evaluate whether the correlation between A and B is statistically significant, by measuring the probability of there being a significant difference between the predicted versus actual result (e.g., the presence of a cholesterol binding site). For CholMine cholesterol site prediction, the lift value was 7.7, indicating CholMine is almost 8 times as effective as random prediction of cholesterol sites. The very small chi-squared $P$-value of $1.05 \times 10^{-13}$ indicates significant correlation between CholMine


DOI: 10.1021/ci5006542
J. Chem. Inf. Model. 2015, 55, 747−759

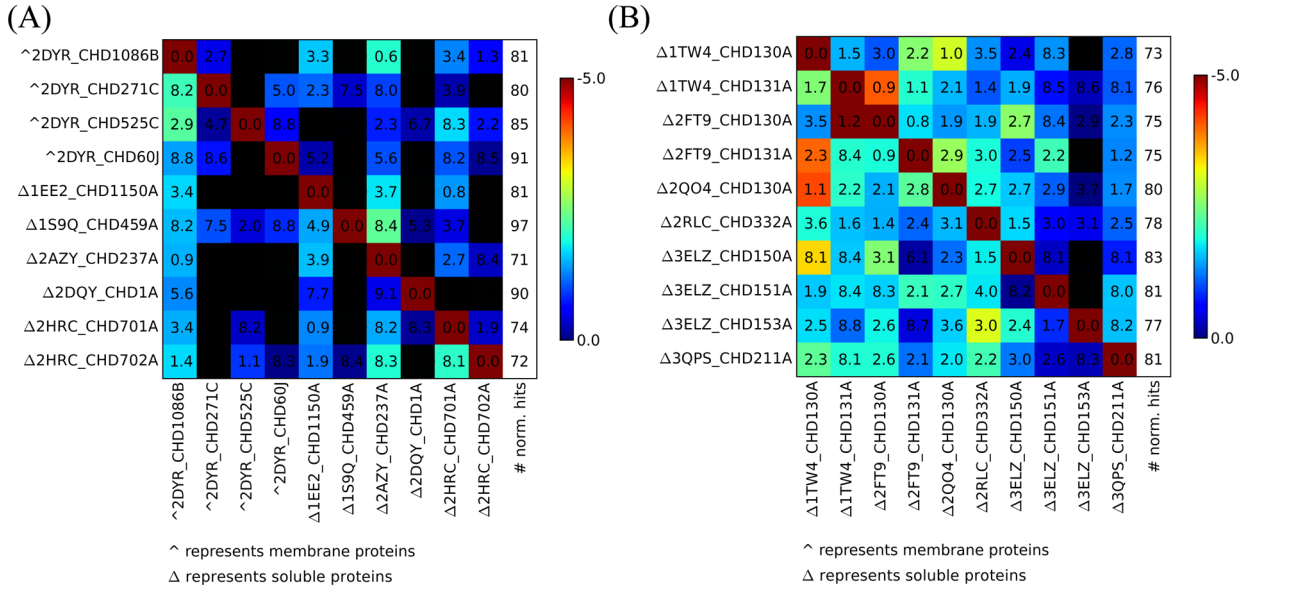



**Figure 5.** Pairwise alignment and similarity scoring. (A) All-against-all SimSite3D similarity comparison for the first data set, which includes 4 membrane cholate binding sites and 6 soluble cholate binding sites. (B) All-against-all comparison for the second data set, which includes another 10 soluble cholate binding sites unrelated to the first set (See Figure 4 legend for additional details.).

**Table 4. Prediction Results from Using Cholate Sites 2DYR_CHD525C (best representative from a membrane protein) and 2QO4_CHD130A (best representative from a soluble protein in the second set) for Alignment and Scoring to Predict Cholate Binding Sites in Other Proteins and Assess False Positive Rate in a Set of 140 Non-Cholate Sites[a]**

| query ID | true positive rate for training data set | unbiased true positive rate for test data set | false positive rate for diverse data set |
|---|---|---|---|
| 2DYR_CHD525C | 6/9 (67%) | 7/10 (70%) | 17/140 (12%) |
| 2QO4_CHD130A | 6/9 (67%) | 1/10 (10%) | 2/140 (1.4%) |

[a]Query self-matches were excluded from the results. The training and test sets were inverted relative to Table 2 when the 2QO4 query was used.

prediction and cholesterol binding. For CholMine prediction of cholate sites, the lift value is also significant (3.6), with a very small chi-squared *P*-value of $2.53 \times 10^{-8}$.

