# Lipid classification, structures and tools*

**Eoin Fahy**[*], **Dawn Cotter**, **Manish Sud**, and **Shankar Subramaniam**
University of California, San Diego, 9500 Gilman Dr., La Jolla, CA 92093-0411, USA

## Abstract

The study of lipids has developed into a research field of increasing importance as their multiple biological roles in cell biology, physiology and pathology are becoming better understood. The Lipid Metabolites and Pathways Strategy (LIPID MAPS) consortium is actively involved in an integrated approach for the detection, quantitation and pathway reconstruction of lipids and related genes and proteins at a systems-biology level. A key component of this approach is a bioinformatics infrastructure involving a clearly defined classification of lipids, a state-of-the-art database system for molecular species and experimental data and a suite of user-friendly tools to assist lipidomics researchers. Herein, we discuss a number of recent developments by the LIPID MAPS bioinformatics core in pursuit of these objectives. This article is part of a Special Issue entitled Lipodomics and Imaging Mass Spectrometry.

### Keywords

Lipids; Lipidomics; Classification; Mass spectrometry; Bioinformatics

## 1. Introduction

Lipids are a diverse and ubiquitous group of compounds which have many key biological functions, such as acting as structural components of cell membranes, serving as energy storage sources and participating in signaling pathways. A comprehensive analysis of lipid molecules, "lipidomics," in the context of genomics and proteomics is crucial to understanding cellular physiology and pathology; consequently, lipid biology [1,2] has become a major research target of the post-genomic revolution and systems biology. The word "lipidome" is used to describe the complete lipid profile within a cell, tissue or organism and is a subset of the "metabolome" which also includes the three other major classes of biological molecules: amino-acids, sugars and nucleic acids. Lipidomics is a relatively recent research field that has been driven by rapid advances in a number of analytical technologies, in particular mass spectrometry (MS), and computational methods, coupled with the recognition of the role of lipids in many metabolic diseases such as obesity, atherosclerosis, stroke, hypertension and diabetes. This rapidly expanding field complements the huge progress made in genomics and proteomics, all of which constitute the family of systems biology. The diversity in lipid function is reflected by an enormous

---

[*]Corresponding author. efahy@ucsd.edu (E. Fahy).

variation in the structures of lipid molecules. Unlike the case of genes and proteins which are primarily composed of linear combinations of 4 nucleic acids and 20 amino acids, respectively, lipid structures are generally much more complex due to the number of different biochemical transformations which occur during their biosynthesis. This level of diversity makes it important to develop a comprehensive classification, nomenclature, and chemical representation system to accommodate the myriad lipids that exist in nature. A modern, robust classification system in turn paves the way for creation of a comprehensive bioinformatics infrastructure that includes databases of lipids and lipid-associated genes, tools for representing lipid structures, interfaces for analyzing lipidomic experimental data and methodologies for studying lipids at a systems-biology level. This article summarizes the efforts of the LIPID MAPS consortium to address these informatics challenges and play a part in the establishment of lipidomics as a field of emerging importance in biology.

## 2. Lipid classification and nomenclature

The term "lipid" has been loosely defined as any of a group of organic compounds that are insoluble in water but soluble in organic solvents [3]. These chemical features are present in a broad range of molecules such as fatty acids, phospholipids, sterols, sphingolipids, terpenes and others. In view of the fact that lipids comprise an extremely heterogeneous collection of molecules from a structural and functional standpoint, it is not surprising that there are significant differences with regard to the scope and organization of current classification schemes. A number of sources such as The Lipid Library [4] and Cyberlipids [5] segregate lipids into "simple" and "complex" groups, with simple lipids being those yielding at most two types of distinct entities upon hydrolysis (e.g., acylglycerols: fatty acids and glycerol) and complex lipids (e.g., glycerophospholipids: fatty acids, glycerol, and headgroup) yielding three or more products upon hydrolysis. In contrast, the LipidBank [6,7] database in Japan defines a third major group called "derived" lipids (alcohols and fatty acids derived by hydrolyzing the simple lipids) and includes 26 top-level categories in their classification scheme, covering a wide variety of animal and plant sources. In 2005, the International Lipid Classification and Nomenclature Committee on the initiative of the LIPID MAPS Consortium developed and established a comprehensive classification system for lipids based on well-defined chemical and biochemical principles and using a framework designed to be extensible, flexible, scalable and compatible with modern informatics technology [8,9].

### 2.1. Lipid classification

The LIPID MAPS classification system is based on the concept of 2 fundamental "building blocks": ketoacyl groups and isoprene groups (Fig. 1). Consequently, lipids are defined as hydrophobic or amphipathic small molecules that may originate entirely or in part by carbanion based condensations of ketoacyl thioesters and/or by carbocation based condensations of isoprene units (Fig. 2). Based on this classification system, lipids have been divided into eight categories: fatty acyls, glycerolipids, glycerophospholipids, sphingolipids, saccharolipids and polyketides (derived from condensation of ketoacyl subunits); and sterol lipids and prenol lipids (derived from condensation of isoprene subunits) (Fig. 3). Each category is further divided into classes, subclasses and, in the case

of some subclasses of prenol lipids, 4th-level classes. Each lipid is assigned a unique 12- or 14-character identifier (LIPID MAPS ID or "LM ID") based on this classification scheme. The format of the LM ID contains the classification information, provides a systematic means of assigning a unique identification to each lipid molecule and allows for the addition of large numbers of new categories, classes, and subclasses in the future. The last four characters of the LM ID comprise a unique identifier within a particular subclass and are randomly assigned (Table 1). The fatty acyls (FA) are a diverse group of molecules synthesized by chain elongation of an acetyl-CoA primer with malonyl-CoA (or methylmalonyl-CoA) groups that may contain a cyclic functionality and/or are substituted with heteroatoms. It is notable that this category contains not just fatty acids but several other functional variants such as alcohols, aldehydes, amines and esters. Structures with a glycerol group are represented by two distinct categories: the glycerolipids (GL), which include acylglycerols but also encompass alkyl and 1Z-alkenyl variants, and the glycerophospholipids (GP), which are defined by the presence of a phosphate (or phosphonate) group esterified to one of the glycerol hydroxyl groups. The sterol lipids (ST) and prenol lipids (PR) share a common biosynthetic pathway via the polymerization of dimethylallyl pyrophosphate/isopentenyl pyrophosphate but have key differences in terms of their eventual structure and function. In particular the presence of a unique fused ring structure for the sterols distinguishes them from classes of cyclic triterpenes such as the protostanes and fusidanes, which at first glance appear to be sterols but in fact have different fused ring stereochemistry and methylation patterns as a consequence of alternative folding of the linear precursor during biosynthesis. Another well-defined category is the sphingolipids (SP), which contain a long-chain nitrogenous base as their core structure. The saccharolipid (SL) category was created to account for lipids in which fatty acyl groups are linked directly to a sugar backbone. This SL group is distinct from the term "glycolipid" that was defined by the International Union of Pure and Applied Chemists (IUPAC) as a lipid in which the fatty acyl portion of the molecule is present in a glycosidic linkage [10]. It should be noted that glycosylated derivatives of the other 7 lipid categories are classified as members (classes/subclasses) of those appropriate categories and have a major contribution to the structural diversity of lipids in general. The final category is the polyketides (PK), which are a diverse group of metabolites from animal, plant and microbial sources.

This classification system has been the foundation for the LIPID MAPS Structure Database (LMSD) [11], which is discussed in more detail below. The most convenient way of viewing the classification hierarchy is through the LIPID MAPS Nature Lipidomic Gateway [12] where one may conveniently browse examples of categories, classes and subclasses stored in the LMSD.

