Development of Computational Methods for Systematic Analysis of Lipids and Lipidomes

thesis submitted for the title

Doctor of Natural Science

(Dr. rer. nat.)

by

Chakravarthy Marella

from Gunturu, India

at

Department of Chemistry

Faculty of Mathematics, Informatics and Natural Sciences

University of Hamburg

April 20, 2017

Hamburg

Evaluators

Dr. Dominik Schwudke

Prof. Andrew Torda

Date of Oral Defense: June 9, 2017
Approved for Publication: June 14, 2017

Contents

Publications

1. Chakravarthy Marella, Andrew E. Torda, and Dominik Schwudke. “The LUX Score: A Metric for Lipidome Homology.” PLOS Comput Biol 11, 9:e1004511. doi:10.1371/journal.pcbi.1004511.

2. Correa Wilmar, Marcela Manrique-Moreno, Jochen Behrends, Edwin Patiño, Chakravarthy Marella, Carlos Peláez-Jaramillo, Patrick Garidel, Thomas Gutsmann, Klaus Brandenburg and Lena Heinbockel. Galleria Mellonella Native and Analogue Peptides Gm1 and ΔGm1. II) Anti-Bacterial and Anti-Endotoxic Effects. Biochimica et Biophysica Acta (BBA) - Biomembranes 1838:2739–44. doi:10.1016/j.bbamem.2014.07.005.

Table 1: List of Abbreviations

CACTVS	Chemical Algorithms Construction, Threading and Verification System
LIPID MAPS	LIPID Metabolites and Pathways Strategy
LMSD	LIPID MAPS Structure Database
SDF	Structure Data File
SMILES	Simplified Molecular Input Line Entry Specification
PCA	Principal Component Analysis
PC1	Principal Component 1
PC2	Principal Component 2
SMILIGN	SMILES Multiple Sequence Alignment
LUX	Lipidome jUXtaposition
CL	Cardioplipins
Cer	Ceramides
DAG	Diacylglycerol
TAG	Triacylglycerol
GSL	Glycosphingolipids
IPC	Inositol phosphorylceramides
MIPC	Mannose-inositolphospho-ceramide
M(IP)$_2$C, M(IP)2C	Mannose-bis(inositolphospho)ceramide
PA	Phosphatidic acids
PC	Phosphatidylcholines
PE	Phosphatidylethanolamines
PG	Phosphatidylglycerols
PI	Phosphatidylinositols
PS	Phosphatidylserines
CerPE	Phosphorylethanolamine ceramides
sn	stereospecific numbering
HexCer	Hexosyl Ceramides
LCB	Long Chain Base
MDL MOL	Molecular Design Limited MOL format

Zusammenfassung

Lipide sind Botenstoffe, Energiespeicher-Moleküle und eine strukturelle Komponente von biologischen Membranen; gehören somit zu den wichtigsten Biomolekülen. Bekannte Vertreter sind z.B. Cholesterin, Vitamin A, Omega-3-Fettsäure, Sphingomyelin. Angesichts der Tatsache, dass Lipide diese vielfältigen Rollen spielen, ist es keine Überraschung, dass sie bei vielen Krankheiten, einschließlich Alzheimern und Krebs, verändert sind. Organismen wie Hefe, Fruchtfliege werden als Modelle verwendet, um den Stoffwechsel zu verstehen, aber ihre Lipidstrukturen unterscheiden sich von denen des Menschen.

Lipide können durch eine Reihe biochemischer Techniken bestimmt (oder gemessen) werden, die als "Lipidomik" zusammengefasst werden. Die hohe Durchsatzrate der aktuellen Lipidomik-Plattformen erlaubt zur Identifizierung von Hunderten von Lipiden aus einem gegebenen biologischen Material, das als Lipidom bezeichnet wird. Die Fortschritte in der Massenspektrometrie-Instrumentierung führten zu einer Erhöhung der Anzahl katalogisierter Lipide, was sich in der erweiterten LIPID-MAPS-Strukturdatenbank widerspiegelt, aber Berechnungsmethoden zur Analyse dieser Hochdurchsatzdaten sind begrenzt. Vor allem gibt es keine Methoden, die Lipidstrukturunterschiede verwenden, um Lipidome von Modellorganismen und Menschen zu vergleichen.