**GPCR Cholesterol Binding Site Prediction.** Putative cholesterol sites in class A GPCRs were analyzed as one way of testing the predictive ability of CholMine on additional cholesterol sites. The consensus motif (CCM) found in the cholesterol-binding site of human $\beta$2-adrenergic receptor (labeled as residue 412 in PDB code, 2RH1) is matched by the sequences in 44% of human class A G protein coupled receptors.[13] To assess the ability of CholMine to find sites matching the sequence-based consensus motif, prediction was performed on the structures available for 11 of these receptors (PDB codes: 3EML, 3PBL, 2KS9, 2Y00, 3RZE, 1U19, 2Z73, 3ODU, 3V2W, 3UON, and 4DJH; Table S2 in the Supporting Information). A total of 82% of these proteins were predicted by CholMine to bind cholesterol in the region corresponding to cholesterol 412 in PDB entry 2RH1, in PDB entries 3EML, 2KS9, 2Y00, 3RZE, 1U19, 3ODU, 3V2W, and 3UON. In addition, for the 1.8 Å resolution crystal structure of the human A2a adenosine receptor (PDB entry 4EIY), which contains 3 cholesterol-bound sites unrelated to each other by symmetry or amino acid sequence, two of the three sites were predicted by CholMine (labeled as residues 404 and 405 in PDB entry 4EIY).

**Comparison of CholMine Structure-Based Predictions with Sequence-Based Predictions Using the CCM, CRAC, and GXXXG Motifs.** To compare the predictive ability of previously published cholesterol binding sequence motifs with that of CholMine, Sequery[45] was applied to identify sequences matching each motif in crystal structures of the same proteins used for CholMine prediction (Table 1 and Tables S1 and S2 in the Supporting Information). Matching the CCM, CRAC, and GXXXG sequence motifs predicted the membrane protein cholesterol binding sites well (80−100% of these sites were predicted), predicted soluble sites less well (40−80%), and resulted in an unacceptable rate of false positives in the diverse data set: 100 or more cholesterol sites were predicted in 139 sites known to bind a different ligand (Table 5).

**Table 5. Comparison of Cholesterol Site Prediction in True versus Non-Cholesterol Binding Sites by the CholMine Conserved Spatial Motif versus Sequence Motif Matching**

| | relaxed CCM[α] | CCM[α] | CCM + surface accessibility | CRAC[α] | GXXXG[α] | CholMine predictor |
|---|---|---|---|---|---|---|
| membrane set | 5/5 (100%) | 4/5 (80%) | 2/5 (40%) | 5/5 (100%) | 4/5 (80%) | 5/6 (83%) |
| soluble set | 4/5 (80%) | 2/5 (40%) | 1/5 (20%) | 3/5 (60%) | 3/5 (60%) | 4/5 (80%) |
| GPCRs | 11/11 (100%) | 10/11 (91%) | 6/11 (54%) | 11/11 (100%) | 5/11 (45%) | 9/11 (82%) |
| diverse data set (false positives) | 130/139 (94%) | 105/139 (75%) | 33/139 (24%) | 116/139 (83%) | 100/139 (72%) | 7/139 (5%) |

[α]Relaxed CCM, R/K-(X)$_{1-7}$-I/V/L-(X)$_{1-3}$-W/Y;[3,13] CCM, R/K-(X)$_{2-6}$-I/V/L-(X)$_3$-W/Y;[13] CRAC, L/V-(X)$_{1-5}$-Y-(X)$_{1-5}$-R/K;[15−17] G(X)$_3$G.[22]



DOI: 10.1021/ci5006542
J. Chem. Inf. Model. 2015, 55, 747−759

One of the problems with sequence motif based prediction is that it does not assess the surface accessibility of the motif, which is required for cholesterol to access the site. To test whether including solvent accessibility as an additional criterion for sequence motif-based cholesterol site prediction can solve the overprediction problem, a solvent accessible surface threshold was set at 29 Å$^2$ for matching each residue in the CCM motif, corresponding to the minimum exposed surface area per residue in the cholesterol site of human $\beta_2$ adrenergic receptor (PDB entry: 2RH1). The results show that the true positive rate for membrane protein cholesterol sites decreased from 80% to 40%, for soluble protein sites from 40% to 20%, and for GPCRs from 91% to 54% (Table 5, CCM + Surface Accessibility column). The false positive rate decreased from 75% to 24%, while still resulting in 33 false positives in 139 proteins. Overall, even when surface accessibility is considered, sequence motif prediction has an unacceptably high false positive rate for cholesterol prediction (24%) and a moderate rate of true positive prediction (20−40%), whereas CholMine structure-based prediction results in few false positives (5%) and a high true positive rate (80−83%).