## 2.2. Lipid nomenclature

The nomenclature of lipids falls into two main categories: systematic names and common or trivial names. The latter includes abbreviations which are a convenient way to define acyl/alkyl chains in glycerolipids, sphingolipids and glycerophospholipids. The generally accepted guide-lines for lipid systematic names were initially defined by the International Union of Pure and Applied Chemists and the International Union of Biochemistry and Molecular Biology (IUPAC-IUBMB) Commission on Biochemical Nomenclature in 1976,

and their recommendations are summarized on the IUPAC website [13]. The nomenclature adopted by the LIPID MAPS consortium follows existing IUPAC-IUBMB rules closely. The main differences involve (a) clarification of the use of core structures to simplify systematic naming of some of the more complex lipids, and (b) provision of systematic names for recently discovered lipid classes. Key features of our lipid nomenclature scheme are as follows:

**a.** The use of the stereospecific numbering (*sn*) method to describe glycerolipids and glycerophospholipids. The glycerol group is typically acylated or alkylated at the *sn*1 and/or *sn*2 position with the exception of some lipids which contain more than one glycerol group and archaebacterial lipids in which *sn*2 and/or *sn*3 modification occurs.

**b.** Definition of sphinganine and sphing-4-enine as a core structure for the sphingolipid category where the D-erythro or 2S,3R configuration and 4E geometry (in the case of sphing-4-enine) are implied. In molecules containing stereochemistry other than the 2S,3R configuration, the full systematic names are to be used instead (e.g., 2R-amino-1,3R-octadecanediol).

**c.** The use of core names such as cholestane, androstane, and estrane, for sterols.

**d.** Adherence to the names for fatty acids and acyl-chains (formyl, acetyl, propionyl, butyryl, etc.) defined in Appendix A and B of the IUPAC-IUBMB recommendations.

**e.** The adoption of a condensed text nomenclature for the glycan portions of lipids, where sugar residues are represented by standard IUPAC abbreviations, and where the anomeric carbon locants and stereochemistry are included but where the parentheses are omitted. This system has also been proposed by the Consortium for Functional Glycomics (http://www.functionalglycomics.org/static/index.shtml).

**f.** The use of E/Z designations (as opposed to trans/cis) to define double-bond geometry.

**g.** The use of R/S designations (as opposed to a/β or D/L) to define stereochemistries. The exceptions are those describing substituents on glycerol (*sn*) and sterol core structures, and anomeric carbons on sugar residues. In these latter special cases, the α/β format is firmly established.

**h.** The common term 'lyso', denoting the position lacking a radyl group in glycerolipids and glycerophospholipids, will not be used in systematic names, but will be included as a synonym.

**i.** The proposal for a single nomenclature scheme to cover the prostaglandins, isoprostanes, neuroprostanes, and related compounds where the carbons participating in the cyclopentane ring closure are defined and where a consistent chain numbering scheme is used.

**j.** The "d" and "t" designations used in shorthand notation of sphingolipids refer to 1,3 dihydroxy and 1,3,4-trihydroxy long-chain bases, respectively.

**k.** The glycerophospholipid abbreviations (Table 2) are used to refer to species with one or two radyl side-chains where the structures of the side chains are indicated within parentheses in the 'Headgroup (*sn1*/*sn2*)' format (e.g. PC(16:0/18:1(9Z)). By default, the C2 carbon of glycerol has R stereochemistry and attachment of the headgroup at the *sn3* position. In a similar fashion, the glycerolipid abbreviations (MG,DG,TG for mono-, di- and triradyglycerols respectively) are used to refer to species with one to three radyl side-chains where the structures of the side chains are indicated within parentheses in the 'Headgroup(*sn1*/*sn2*/*sn3*)' format (e.g. TG (16:0/18:1(9Z)/16:0)).

**l.** The alkyl ether linkage is represented by the "O-" prefix, e.g. DG(O-16:0/18:1(9Z)/0:0) and the (1Z)-alkenyl ether (neutral Plasmalogen) species by the "P-" prefix, e.g. DG(P-14:0/18:1(9Z)/0:0). The same rules apply to the head-group classes within the Glycerophospholipids category. In cases where glycerolipid total composition is known, but side-chain regiochemistry and stereochemistry is unknown, abbreviations such as TG(52:1) and DG(34:2) may be used, where the numbers within parentheses refer to the total number of carbons and double bonds of all the chains.

## 3. Lipid structures

Lipids display remarkable structural diversity, driven by factors such as variable chain length, a multitude of oxidative, reductive, substitutional and ring-forming biochemical transformations as well as modification with sugar residues and other functional groups of different biosynthetic origin. There are no reliable estimates of the number of discrete lipid structures in nature, due to the technical challenges of elucidating chemical structures. Estimates in the order of 200,000, based on acyl/alkyl chain and glycan permutations for glycerolipids, glycerophospholipids and sphingolipids are almost certainly on the conservative end [14]. This number is actually exceeded by the list of known natural products, most of which are of either prenol or polyketide origin [15]. Given the importance of these molecules in cellular function and pathology, it is essential to have well-organized databases of lipids with relevant structural information and related features.

### 3.1. Lipid structure databases

Modern bioinformatics has become increasingly sophisticated to permit complex database schemas and approaches which enable efficient data acquisition, storage and dissemination. Repositories such as GenBank [16] SwissProt [17] and ENSEMBL [18] support nucleic acid and protein databases; however until recently there was little effort dedicated to cataloging lipids. A number of online resources have currently become available for data basing of lipid structures as well as relevant spectral, biochemical information and references. The LIPIDAT database [19] focuses on the biophysical properties of glycerophospholipids, glycerolipids and sphingolipids and contains over 12,000 unique molecular structures. However it has not been updated in several years and lacks many of the modern online search capabilities. The Lipid Bank database [6] contains over 7000 structures across a broad range of lipid classes and is a rich resource for associated spectral data, biological properties and literature references. Like LIPIDAT, it has not been updated recently during a

period which has seen a considerable increase in publication of new lipid structures. Further it uses an unusual classification system that has not been widely adopted by the lipids research community. The emergence of the LIPID MAPS classification system in 2005 laid the foundation for a comprehensive object-relational database of lipids known as the LIPID MAPS Structure Database (LMSD) [11]. The LMSD is available on the LIPID MAPS Nature Lipidomics Gateway website [12] with browsing and searching capabilities. It currently contains over 30,000 structures which are obtained from a variety of sources: LIPIDMAPS Consortium's core laboratories and partners; lipids identified by LIPID MAPS experiments; computationally generated structures for appropriate lipid classes; biologically relevant lipids manually curated from Lipid Bank, LIPIDAT, Cyberlipids and other public databases; peer-reviewed journals and book chapters describing lipid structures. After lipids have been selected for inclusion into LMSD, they are classified following the LIPID MAPS classification scheme and assigned a unique LIPID MAPS identifier (LM ID). Structures are stored as binary large objects (BLOBs) within an Oracle database [20] and may be viewed online in multiple formats as GIF images (default), Java applets and ChemDraw objects. In addition to structures the LMSD also contains all relevant information for that molecule such as, common and systematic names, synonyms, molecular formula, exact mass, classification hierarchy, InChIKey [21] and cross-references (if any) to other databases.