Ziel dieser Studie ist es, rechnerische Methoden zur Verfügung zu stellen, die einen besseren Einblick in die Lipidomik-Daten bieten. In dieser Studie wurde ein metrisches Raummodell von Lipiden und Lipidomen entwickelt, das aus drei Schritten besteht. Zuerst wird eine Stringdarstellung von Lipiden, SMILES, untersucht. Zweitens wurden Methoden zur Bestimmung der strukturellen Ähnlichkeit von Lipid-SMILES getestet. Drittens werden Strategien zur Visualisierung und Gegenüberstellung von Lipidomen vorgestellt. Lipidom-Nebeneinanderstellung (LUX), die als Teil dieser Studie entwickelt wurde, zielt auf den globalen Vergleich von Lipidprofilen, insbesondere zwischen Modellorganismen und Menschen, ab. Diese Studie ergänzt die vorhandenen Downstream-Datenanalyse-Techniken, indem sie LUX-Score als eine neue Maßnahme der Lipidom-Divergenz vorschlägt.

Abstract

Lipids are important biomolecules. They are signal messengers, energy storage molecules and a major structural component of biological membranes, e.g. cholesterol, vitamin A, ω-3-fatty acid, sphingomyelin etc. Given that lipids play these diverse roles, it is no surprise they are disrupted in many diseases, including Alzheimer's and cancer. Organisms such as yeast, fruit fly are used as models to understand disease metabolism but their lipid structures are different from humans.

Lipidomics is the study of the structure and function of the complete set of lipids (the lipidome) produced in a given cell or organism as well as their interactions with other lipids, proteins and metabolites. The advances in mass spectrometry based lipidomics has led to an increase in the number of cataloged lipids, which is reflected in the expanded LIPID MAPS Structure Database, but computational methods for analyzing this high-throughput data are limited. Expecially, there are no methods that use lipid structure differences to compare lipidomes of model organisms and humans. Statistical methods such as correlation coefficient and multi variate regression models are increasingly used to find patterns in lipidomics datasets, but the problem is that only lipid abundances (rather than structures) were used for comparison and clustering.

The aim of this study is to provide computational methods that offer better insights into the lipidomics data. A metric space model of lipids and lipidomes was developed in this study, which was achieved in three steps. First, a string representation of lipids, SMILES is throughly examined. Second, methods to determine structural similarity from lipid SMILES were tested. Third, strategies to visualize and juxtapose lipidomes are presented. Lipidome juxtaposition (LUX) score that was developed in this study is aimed at global comparison of lipid profiles, especially between model organisms and humans. This study complements the existing down stream data analysis techniques by suggesting LUX score as a new measure of lipidome divergence.

Introduction

What is a 'lipid'? or the difficulty with definition

The term 'lipid' has no universally accepted definition [1–4]. In the International Union of Pure and Applied Chemistry (IUPAC) nomenclature, chemical compounds were grouped by the presence of a distinguishing structural feature or a functional group [5]. Surprisingly for lipids, the grouping is not based on structure but on the basis of solubility. IUPAC defines lipids as "substances of biological origin that are soluble in non-polar solvents" [5]. However, the solubility-based definition is not adequate because some lipids (example, gangliosides) are soluble in polar solvents [6]. A definition based on function and biosynthesis was presented by Christie *et al.* "lipids are fatty acids and their derivatives, and substances related biosynthetically or functionally to these compounds" [7]. But this definition does not include steroid hormones and polyketides. Recently, Fahy *et al.* defined lipids as "hydrophobic or amphipathic small molecules that may originate entirely or in part by carbanion based condensations of thioesters (fatty acyls, glycerolipids, glycerophospholipids, sphingolipids, saccharolipids, and polyketides) and/or by carbocation-based condensations of isoprene units (prenol lipids and sterol lipids)" [8]. The last definition encompasses many heterogeneous organic compounds and was formulated for lipid classification [Table 1.1].