**Deciphering the Determinants of Cholesterol Binding.** For cholesterol binding site prediction in membrane proteins, all the conserved site map points representing favorable cholesterol contacts derive from hydrophobic groups, more specifically, Ile D35, Leu D36, Tyr D39, Tyr D43, Glu C840, Ile C843, Tyr C847, and Met C852 in the representative query site, 3KDP_CLR3001D (Figures 2 and 6A). A smaller but similar set of interactions with cholesterol at this site is identified when the single 3KDP crystal structure is analyzed by LigPlot and LigPlot$^+$[46,47] (Figure 6B,C). Compared with the CCM (R/K-$(X)_{1-7}$-I/V/L-$(X)_{1-3}$-W/Y) and CRAC (L/V-$(X)_{1-5}$-Y-$(X)_{1-5}$-R/K) motifs, the CholMine spatially conserved binding motif exemplified by this site contains an I-L-$(X)_2$-Y motif, which matches the residues at the end of the CCM and the beginning of the CRAC motif. CholMine's conserved interaction points surround atoms on the steroid ring observed to have the highest frequency of protein interaction (Figure 6A). There may be several reasons for the observed lack of conserved polar interactions with cholesterol. First, there is only a single polar group, the A-ring hydroxyl substituent, in cholesterol. In seven cholesterol sites evaluated (two sites in 2RH1 and 3AM6 and one each in 2ZXE, 3KDP, and 4DKL), there was only a single direct protein hydrogen bond to the cholesterol hydroxyl group, with water-mediated interactions to cholesterol in another structure, and no protein hydrogen bonds to the cholesterol hydroxyl group observed in any of the other cases. This suggests that the hydroxyl group may help position cholesterol correctly at the interface between the lipid bilayer and bulk solvent, rather than being a recognition determinant for binding to proteins. Also supportive of a lesser role for polar group recognition is the observation that the arginine or lysine residue in the CCM is only 22% conserved in class A GPCRs; thus interactions of this residue with cholesterol are only mildly conserved.[13]

In soluble protein cholesterol binding sites, both faces of cholesterol are surrounded in the pocket, forming additional interactions with the protein. However, the conserved interaction points from soluble protein cholesterol binding sites perform less well than those from membrane proteins in predicting cholesterol sites in general (Table 3). The conserved membrane protein cholesterol interactions (Figure 6A) can predict and are characteristic of both membrane and soluble

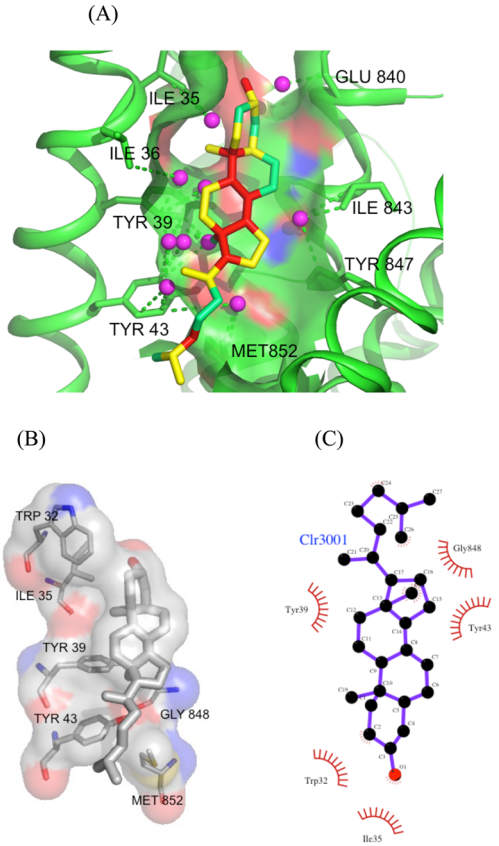



**Figure 6.** (A) Sodium/potassium-transporting ATPase cholesterol site (PDB entry 3KDP, residue D3001) used as the representative query for CholMine predictions. Purple spheres represent conserved interaction points in the membrane proteins binding cholesterol (from Figure 2), displayed in the context of the representative site from 3KDP. The green dashed lines connect the conserved interaction points to corresponding protein atoms. Cholesterol atoms colored in green contact a protein atom in >60% of the training set sites, atoms colored yellow have a 30−60% frequency of contact, and atoms colored in red contact the protein in <30% of the sites. (B) For comparison, LigPlot$^+$ three-dimensional view (shown with PyMOL; Schrödinger, New York, NY; http://pymol.org) of key sodium/potassium-transporting ATPase cholesterol interactions identified in just the single structure of 3KDP. (C) Alternative LigPlot two-dimensional view of these interactions.

sites in unrelated proteins and are the basis for CholMine cholesterol site prediction.

**CholMine Distinguishes Cholesterol Sites from Sites Occupied by Acyl Chain Lipids.** CholMine was also applied to diverse lipid binding sites: the 22 independent acyl lipid sites in the adenosine receptor (PDB code, 4EIY) and five phosphatidylethanolamine and analogue sites in PDB entries 3DDL, 2Z73, 3UTW, 3UTV (Table S3 in the Supporting Information). CholMine correctly predicted that 21 out of 22 sites in the adenosine receptor do not bind cholesterol and the same for all five of the phosphatidylethanolamine sites.

**Discriminating Cholesterol and Cholate Sites from Other Steroid Sites.** To test whether CholMine can distinguish cholesterol sites from steroid binding sites in general, a variety of nonhomologous crystal structures were tested: the progesterone sites in PDB entries 1A28, 2AA6, 2BAB, and 2HZQ, the estradiol sites in 1AQU, 1E6W, 1JGL, 1LHU, and 3OLL, and the testosterone sites in 2AM9, 1J96, and 3KDM (Table S3 in the Supporting Information). A total


DOI: 10.1021/ci5006542
J. Chem. Inf. Model. 2015, 55, 747−759

of 10 out of the 12 sites were predicted as noncholesterol sites, with two false positives, in 1AQU and 1J96. The cholesterol site predictor was also applied to the cholate training and test sets (Table 2) and vice versa (Table 1). The cholesterol site predictor predicts 30% of the training and 30% of the test set of cholate sites. The cholate site predictor predicts 57% of the membrane cholesterol sites and 80% of the soluble sites. Thus, cholesterol and cholate sites are harder to discriminate than cholesterol and steroid sites in general, and again we see a higher level of discrimination of cholesterol relative to cholate sites. Reasons for this are discussed below in the section below, Comparison of Cholesterol and Cholate Binding Site Conservation.