## 3.2. Lipid structure representation

In addition to having rules for lipid classification and nomenclature, it is important to establish clear guidelines for drawing lipid structures. Currently, there is a large diversity in the way molecules are represented [4–6,18] which tends to create confusion especially in the case of the more complex lipid structures. For example, usage of the Simplified Molecular Line Entry Specification (SMILES) [22,23] format to represent lipid structures, while being very compact and accurate in terms of bond connectivity, valence and chirality, causes problems when the structure is rendered. In order to address these issues, the LIPID MAPS consortium proposed a consistent framework for representing lipid structures [8,9,11]. In general, the acid/acyl group or its equivalent is drawn on the right side and hydrophobic chain is on the left. Thus for glycerolipids and glycerophospholipids, the radyl hydrocarbon chains are drawn to the left; the glycerol group is drawn horizontally with stereochemistry defined at *sn* carbons and the headgroup for glycerophospholipids is depicted on the right. For sphingolipids, the C1 hydroxyl group of the long-chain base is placed on the right and the alkyl portion on the left; the headgroup of sphingolipids ends up on the right. The linear prenols and isoprenoids are drawn like fatty acids with the terminal functional groups on the right. A number of structurally complex lipids – polycyclic isoprenoids, and polyketides – cannot be drawn using these simple rules; these structures are drawn using commonly accepted representations. Structures of all lipids in LMSD adhere to the structure drawing rules proposed by the LIPID MAPS consortium. Fig. 4 shows representative structures for several lipid categories. This approach has the advantage of making it easier to recognize lipids within a category or class and also to highlight structural similarities between classes. For example the glycerophosphocholines (PC) and sphingomyelins (SM) have distinctly different biosynthetic origins but their structural similarities as amphipathic membrane lipids are evident when displayed as shown in Fig. 5.

# 4. Bioinformatics tools for lipidomic analysis

Recent advances in the field of lipidomics have largely been driven by the development of new mass spectrometric tools and protocols as well as dramatic improvements in informatics infrastructure and databasing capability [24,25]. As we enter the era of high-throughput lipidomics analysis there is an emerging need for a variety of software tools to handle key tasks such as lipid identification (mainly via MS analysis), quantitation, data processing and databasing, statistical analysis, pathway analysis/systems biology modeling as well as user-friendly information resources. The following section discusses a number of lipidomic tools and resources recently developed by the LIPID MAPS consortium with an objective toward addressing these needs.

## 4.1. Online lipidomic display tools

The growth of informational resources for modern biology has been greatly enhanced over the last two decades taking advantage of the dramatic developments in database technology and the evolution of the world-wide-web. As far as lipidomics is concerned, these technologies allow researchers to obtain and display information on structures, experimental data and protocols much more conveniently using a web browser. A case in point is the LMSD structure database on the LIPID MAPS website where users may search for structures in several different ways such as classification-based, text/structural feature-based, and structure-based methods. The classification-based method provides the capability to browse lipid categories and classes based on the LIPID MAPS classification scheme. After the user selects one of the main categories of lipids, options to specify a particular class or subclass within that category are provided. In this fashion one may "drill down" to the set of lipids of interest and view their structures and related information through their detail pages.

Another very useful method is a text-based search where a user enters an entire or a partial lipid name or a synonym in a search form. Other search terms include molecular formula, lipid mass (within a specified range), classification term and LM ID (if known). More than one term may be used simultaneously (e.g. lipid name and mass range) to further limit the search, and results are displayed in tabular format with links to the detail-view for each molecule. An issue of major importance in dealing with lipid structures is the huge diversity of chemical functional groups. This presents problems in explicitly classifying certain lipids containing multiple functional groups since assignment of a structure to a particular subclass may be somewhat subjective. For example, a fatty acid containing both epoxy and hydroxyl groups could be assigned to either an epoxy- or a hydroxy fatty acids subclass. To address this problem, the LIPID MAPS bioinformatics group has developed software methods which calculate the number of functional groups, number of rings and other structural information from a Molfile representation of a molecular structure. This information in turn is used to create a table of structural features which may then be incorporated into the database infrastructure. This enables the use of a feature-based text search where a user may choose to search for lipids containing certain functional groups, number of carbons, rings, etc., irrespective of their classification designation. A web-based implementation of this type of feature-based search has been implemented on the LIPID MAPS website.

A structure-based search option provides the capability to search the LMSD by performing a substructure or exact match using the structure drawn by the user. Three supported structure drawing tools are MarvinSketch [26], JME [27] and ChemDrawPro [28]. The first two of these structure drawing tools are Java applets and require only applet support in the browser. In addition, the user may also specify a partial LM ID and/or name or a synonym to further refine the search.

The record details page, in addition to displaying the structure for the selected lipid, also contains all relevant information for that molecule such, common and systematic names, synonyms, molecular formula, exact mass, classification hierarchy and cross-references (if any) to other databases. Fig. 6 shows some screen shots of the LMSD user interface for lipid classification-based browsing, text-based and structure-based searching. The details page also includes the IUPAC International Chemical Identifier [21] (InChI) string and its compressed version (InChIKey) for each lipid, generated with software available from InChI website. The InChIKey is a fixed length (25-character) condensed digital representation of the InChI which encodes structural information in a layered format and is specially designed to allow for easy web searches of chemical compounds. Of particular relevance to lipids is that the first 14 characters of the InChIKey (the main layer) define the molecular formula and bond connectivity of the molecule, whereas the remainder of the InChIKey defines the stereochemistry, double-bond geometry and isotopic substitution (if any). Therefore one may use an entire InChIKey to perform an "exact" search for that particular structure in the LMSD database or any other online resource that exposes InChIKeys for searching purposes. Secondly, using only the 14-character string of the main layer, one can search for all stereoisomers and/or cis/trans geometric isomers of a molecule of interest. This feature is enabled in the LMSD detail view pages as a hyperlink in the InChIKey section and is exemplified in Fig. 7 where in the case of cholic acid all 16 stereoisomers corresponding to the 3 hydroxyl groups and C5 stereocenter share the same main layer.

Finally, the advent of simple and powerful online search engines such as Google has popularized the notion of a "power-search" option for many online data repositories where a user can enter a query from the home page and search multiple databases and/or resources concurrently. Accordingly, a "Quick Search" feature has been developed for the LIPID MAPS website where a user may search the lipid database, the lipid-related gene/protein database (LMPD) [29], the lipid standards database as well as all text pages on the website by entering a complete or partial lipid class, common name, systematic name or synonym, InChIKey, LIPID MAPS ID, gene or protein term. Results are then organized and displayed in a context-specific format.

## 4.2. Structure drawing tools

The structures of large and complex lipids are difficult to represent in drawings, which leads to the use of many custom formats that often generate more confusion than clarity among members of the lipid research community. Additionally the structure drawing step is typically the most time-consuming process in creating molecular databases of lipids. However, many classes of lipids lend themselves to automated structure drawing paradigms, due to their consistent 2-dimensional layout. The LIPID MAPS consortium has developed

and deployed a suite of structure drawing tools [30] that greatly increase the efficiency of data entry into lipid structure databases and permit "on-demand" structure generation. Following the structural representation guidelines [8,9] discussed above enables a more consistent, error-free approach to drawing lipid structures and has been used extensively in populating the LMSD, which currently contains over 22,500 molecules. LMSD structures are either drawn manually using ChemDraw or generated automatically by structure drawing tools developed by LIPID MAPS consortium for various subclasses in fatty acyls, glycerolipids, glycerophospholipids, sphingolipids, and sterols. The structure drawing tools themselves are scripts written in the Perl [31] programming language which can generate a large number of structures relatively quickly via command-line or web-based interface. We have adopted an approach where "core" structures such as diacetyl glycerol (glycerolipids) and formic acid (fatty acyls) are represented as text-based MDLMOL files, and these MOL file templates are then manipulated to generate a variety of structures containing that core and other appropriate modifications such as addition of acyl chains (Fig. 8).

The LIPID MAPS website currently contains a suite of structure drawing tools for the following lipid categories: fatty acyls, glycerolipids, glycerophospholipids, cardiolipins, sphingolipids, sterols, and sphingolipid glycans. All major lipid categories contain glycoslyated forms whose glycan substituents can be challenging to draw in full chair conformation. The glycan structure drawing tools support the generation of a wide variety of polymers by specifying the constituent sugars using the Consortium for Functional Glycomics nomenclature [32]. In general, the online layout (Fig. 9) consists of a "core" structure and pull-down menus arranged in locations appropriate for that structure. For example, in the case of the glycerophospholipid drawing tool, a central glycerol core is surrounded by pull-down menus allowing the end-user to choose from a list of headgroups and *sn*1 and *sn*2 acyl side-chains. The list of acyl chains represents the more common species found in mammalian cells, and may easily be modified to include additional chains. The structure is rendered in the web browser as a Java-based applet or Chem Draw object. The fatty acyl structure drawing tool has a different user-input format where the user enters a valid fatty acyl LIPID MAPS abbreviation representing acyl chain length, presence of double or triple bonds and substituents on the acyl chain. Examples are "18:1(9Z)" (oleic acid) and "20:4(5Z,8Z,11E,14Z)(11OH[S])" (11S-hydroxy-5Z,8Z,11E,14Z-eicosatetraenoic acid). The fatty acyl drawing tool is capable of drawing chiral centers and ring structures.

Command-line versions of these structure drawing tools are also extremely useful in the area of bioinformatics because structures and related information such as formulae, masses and abbreviations may be generated rapidly for large permutations of side-chain substituents. For example, one may define a set of 30 acyl/alkyl chains and create a "virtual database" glycerolipids comprised of all $30^3$ chain permutations (27,000 discrete molecules if *sn* stereochemistry is considered) which may be used for purposes such as a back-end database for MS searching, as discussed in the next section. The command-line tools are available from the LIPIDMAPS website along with detailed documentation on the methods and functions used by these programs.

### 4.3. MS search tools

The detection and quantitation of lipids has been greatly improved in recent years by the development of advanced mass spectrometry techniques, in particular matrix-assisted desorption/ionization (MALDI), atmospheric pressure chemical ionization (APCI) and electrospray ionization (ESI) methods in conjunction with automated liquid chromatography (LC) [33–38]. However it is extremely challenging to identify large numbers of lipids from MS experiments of biological samples where analysis is complicated by the presence of many analytes with similar mass-to-charge (m/z) ratios. Accordingly, considerable effort has been made over the last 10 years to develop software to deconvolute and process MS data from lipidomics experiments [24,39,40]. In general there are 3 predictive approaches that may be used: (a) identification with neutral loss or product ion scanning, (b) identification with MS/MS spectral database searching and (c) identification by accurate mass or a combination of mass and retention time in LC–MS experiments [41,42]. All of these methods have limitations with regard to precise structure elucidation to a greater or lesser extent due to inability to distinguish molecular features such as chiral centers, position of functional groups and double-bond location/geometry, although MS/MS and $MS^n$ analysis may provide key information about regiochemistry. In target lipidomics approaches, the use of multiple reaction monitoring (MRM) has been employed to identify precursor ion/product ion pairs with high sensitivity, for example, to detect and quantitate large numbers of eicosanoid species [43]. Fragment ion spectra of lipids from MS/MS experiments contain a wealth of information as compared to precursor ion scanning but suffer from the drawback that product ion detection and intensity is highly dependent on the type of instrument and acquisition parameters. Nevertheless, empirical databases of MS/MS data exist which can be used as part of fragment ion matching algorithms. For example, the LipidQA program [44] utilizes fragment ion database composed of calculated and acquired reference tandem spectra of glycerophospholipids of all major head-group classes. Users may then compare their raw MS/MS data to the set of theoretical fragment ion spectra in the database to assist in identification the lipid molecular species. With the advent of high-resolution mass spectrometers, such as Orbitrap and Fourier transform–ion cyclotron resonance (FT-ICR) instruments [38], another powerful approach for identification of individual lipid species is the use of accurate mass-matching where precursor ion masses are compared to a list of known masses from either (a) a database of lipids known to exist in biological samples, or (b) a "virtual" database which is artificially created by considering all possible chain permutations for a particular class of lipids (where the list of chains are those that are likely to be encountered in the sample of interest). In many cases, these instruments have sufficient mass resolution and mass accuracy to determine elemental composition but still suffer from the drawbacks mentioned above with regard to distinguishing between regioisomers and stereoisomers in isobaric mixtures. Certain classes of lipids such as glycerolipids and glycerophospholipids composed of an invariant core (glycerol and headgroups) and one or more acyl/alkyl substituents are good candidates for MS computational analysis. At least in mammalian systems, one can readily create databases composed of all permutations of the commonly occurring acyl/alkyl chains and calculate their predicted precursor ions for various ion modes and adducts. Additionally these classes generally tend to fragment in a predictable fashion in collision-induced experiments leading to loss of acyl side-chains, neutral loss of fatty acids, and loss of water and other diagnostic ions [45], depending on the

nature of the headgroup. The LIPID MAPS bioinformatics group has developed a suite of online MS search tools [12] that allows a user to enter an m/z value of interest and view a list of matching structure candidates, along with a list of calculated "high-probability" product ions where appropriate. The MS prediction tools are currently available for a number of different categories of lipids: glycerolipids, glycerophospholipids, cardiolipins, cholesteryl esters, fatty acids and sphingolipids. The MS prediction tools for glycerolipids and glycerophospholipids have been extended by computing product ion masses for commonly observed fragments corresponding to acyl chain ions, neutral loss of acyl chains, loss of water, headgroup-specific fragmentations and combinations of the above. The MS prediction tools generally accept an m/z value from the user for the precursor ion and offer a menu to allow selection of the ion mode ([M+H]$^+$, [M+NH4]$^+$, [M−H]$^−$, etc.). In addition, a mass tolerance window and a headgroup (in the case of glycerophospholipids) may be specified to limit the number of matches. The mass tolerance setting used is dictated by the instrument used to acquire the spectral data—low mass tolerance values appropriate for high resolution data will lead to a smaller number of possible matches. The list of matches may also be filtered by specifying a particular set of radyl chains (for example, only chains with even numbers of carbon atoms). On completion of a search, the output format (Fig. 10) contains a list of structures that (a) satisfy the input criteria and (b) whose side chains belong to the list of radyl chains used to populate the database. The predicted masses of the fragment ions are computed at run-time by the online application. All entries in the result set are hyperlinked to the structure-drawing application, enabling "on-demand" visualization of the molecular structures. Isotopic distribution profiles for each structure may also be viewed online. The online tools allow batch-mode searches of lists of precursor ions and intensity values which may be copied and pasted into the user interface. Users may perform searches where the matched ions are displayed in "bulk" format (e.g. PE(34:1), TG(54:2)) or as discrete molecular species (e.g. PE(16:0/18:1(9Z)), TG(18:0/18:1(9Z)/18:1(9Z))). The bulk format option has the advantage of simplifying the results table in cases where many isobaric species are present. Additionally, in the case of experimental samples where the relative amounts of the acyl groups of glycerolipids and glycerophospholipids are already known (e.g. from fatty acid methyl ester (FAME) analysis by GC), these data may be entered and a scoring algorithm then ranks the matched species based on the relative abundance of those acyl chains in each lipid. There are ongoing efforts to expand the scope of these MS analysis tools to cover additional lipid classes from sources such as plants and bacteria.