**Bacterial Membrane Proteins for Evaluating False Positive Predictions.** Bacteria contain no cholate or cholesterol. Thus, known ligand sites, mostly lipid-binding, were analyzed in 109 low-homology bacterial membrane protein structures (Table S4 in the Supporting Information) as an additional stringent test of the false positive rate for cholesterol and cholate site prediction. A total of 11 of the 109 sites, or 10%, were falsely predicted as potential cholesterol sites. When analyzed as potential cholate sites, 14 (13%) of the sites were predicted. Though nominally these are false positives, eubacteria are known to contain sterol-like molecules including cyclic hopanoids, tetrahymanol, and squalene.[48,49] Thus, it remains possible that some sites that were occupied by unnatural molecules in the bacterial crystal structures may natively bind sterol-like molecules.

**Cholate Binding Determinants.** Cholate is an important detergent for membrane proteins and also a representative of bile acids that act as hormones, pheromones, and important metabolites of cholesterol. CholMine was trained for cholate site prediction similarly to the protocol for cholesterol, and the determinants for cholate binding in membrane proteins were found to differ somewhat from those in soluble proteins. For membrane protein cholate binding sites, the conserved interaction points were all hydrophobic. In the representative 2DYR_CHD525C (cytochrome c oxidase) site used for CholMine prediction, these interactions arise from TrpC99, HisA233, TrpA288, TyrA304A, and PheA305 (Figure 7). The latter trio of residues serve to anchor cholate in the binding pocket. Out of the 10 training set cholate molecules, half of the O3 hydroxyl groups (on the A ring of cholate) formed water-mediated and two formed direct hydrogen bonds to the protein. The O7 and O12 hydroxyls (on the B and C rings) formed fewer hydrogen bonds to protein: two O7 and four O12 water-mediated hydrogen bonds were observed, and 1 direct hydrogen bond was found in the 10 sites, with a low degree of conservation. The tail carboxylate oxygens formed 7 direct H-bonds overall, which were spatially varied in position.

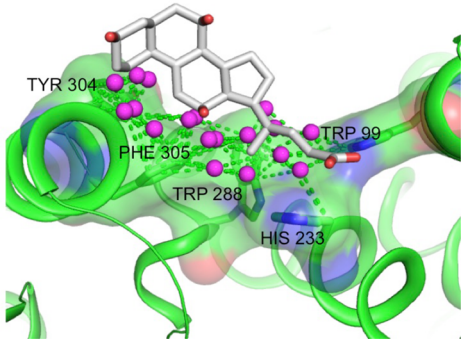



**Figure 7.** Conserved interaction points for CholMine cholate site prediction (purple spheres) are shown in the context of the interactions between the representative membrane protein query site 2DYR_CHD525C from cytochrome c oxidase and its bound cholate molecule (white tubes with oxygen atoms in red). Essential residues contributing to the conserved interaction are labeled.

**Comparison of Cholesterol and Cholate Binding Site Conservation.** To understand why the number of conserved interaction points is greater for cholate sites (Figure 7) compared with cholesterol (Figure 6), the crystallographic mobility of atoms in these ligands was compared. In the training set of 10 cholate sites, the crystallographic *B*-factor average for cholate atoms was 48 Å$^2$, whereas in the training set of 7 cholesterol sites, the *B*-factor average for cholesterol atoms was 1.5 times as high (74 Å$^2$), reflecting significant mobility. Higher atomic mobility is thus likely the reason for fewer spatially conserved interactions in cholesterol sites.

A generally similar pattern is seen in the edges and faces of cholate and cholesterol that predominate in forming conserved interactions with protein sites (Figure 8). Discrimination between cholesterol and cholate binding is not via polar interactions (which are not conserved across cholate or cholesterol sites), but by conserved interactions at the bend between the steroid A and B rings and near the center of the tail in cholate versus a paucity of conserved interactions at the A−B ring junction or hydrophobic tail region in cholesterol. The conformational diversity of the tails when cholate and cholesterol bind to different sites results in their termini not being well conserved spatially, whereas they still experience different chemical environments. Detecting differences in the general protein environments of the alpha face of the steroid ring (upper face in Figure 8) and the tail termini in cholate (polar) versus cholesterol (hydrophobic) sites will be a focus for enhancements in CholMine as well as expanding the training data sets.

**Computational Efficiency of the CholMine Server.** For the 261 cholesterol, cholate, and other ligand sites analyzed here, the maximum protein volume for site map generation was <10 000 Å$^3$ (a box with edges of ∼21 Å), and each prediction completed in less than 5 min (the time to exhaustively check and score all orientations of the user-defined cleft versus the representative site, then filter for conserved interaction matches). For the majority of cases, the server elapsed time was <3 min per site.

## CONCLUDING DISCUSSION

CholMine, a predictor for cholesterol and cholate binding in protein three-dimensional structures, has been established as a free Web server at http://cholmine.bmb.msu.edu. This approach is based on the determination of conserved interactions for cholesterol and cholate binding to non-homologous membrane and soluble protein sites in PDB structures. SimSite3D alignment and scoring of site similarity serves as the first layer of prediction, considering the chemical interactions that can be made with the protein and their degree of surface match, independent of ligand information or protein structural conservation. This approach allows CholMine to focus on spatial conservation of chemical interactions rather than residue conservation. Requiring 70% match of the conserved spatial interactions of known cholesterol or cholate sites serves as the second layer of prediction, ruling out the vast majority of false positives in a data set of diverse soluble ligand


DOI: 10.1021/ci5006542
J. Chem. Inf. Model. 2015, 55, 747−759

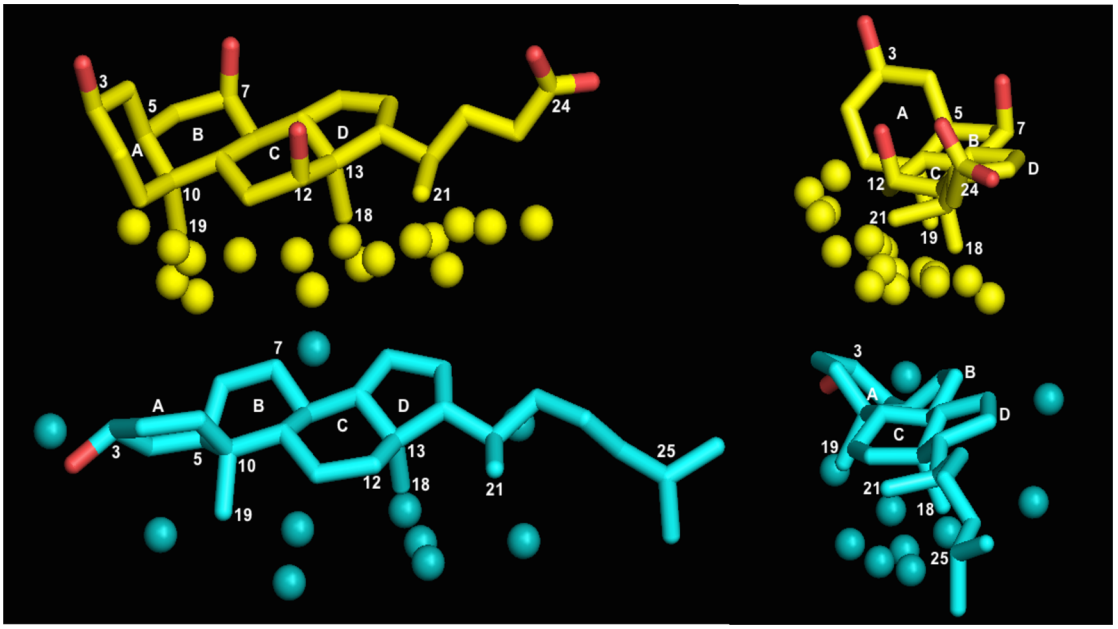



**Figure 8.** SimSite3D-identified conserved interactions for cholate (yellow) and cholesterol (blue) recognition abound along the groove formed between the row of C18, C19, and C21 methyl groups on the beta (lower) face of the steroid and the edge of the steroid ring system. The view on the right is rotated roughly 90 deg about a vertical axis through the center of each molecule. Cholate sites are distinguished from cholesterol primarily based on interactions with the relatively conserved C22−C23 tail orientation in cholate and numerous conserved interactions associated with the strongly bent (5-beta configuration) joint between the A and B rings of the cholate steroid ring system. Because the tail configurations are conformationally diverse in different binding sites, conserved interactions are absent in the C24−C25 region.

sites (resulting in a 5% false positive rate for cholesterol and 12% for cholate sites) and a slightly higher rate when applied to a data set of diverse membrane proteins (10% for cholesterol and 13% for cholate sites). CholMine can predict 80% of known cholesterol and 70% of known cholate binding sites in diverse protein families including soluble and membrane proteins from different species, when applied to sites unrelated to those used in training. CholMine can discriminate ∼75% of sites containing other steroids from cholesterol binding sites. Cholate site prediction is less steroid-selective; it also predicts two-thirds of the known cholesterol sites, likely due to the limited availability of nonhomologous cholate sites for training the predictor. This problem can be addressed by periodic updating of the training set. However, the false positive rate of cholate site predictions on diverse membrane protein sites is relatively low, 13%.