## 5. Conclusion

The field of lipidomics has made rapid progress on many fronts over the past two decades although it has still to achieve the same level of advancement and knowledge as genomics and proteomics. The diversity of lipid chemical structures presents a challenge both from the experimental and informatics standpoints. The need for a robust, scalable bioinformatics infrastructure is high at a number of different levels: (a) establishment of a globally accepted classification system, creation of databases of lipid structures, lipid-related genes and proteins, (c) efficient analysis of experimental data, (d) efficient management of metadata and protocols, (e) integration of experimental data and existing knowledge into metabolic

and signaling pathways, (f) development of informatics software for efficient searching, display and analysis of lipidomic data. The study of mammalian lipdomes has been complemented in recent years by comprehensive lipidomic analyses of yeast, mycobacteria, archaebacteria and plants, each with its own set of challenges and insights which will need to be addressed by collaborative efforts between biology, chemistry and bioinformatics.

## Acknowledgments

## References

1. Oresic M, Hänninen VA, Vidal-Puig A. Lipidomics: a new window to biomedical frontiers. Trends Biotechnol. 2008; 26:647–652. [PubMed: 18951641]

2. Watson AD. Lipidomics: a global approach to lipid analysis in biological systems. J Lipid Res. 2006; 47:2101–2111. [PubMed: 16902246]

3. Smith, AD. Oxford Dictionary of Biochemistry and Molecular Biology. Oxford University Press; Oxford: 2000. Rev

4. Lipid Library website. http://lipidlibrary.aocs.org

5. Cyberlipid Center website. http://www.cyberlipid.org

6. LIPID BANK website. www.lipidbank.jp

7. Watanabe K, Yasugi E, Oshima M. How to search the glycolipid data in "LIPIDBANK for Web", the newly developed lipid database in Japan. Trends Glycosci Glycotechnol. 2000; 12:175–184.

8. Fahy E, Subramaniam S, Brown HA, Glass CK, Merrill AH Jr, Murphy RC, Raetz CRH, Russell DW, Seyama Y, Shaw W, Shimizu T, Spener F, van Meer G, Van Nieuwenhze MS, White SH, Witztum JL, Dennis EA. A comprehensive classification system for lipids. J Lipid Res. 2005; 46:839–861. [PubMed: 15722563]

9. Fahy E, Subramaniam S, Murphy RC, Raetz CRH, Nishijima M, Shimizu T, Spener F, van Meer G, Wakelam MJ, Dennis EA. Update of the LIPID MAPS comprehensive classification system for lipids. J Lipid Res. 2009; 50(Suppl):S9–S14. [PubMed: 19098281]

10. Chester MA. Nomenclature of glycolipids. Pure Appl Chem. 1997; 69:2475–2487.

11. Sud M, Fahy E, Cotter D, Brown HA, Dennis EA, Glass CK, Merrill AH Jr, Murphy RC, Raetz CRH, Russell DW, Subramaniam S. LMSD: LIPID MAPS structure database. Nucleic Acids Res. 2007; 35:D527–D532. [PubMed: 17098933]

12. LIPID MAPS Nature Lipidomics Gateway. http://www.lipidmaps.org

13. IUPAC, website covering lipid nomenclature. http://www.chem.qmul.ac.uk/iupac/

14. Yetukuri L, Ekroos K, Vidal-Puig A, Oresic M. Informatics and computational strategies for the study of lipids. Mol BioSyst. 2008; 4:121–127. [PubMed: 18213405]

15. Buckingham, J. Dictionary of Natural Products on CD-ROM, Version 19.1. Chapman and Hall; London: 2010.

16. GenBank website. www.ncbi.nlm.nih.gov/genbank

17. Swiss-Prot protein knowledgebase website. www.expasy.ch/sprot

18. Ensemble website. www.ensembl.org

19. Caffrey M, Hogan J. LIPIDAT: a database of lipid phase transition temperatures and enthalpy changes. Chem Phys Lipids. 1992; 61:1–109. [PubMed: 1315624]

20. Oracle website. www.oracle.com

21. The IUPAC International Chemical Identifier (InChi). website. www.iupac.org/inchi

22. Weininger D. SMILES, a Chemical Language and Information-System. J Chem Inf Comp Sci. 1988; 28:31–36.

23. SMILES website. www.daylight.com/smiles/index.html

24. Yetukuri L, Katajamaa M, Medina-Gomez G, Seppänen-Laakso T, Vidal Puig A, Oresic M. Bioinformatics strategies for lipidomics analysis: characterization of obesity related hepatic steatosis. BMC Syst Biol. 2007; 1:e12.

25. Wheelock CE, Goto S, Yetukuri L, D'Alexandri FL, Klukas C, Schreiber F, Oresic M. Bioinformatics strategies for the analysis of lipids. Methods Mol Biol. 2009; 580:339–368. [PubMed: 19784609]

26. ChemAxon website. www.chemaxon.com

27. Jmol website. http://jmol.sourceforge.net

28. CambridgeSoft website. www.cambridgesoft.com

29. Cotter D, Maer A, Guda C, Saunders B, Subramaniam S. LMPD: LIPID MAPS proteome database. Nucleic Acids Res. 2006; 34:D507–D510. [PubMed: 16381922]

30. Fahy E, Sud M, Cotter D, Subramaniam S. LIPID MAPS online tools for lipid research. Nucleic Acids Res. 2007; 35:W606–W612. [PubMed: 17584797]

31. CPAN — Comprehensive Perl archive network. www.cpan.org

32. Functional Glycomics Gateway. www.functionalglycomics.org

33. Han X, Gross RW. Global analyses of cellular lipidomes directly from crude extracts of biological samples by ESI mass spectrometry: a bridge to lipidomics. J Lipid Res. 2003; 44:1071–1079. [PubMed: 12671038]

34. Han X, Gross RW. Shotgun lipidomics. Mass Spectrom Rev. 2005; 24:367–412. [PubMed: 15389848]

35. Ejsing CS, Duchoslav E, Sampaio J, Simons K, Bonner R, Thiele C, Ekroos K, Shevchenko A. Automated identification and quantification of glycerophospholipid molecular species by multiple precursor ion scanning. Anal Chem. 2006; 78:6202–6214. [PubMed: 16944903]

36. Dennis EA, Deems RA, Harkewicz R, Quehenberger O, Brown HA, Milne SB, Myers DS, Glass CK, Hardiman G, Reichart D, Merrill AH Jr, Sullards MC, Wang E, Murphy RC, Raetz CRH, Garrett TA, Guan Z, Ryan AC, Russell DW, McDonald JG, Thompson BM, Shaw WA, Sud M, Zhao Y, Gupta S, Maurya MR, Fahy E, Subramaniam S. A mouse macrophage lipidome. J Biol Chem. 2010; 285:39976–39985. [PubMed: 20923771]

37. Quehenberger O, Armando AM, Brown AH, Milne SB, Myers DS, Merrill AH, Bandyopadhyay S, Jones KN, Kelly S, Shaner RL, Sullards CM, Wang E, Murphy RC, Barkley RM, Leiker TJ, Raetz CRH, Guan Z, Laird GM, Six DA, Russell DW, McDonald JG, Subramaniam S, Fahy E, Dennis EA. Lipidomics reveals a remarkable diversity of lipids in human plasma. J Lipid Res. 2010; 51:3299–3305. [PubMed: 20671299]

38. Blanksby SJ, Mitchell TW. Advances in mass spectrometry for lipidomics. Annu Rev Anal Chem. 2010; 3:433–465.

39. Katajamaa M, Oresic M. Processing methods for differential analysis of LC/MS profile data. BMC Bioinformatics. 2006; 6:179–190. [PubMed: 16026613]

40. Haimi P, Uphoff A, Hermansson M, Somerharju P. Software tools for analysis of mass spectrometric lipidome data. Anal Chem. 2006; 78:8324–8331. [PubMed: 17165823]

41. Song H, Ladenson J, Turk J. Algorithms for automatic processing of data from mass spectrometric analyses of lipids. J Chromatogr B. 2009; 877(26):2847–2854.