Hydrophobic interactions focused along the groove between the steroid methyl group substituents and the ring system itself are found to be the major conserved determinants for the recognition of both cholesterol and cholate, with their polar groups not contributing to conserved interactions. Classical motifs for cholesterol site prediction have focused on amino acid residue conservation and tend not to generalize well to other protein families, with particularly limited performance for predicting known binding sites in soluble proteins. Sequence motif-based prediction also results in many false positives (with 70% or more of 139 diverse noncholesterol, noncholate binding sites falsely predicted), which overwhelms the number of true positive predictions. The enhanced predictive specificity and selectivity of CholMine is based on inferring shared three-dimensional shape and chemical information from non-homologous sites. This approach is now being generalized to create a LigPattern server that discovers the shared interaction determinants of other important regulatory ligands and substrates, including polar molecules such as adenosine.

## ■ ASSOCIATED CONTENT

### Supporting Information

Tables containing test data sets used in the CholMine validation. This material is available free of charge via the Internet at http://pubs.acs.org.

## ■ AUTHOR INFORMATION

### Corresponding Author

*Phone: +1-517-353-8745. Fax: +1-517-353-9334. E-mail: KuhnL@msu.edu.

### Present Address

[‖]J.R.V.V.: Symantec Corporation, 2815 Cleveland Ave. North, Minneapolis, MN 55113.

### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

This research was partially supported by funding from the Great Lakes Fishery Commission to L.K., to guide the discovery of bile acid mimics as ligands for G protein coupled receptors. We thank OpenEye Scientific Software, LLC (Santa Fe, NM) for academic licensing of their QuacPac/molcharge software and the OEChem Toolkit used in this research. We are grateful to Shelagh Ferguson-Miller and Fei Li (Michigan State University) for their suggestions on CholMine software features to support experimental follow-up, and their feedback on this manuscript. The CholMine predictor is available for use at http://cholmine.bmb.msu.edu



DOI: 10.1021/ci5006542
J. Chem. Inf. Model. 2015, 55, 747−759

## REFERENCES


(1) Lund, S.; Orlowski, S.; Foresta, B. de; Champeil, P.; Maire, M. Le; Møbller, J. V. Detergent Structure and Associated Lipid as Determinants in the Stabilization of Solubilized $Ca^{2+}$-Atpase from Sarcoplasmic Reticulum. *J. Biol. Chem.* **1989**, *264*, 4907−4915.

(2) Seddon, A. M.; Curnow, P.; Booth, P. J. Membrane Proteins, Lipids and Detergents: Not Just a Soap Opera. *Biochim. Biophys. Acta* **2004**, *1666*, 105−117.

(3) Contreras, F.-X.; Ernst, A. M.; Wieland, F.; Brügger, B. Specificity of Intramembrane Protein-Lipid Interaction. *Cold Spring Harbor Perspect. Biol.* **2011**, *3*, 1−18.

(4) Ernst, A. M.; Contreras, F.-X.; Brügger, B.; Wieland, F. Determinants of Specificity at the Protein−Lipid Interface in Membranes. *FEBS Lett.* **2010**, *584*, 1713−1720.

(5) Hite, R. K.; Li, Z.; Walz, T. Principles of Membrane Protein Interactions with Annular Lipids Deduced from Aquaporin-0 2D Crystals. *EMBO J.* **2010**, *29*, 1652−1658.

(6) Shinzawa-Itoh, K.; Aoyama, H.; Muramoto, K.; Terada, H.; Kurauchi, T.; Tadehara, Y.; Yamasaki, A.; Sugimura, T.; Kurono, S.; Tsujimoto, K.; Mizushima, T.; Yamashita, E.; Tsukihara, T.; Yoshikawa, S. Structures and Physiological Roles of 13 Integral Lipids of Bovine Heart Cytochrome C Oxidase. *EMBO J.* **2007**, *26*, 1713−1725.

(7) Qin, L.; Hiser, C.; Mulichak, A.; Garavito, R. M.; Ferguson-Miller, S. Identification of Conserved Lipid Detergent-Binding Sites in a High-Resolution Structure of the Membrane Protein Cytochrome C Oxidase. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 16117−16122.

(8) Munro, S. Lipid Rafts: Elusive or Illusive? *Cell* **2003**, *115*, 377−388.

(9) Burger, K.; Gimpl, G.; Fahrenholz, F. Regulation of Receptor Function by Cholesterol. *Cell. Mol. Life Sci.* **2000**, *57*, 1577−1592.

(10) Schroeder, C. Cholesterol-Binding Viral Proteins in Virus Entry and Morphogenesis. In *Cholesterol Binding and Cholesterol Transport Proteins: Structure and Function in Health and Disease*; Harris, J. R., Ed.; Springer: Dordrecht, The Netherlands, 2010; Vol. *51*, pp 77−108.

(11) Huber, T. B.; Schermer, B.; Müeller, R. U.; Höhne, M.; Bartram, M.; Calixto, A.; Hagmann, H.; Reinhardt, C.; Koos, F.; Kunzelmann, K.; Shirokova, E.; Krautwurst, D.; Harteneck, C.; Simons, M.; Pavenstädt, H.; Kerjaschki, D.; Thiele, C.; Walz, G.; Chalfie, M.; Benzing, T. Podocin and MEC-2 Bind Cholesterol to Regulate the Activity of Associated Ion Channels. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 17079−17086.

(12) Hulce, J. J.; Cognetta, A. B.; Niphakis, M. J.; Tully, S. E.; Cravatt, B. F. Proteome-Wide Mapping of Cholesterol-Interacting Proteins in Mammalian Cells. *Nat. Methods* **2013**, *10*, 259−264.

(13) Hanson, M. A.; Cherezov, V.; Griffith, M. T.; Roth, C. B.; Jaakola, V. P.; Chien, E. Y.; Velasquez, J.; Kuhn, P.; Stevens, R. C. A Specific Cholesterol Binding Site is Established by the 2.8 Å Structure of the Human B2-Adrenergic Receptor. *Structure* **2008**, *16*, 897−905.

(14) Adamian, L.; Naveed, H.; Liang, J. Lipid-Binding Surface of Membrane Proteins: Evidence from Evolutionary and Structure Analysis. *Biochim. Biophys. Acta* **2011**, *1808*, 1092−1102.

(15) Li, H.; Papadopoulos, V. Peripheral-Type Benzodiazepine Receptor Function in Cholesterol Transport. Identification of a Putative Cholesterol Recognition/Interaction Amino Acid Sequence and Consensus Pattern. *Endocrinology* **1998**, *139*, 4991−4997.

(16) Takeda, K.; Tonthat, N. K.; Glover, T.; Xu, W.; Koonin, E. V.; Yanagida, M.; Schumacher, M. A. Implications for Proteasome Nuclear Localization Revealed by the Structure of the Nuclear Proteasome Tether Protein Cut8. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 16950−16955.

(17) Li, F.; Liu, J.; Valls, L.; Ferguson-Miller, S. Identification of a Key Cholesterol Binding Enhancement Motif in Translocator Protein 18 Kda (TSPO). *Biochemistry* **2015**, *54*, 1441−1443.

(18) Baier, C. J.; Fantini, J.; Barrantes, F. J. Disclosure of Cholesterol Recognition Motifs in Transmembrane Domains of the Human Nicoticin Acetylcholine Receptor. *Sci. Rep.* **2011**, *1*, 1−7.

(19) Fantini, J.; Yahi, N. Molecular Basis for the Glycosphingolipid-Binding Specificity of α-Synuclein: Key Role of Tyrosine 39 in Membrane Insertion. *J. Mol. Biol.* **2011**, *408*, 654−669.

(20) Fantini, J.; Barrantes, F. J. How Cholesterol Interacts With Membrane Proteins: an Exploration of Cholesterol-Binding Sites Including CRAC, CARC, and Tilted Domains. *Front Physiol.* **2013**, *4*, 1−9.

(21) Palmer, M. Cholesterol and the Activity of Bacterial Toxins. *FEMS Microbiol. Lett.* **2004**, *238*, 281−289.

(22) Barrett, P. J.; Song, Y.; Van Horn, W. D.; Hustedt, E. J.; Schafer, J. M.; Hadziselimovic, A.; Beel, A. J.; Sanders, C. R. The Amyloid Precursor Protein Has a Flexible Transmembrane Domain and Binds Cholesterol. *Science* **2012**, *336*, 1168−1171.

(23) Song, Y.; Kenworthy, A. K.; Sanders, C. R. Cholesterol as a Co-Solvent and a Ligand for Membrane Proteins. *Protein Sci.* **2014**, *23*, 1−22.

(24) Farrand, A. J.; LaChapelle, S.; Hotze, E. M.; Johnson, A. E.; Tweten, R. K. Only Two Amino Acids are Essential for Cytolytic Toxin Recognition of Cholesterol at the Membrane Surface. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 4341−4346.

(25) Chiang, J. Y. Bile Acid Regulation of Gene Expression: Roles of Nuclear Hormone Receptors. *Endocr. Rev.* **2001**, *23*, 443−463.

(26) Russell, D. W.; Setchell, K. D. R. Bile Acid Biosynthesis. *Biochemistry* **1992**, *31*, 4737−4749.

(27) Hofmann, A. F. The Enterohepatic Circulation of Bile Acids in Man. *Clin. Gastroenterol.* **1977**, *6*, 3−24.

(28) Maruyama, T.; Miyamoto, Y.; Nakamura, T.; Tamai, Y.; Okada, H.; Sugiyama, E.; Nakamura, T.; Itadani, H.; Tanaka, K. Identification of Membrane-Type Receptor for Bile Acids (M-BAR). *Biochem. Biophys. Res. Commun.* **2002**, *298*, 714−719.

(29) Kawamata, Y.; Fujii, R.; Hosoya, M.; Harada, M.; Yoshida, H.; Miwa, M.; Fukusumi, S.; Habata, Y.; Itoh, T.; Shintani, Y.; Hinuma, S.; Fujisawa, Y.; Fujino, M. A G Protein-Coupled Receptor Responsive to Bile Acids. *J. Biol. Chem.* **2003**, *278*, 9435−9440.