42. Hartler J, Trötzmüller M, Chitraju C, Spener F, Köfeler HC, Thallinger GG. Lipid Data Analyzer: unattended identification and quantitation of lipids in LC–MS data. Bioinformatics. 2011; 27(4): 572–577. [PubMed: 21169379]

43. Deems, R.; Buczynski, MW.; Bowers-Gentry, R.; Harkewicz, R.; Dennis, EA. Detection and quantitation of eicosanoids via high performance liquid chromatography-electrospray ionization-mass spectrometry. In: Brown, HA., editor. Methods Enzymol. Vol. 432. Academic Press; San Diego: 2007. p. 59-82.

44. Song H, Hsu F, Ladenson J, Turk J. Algorithm for processing raw mass spectrometric data to identify and quantitate complex lipid molecular species in mixtures by data-dependent scanning and fragment ion database searching. J Am Soc Mass Spectrom. 2007; 18:1848–1858. [PubMed: 17720531]

45. Murphy RC, Fiedler J, Hevko J. Analysis of nonvolatile lipids by mass spectrometry. Chem Rev. 2001; 2:479–526. [PubMed: 11712255]
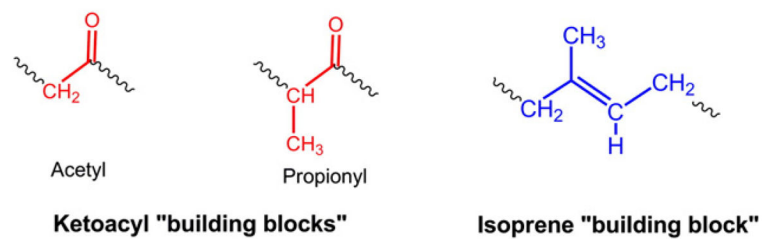
**Fig. 1.**

Lipid building blocks. The LIPID MAPS classification system is based on the concept of 2 fundamental biosynthetic "building blocks": ketoacyl groups and isoprene groups.
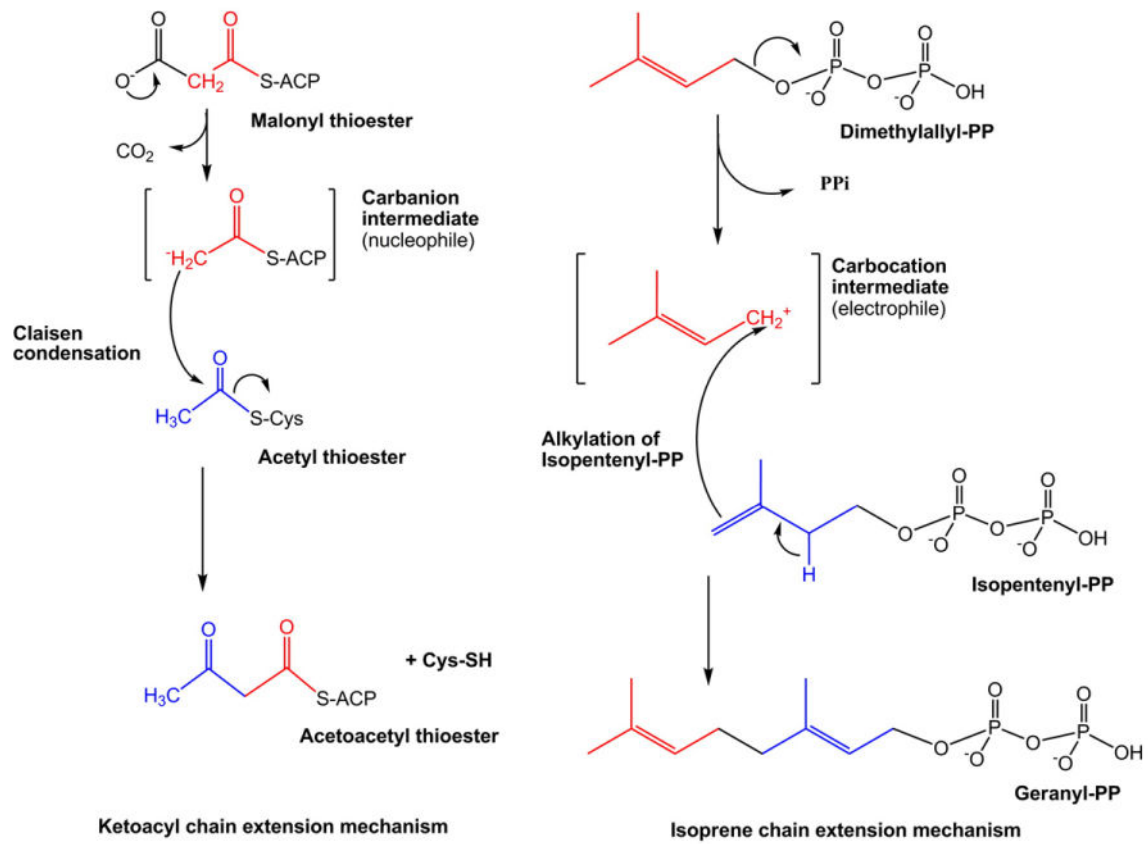
**Fig. 2.**
Mechanisms of lipid biosynthesis. Biosynthesis of ketoacyl- and isoprene-containing lipids proceeds by carbanion and carbocation-mediated chain extension, respectively.
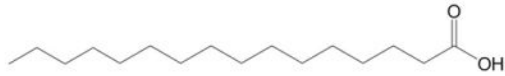
**Fatty Acyls**: hexadecanoic acid
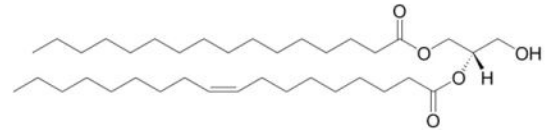
**Glycerolipids**: 1-hexadecanoyl-2-(9Z-octadecenoyl)-*sn*-glycerol
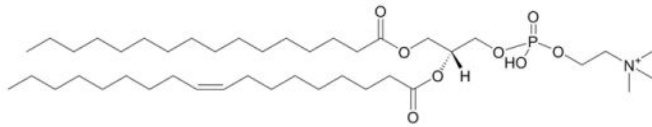
**Glycerophospholipids**:1-hexadecanoyl-2-(9Z-octadecenoyl)-*sn*-glycero-3-phosphocholine
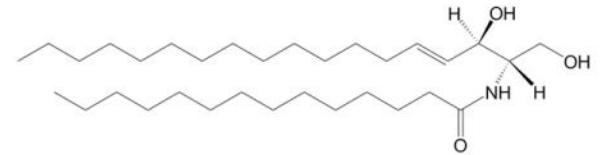
**Sphingolipids**: N-(tetradecanoyl)-sphing-4-enine
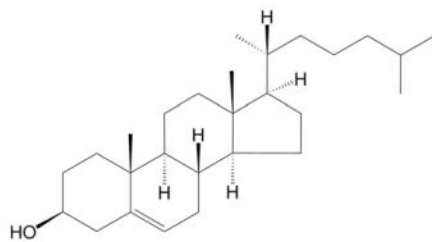
**Sterol Lipids**: cholest-5-en-3β-ol
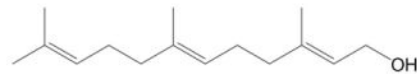
**Prenol Lipids**: 2E,6E-farnesol
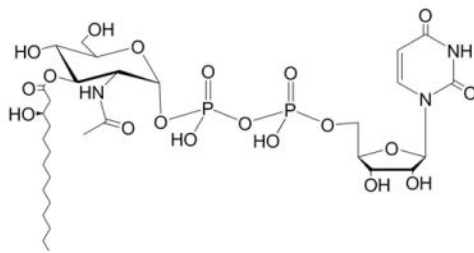
**Saccharolipids**: UDP-3-O-(3R-hydroxy-tetradecanoyl)-αD-N-acetylglucosamine
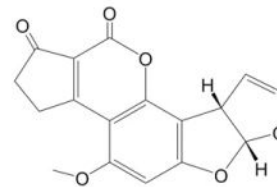
**Polyketides**: aflatoxin B1