(30) Lischka, F.; Kuhn, L. A.; Libants, S.; Wu, H.; Yuan, Q.; Teeter, J.; Li, W. De-Orphanization of Two Vertebrate Pheromone Receptors. In preparation, 2015.

(31) Yau, W.-M.; Wimley, W. C.; Gawrisch, K.; White, S. H. The Preference of Tryptophan for Membrane Interfaces. *Biochemistry* **1998**, *37*, 14713−14718.

(32) Ballesteros, J. A.; Weinstein, H. Integrated Methods for the Construction of Three-Dimensional Models and Computational Probing of Structure-Function Relations in G Protein-Coupled Receptors. *Methods Neurosci.* **1995**, *25*, 366−428.

(33) Hunte, C. Specific Protein-Lipid Interactions in Membrane Proteins. *Biochem. Soc. Trans.* **2005**, *33*, 938−942.

(34) Liu, W.; Chun, E.; Thompson, A. A.; Chubukov, P.; Xu, F.; Katritch, V.; Han, G. W.; Roth, C. B.; Heitman, L. H.; IJzerman, A. P.; Cherezov, V.; Stevens, R. C. Structural Basis for Allosteric Regulation of Gpcrs by Sodium Ions. *Science* **2012**, *337*, 232−236.

(35) Lin, H. H.; Han, L. Y.; Zhang, H. L.; Zheng, C. J.; Xie, B.; Chen, Y. Z. Prediction of the Functional Class of Lipid Binding Proteins from Sequence-Derived Properties Irrespective of Sequence Similarity. *J. Lipid Res.* **2006**, *47*, 824−831.

(36) Xiong, W.; Guo, Y.; Li, M. Prediction of Lipid-Binding Sites Based on Support Vector Machine and Position Specific Scoring Matrix. *Protein J.* **2010**, *29*, 427−431.

(37) Scott, D. L.; Diez, G.; Goldmann, W. H. Prediction-Lipid Interactions: Correlation of a Predictive Algorithm for Lipid-Binding Sites with Three-Dimensional Structural Data. *Theor. Biol. Med. Model.* **2006**, *3*, 1−14.

(38) Van Voorst, J. R.; Finzel, B. C.; Tonero, M. E.; Rai, B.; Narasimhan, L.; Howe, W. J.; Kuhn. L. A. Screening to Identify Similar Ligand-Binding Pockets in Diverse Proteins. In preparation, 2015.

(39) Weill, N.; Rognan, D. Development and Validation of a Novel Protein-Ligand Fingerprint to Mine Chemogenomic Space: Application to G Protein-Coupled Receptors and Their Ligands. *J. Chem. Inf. Model.* **2009**, *49*, 1049−1062.




DOI: 10.1021/ci5006542
*J. Chem. Inf. Model.* 2015, 55, 747−759

(40) Madala, P. K.; Fairlie, D. P.; Boden, M. Matching Cavities in G Protein-Coupled Receptors to Infer Liand-Binding Sites. *J. Chem. Inf. Model.* **2012**, *52*, 1401−1410.
(41) Van Voorst, J. R. *Surface Matching and Chemical Scoring to Detect Unrelated Proteins Binding Similar Small Molecules*. Ph.D. Thesis, Michigan State University, East Lansing, MI, 2011.
(42) Zavodszky, M. I.; Sanschagrin, P. C.; Korde, R. S.; Kuhn, L. A. Distilling the Essential Features of a Protein Surface for Improving Protein-Ligand Docking, Scoring, and Virtual Screening. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 883−902.
(43) Wang, G.; Dunbrack, R. L., Jr. PISCES: A Protein Sequence Culling Server. *Bioinformatics* **2003**, *19*, 1589−1591.
(44) Tan, P. N.; Steinbach, M.; Kumar, V. *Introduction to Data Mining*; Addison-Wesley: Boston, MA, 2006; pp 370−386.
(45) Craig, L.; Sanschagrin, P. C.; Rozek, A.; Lackie, S.; Kuhn, L. A.; Scott, J. K. The Role of Structure in Antibody Cross-Reactivity between Peptides and Folded Proteins. *J. Mol. Biol.* **1998**, *281*, 183−201.
(46) Wallace, A. C.; Laskowski, R. A.; Thornton, J. M. LIGPLOT: a Program to Generate Schematic Diagrams of Protein-Ligand Interactions. *Protein Eng.* **1996**, *8*, 127−134.
(47) Laskowski, R. A.; Swindells, M. B. LigPlot+: Multiple Ligand−Protein Interaction Diagrams for Drug Discovery. *J. Chem. Inf. Model.* **2011**, *51*, 2778−2786.
(48) Barenholz, Y. Cholesterol and Other Membrane Active Sterols: From Membrane Evolution to Rafts. *Prog. Lipid Res.* **2002**, *41*, 1−5.
(49) Majewska, M. D. Steroids and Ion Channels in Evolution: From Bacteria to Synapses and Mind. *Acta Neurobiol. Exp.* **2007**, *67*, 219−233.


DOI: 10.1021/ci5006542
*J. Chem. Inf. Model.* 2015, 55, 747−759