**Fig. 3.**
Examples of lipid categories. Representative structures from each of the 8 LIPID MAPS lipid categories.
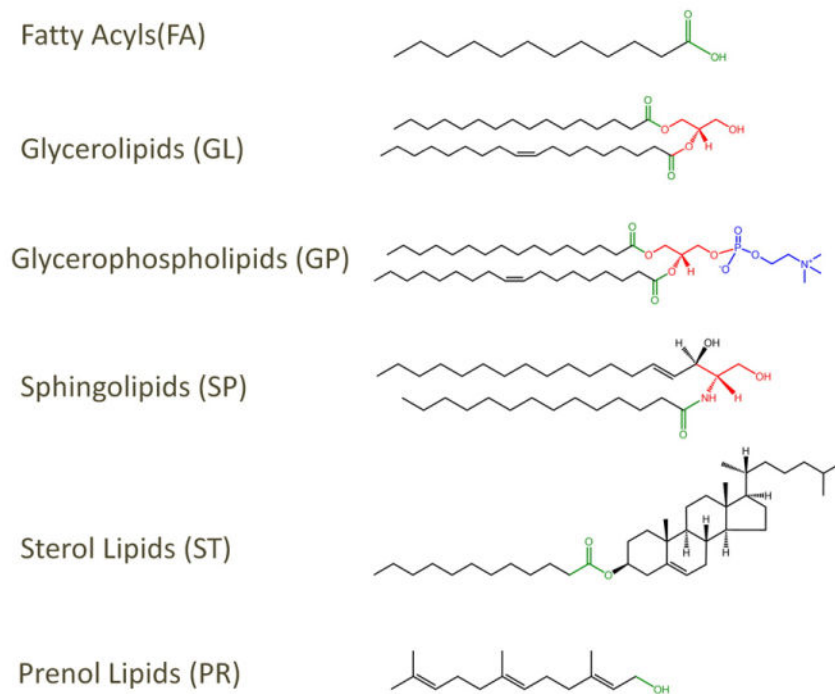
**Fig. 4.**
Lipid structure representation. Examples of structures from a number lipid categories in which acyl or prenyl chains are oriented horizontally with the terminal functional group (in green) on the right and the unsubstituted "tail" on the left.
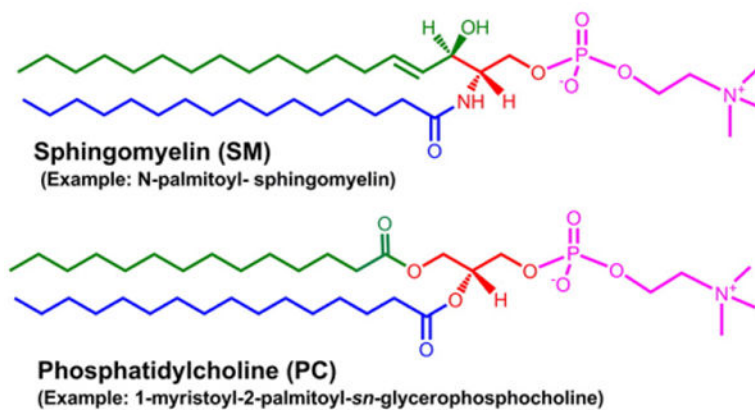
**Fig. 5.**
Structural similarity of PC and SM. Orientation of examples of phosphatidylcholine (PC) and sphingomeyelin (SM) structures highlights their similarity. The sphingosine "backbone" of SM is biosynthetically derived from palmitoyl-CoA (green) and serine (red) whereas PC contains a glycerol core (red) with an additional acyl chain (green). Both molecules contain a phosphocholine (pink) headgroup.

**Fig. 6.**
Searching the LIPID MAPS structure database. A selection of screen shots showing various options for searching LIPID MAPS Structure Database (LMSD).
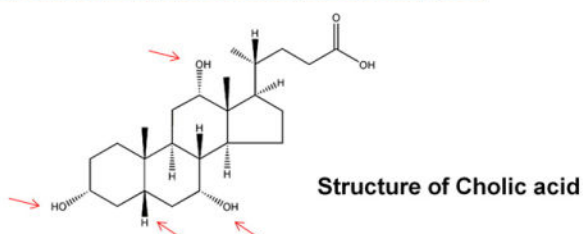
**Fig. 7.**
Use of InchIKeys in lipid structure searching. The first 14 characters of the InChIKey main layer (in red) define the molecular formula and bond connectivity of the molecule. In the case of cholic acid, searching the database with "BHQCQFFYRZLCQQ" will retrieve all 16 bile acid stereoisomers corresponding to the 3 hydroxyl groups and C5 stereocenter.
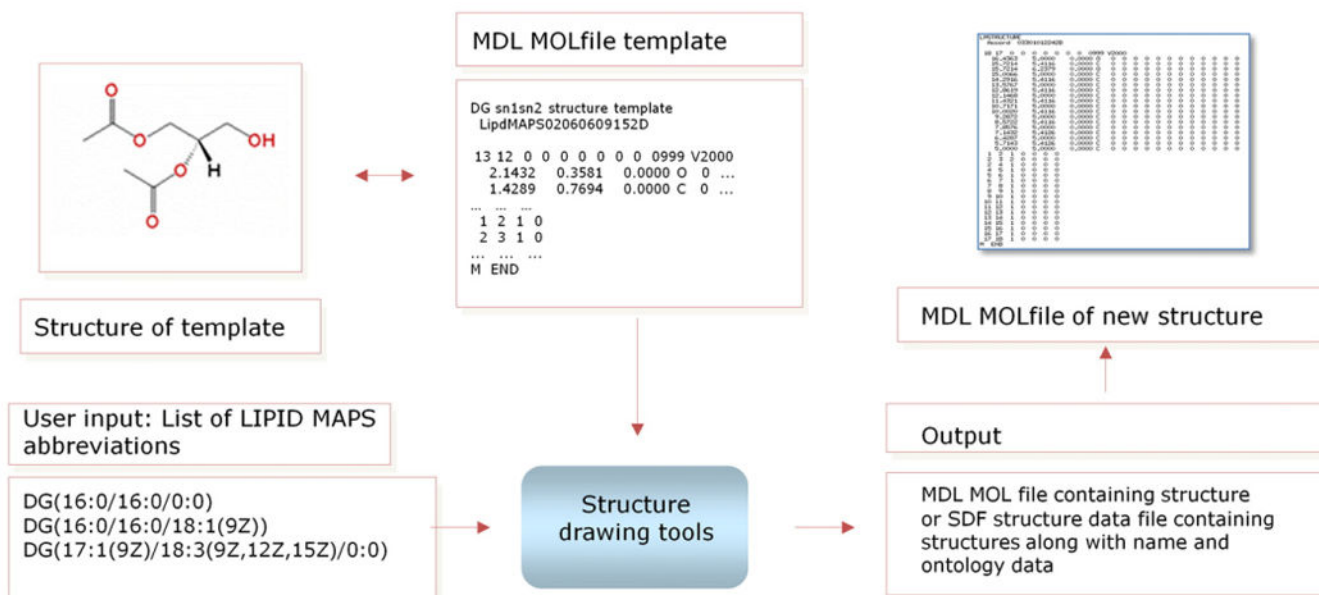
**Fig. 8.**
Overview of LIPID MAPS structure data generation methodology. A set of MOLfile templates are used as a starting point for the lipid structure drawing programs which in turn programmatically manipulate these molfiles to extend radyl chains and/or add various functional groups. The drawing programs are capable of accepting input either from online forms on the LIPID MAPS website or from lists of abbreviations in command-line mode.

**Fig. 9.**
LIPID MAPS structure drawing tools. A selection of screen shots showing the online structure drawing tools and options on the LIPID MAPS website.
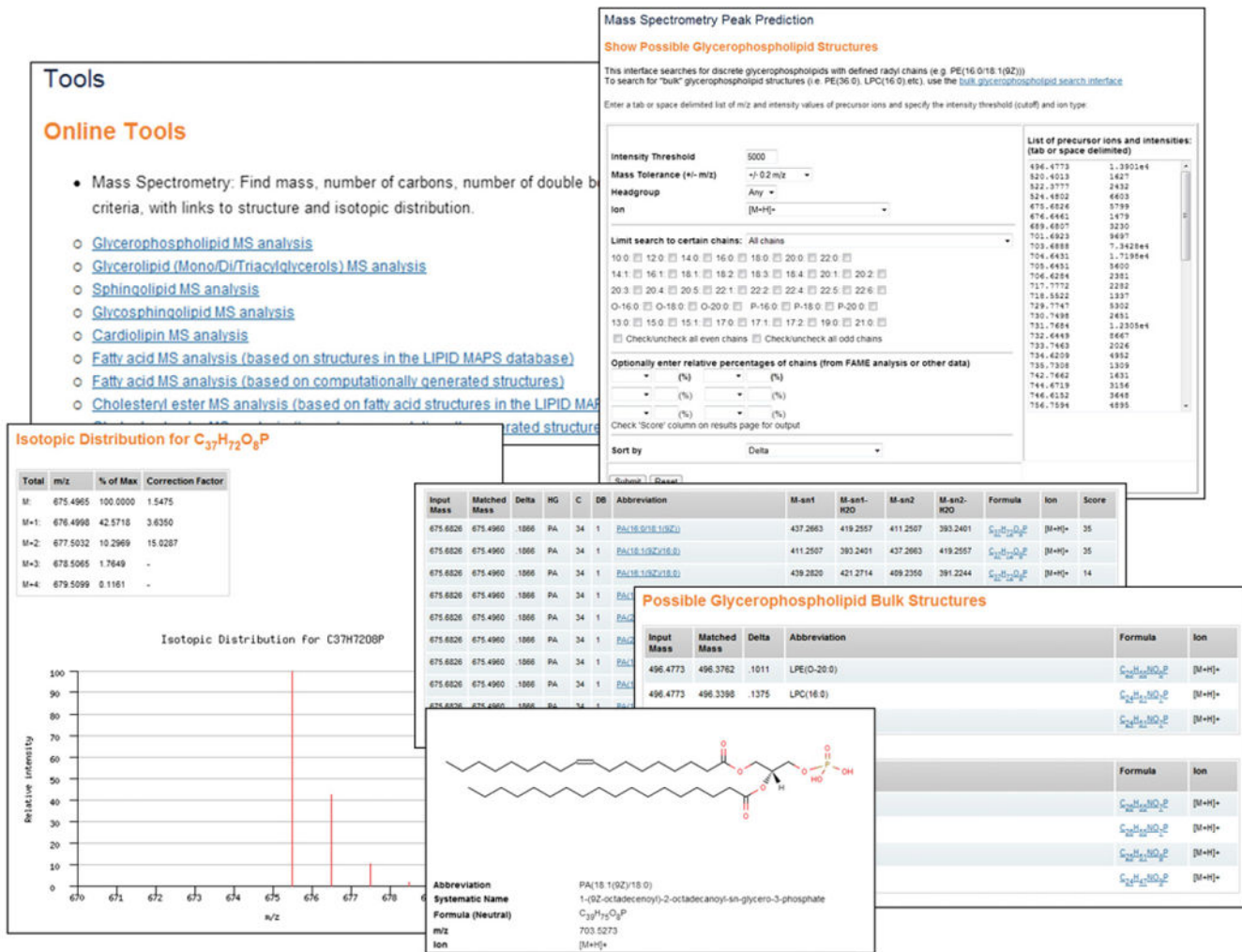
**Fig. 10.**
LIPID MAPS MS prediction tools. Some screen shots showing the online MS prediction tools on the LIPID MAPS website. Several options are available for defining both the search parameters and output format. Results are linked to isotopic distribution profiles and discrete structures (where applicable).

**Table 1**

Anatomy of the LIPID MAPS identifier (LM_ID) which encodes lipid classification within the identifier.

| Characters | Description | Example |
| --- | --- | --- |
| 1–2 | Fixed database designation | LM |
| 3–4 | 2-letter category code | FA |
| 5–6 | 2-digit class code | 03 |
| 7–8 | 2-digit subclass code | 02 |
| 9–12[a] | Unique 4-character identifier within subclass | 0016 |

[a]Certain lipid subclasses (e.g. prenol sesquiterpenes) contain a 4th level of classification in which case characters 9–10 define "Class level 4" and characters 11–16 define the unique identifier for that 4th level class.

**Table 2**

Abbreviations and examples of glycerophospholipid main classes.

| Class | Abbreviation | Examples |
|---|---|---|
| Glycerophosphocholines | PC (LPC for lyso species) | PC(P-16:0/18:2(9Z,12Z)) |
| Glycerophosphoethanolamines | PE (LPE for lyso species) | PE(O-16:0/20:4(5Z,8Z,11Z,14Z)) |
| Glycerophosphoserines | PS (LPS for lyso species) | PS(16:0/18:1(9Z)) |
| Glycerophosphoglycerols | PG (LPG for lyso species) | – |
| Glycerophosphates | PA (LPA for lyso species) | PA(16:0/0:0) or LPA(16:0) |
| Glycerophosphoinositols | PI (LPI for lyso species) | PI(18:0/18:0) |
| Glycerophosphoinositol monophosphates | PIP | – |
| Glycerophosphoinositol bis-phosphates | PIP2 | – |
| Glycerophosphoinositol tris-phosphates | PIP3 | – |
| Glycerophosphoglycerophosphoglycerols (Cardiolipins) | CL | CL(1′-[16:0/18:1(11Z)],3′-[16:0/18:1(11Z)]) |
| Glycerophosphoglycerophosphates | PGP | – |
| Glyceropyrophosphates | PPA | – |
| CDP-glycerols | CDP-DG | – |
| Glycosylglycerophospholipids | [glycan]-GP | – |
| Glycerophosphoinositolglycans | [glycan]-PI | – |
| Glycerophosphonocholines | PnC | – |
| Glycerophosphonoethanolamines | PnE | – |