Lipidomics is relatively new - a historical overview

The term 'lipidome analysis' was introduced by Kishimoto *et al.* [10] to describe an analytical method for determining and comparing the "changed mass of multiple lipid species". Later, Han *et al.* expanded the scope of lipidomics as the use of "multiple techniques to quantitate the hundreds of chemically distinct lipids in cells and determine the molecular mechanisms through which they facilitate cellular function" [11]. Methods to identify proteins and genes from biological samples are fairly advanced, even though methods for lipid analysis are not as advanced, they are fast improving [12–14]. Earlier studies used fluorescent dyes, Thin Layer Chromatography (TLC) and enzyme kits to identify lipids [10]. These methods were

Table 1.1: Lipid classification by Fahy *et al.* [8, 9]

A		Fatty Acyls
	1	Fatty Acids and Conjugates
	2	Octadecanoids
	3	Eicosanoids
	4	Docosanoids
	5	Fatty alcohols
	6	Fatty aldehydes
	7	Fatty esters
	8	Fatty amides
	9	Fatty nitriles
	10	Fatty ethers
	11	Hydrocarbons
	12	Oxygenated hydrocarbons
	13	Fatty acyl glycosides
	00	Other Fatty Acyls
B		Glycerolipids
	1	Monoradylglycerols
	2	Monoacylglycerols
	3	Monoalkylglycerols
	4	Mono-(1Z-alkenyl)-glycerols
	5	Diradylglycerols
	6	Triradylglycerols
	7	Glycosylmonoradylglycerols
	8	Glycosyldiradylglycerols
	00	Other Glycerolipids
C		Glycerophospholipids
	1	Glycerophosphocholines
	2	Glycerophosphoethanolamines
	3	Glycerophosphoserines
	4	Glycerophosphoglycerols
	5	Glycerophosphoglycerophosphates
	6	Glycerophosphoinositols
	7	Glycerophosphoinositol monophosphates
	8	Glycerophosphoinositol bisphosphates
	9	Glycerophosphoinositol trisphosphates
	10	Glycerophosphates
	11	Glyceropyrophosphates
	12	Glycerophosphoglycerophosphoglycerols
	13	CDP-Glycerols
	14	Glycosylglycerophospholipids
	15	Glycerophosphoinositolglycans
	16	Glycerophosphonocholines
	17	Glycerophosphonoethanolamines
	18	Di-glycerol tetraether phospholipids
	19	Glycerol-nonitol tetraether phospholipids
	20	Oxidized glycerophospholipids
	00	Other Glycerophospholipids
D		Sphingolipids
	1	Sphingoid bases
	2	Ceramides
	3	Phosphosphingolipids
	4	Phosphonosphingolipids
	5	Neutral glycosphingolipids
	6	Acidic glycosphingolipids
	7	Basic glycosphingolipids
	8	Amphoteric glycosphingolipids
	9	Arsenosphingolipids
	00	Other Sphingolipids
E		Sterol Lipids
	1	Sterols
	2	Steroids
	3	Secosteroids
	4	Bile acids and derivatives
	5	Steroid conjugates
	00	Other Sterol lipids
F		Prenol Lipids
	1	Isoprenoids
	2	Quinones and hydroquinones
	3	Polyprenols
	4	Hopanoids
	00	Other Prenol lipids
G		Saccharolipids
	1	Acylaminosugars
	2	Acylaminosugar glycans
	3	Acyltrehaloses
	4	Acyltrehalose glycans
	5	Other acyl sugars
	00	Other Saccharolipids
H		Polyketides
	1	Linear polyketides
	2	Halogenated acetogenins
	3	Annonaceae acetogenins
	4	Macrolides and lactone polyketides
	5	Ansamycins and related polyketides
	6	Polyenes
	7	Linear tetracyclines
	8	Angucyclines
	9	Polyether antibiotics
	10	Aflatoxins and related substances
	11	Cytochalasins
	12	Flavonoids
	13	Aromatic polyketides
	14	Non-ribosomal peptide/polyketide hybrids
	00	Other Polyketides

time consuming and identification is only possible at the level of lipid class [Fig. 1.1]. Electron Spray Ionization coupled with Mass Spectrometry (ESI-MS) was used for the separating lipids to the level of molecular species [15]. ESI-MS work flow often involves a solvent extraction step but Matrix Assisted Laser Desorption Ionization (MALDI) technique does not require solvent extraction [16]. In the last two decades, the technological advances in mass spectrometry instrumentation had a positive impact on lipidomics in two complementary ways a. an increase in the number of distinct lipid species identified from a given sample and b. improved structure characterization, such as the acyl chain composition [Fig. 1.1] [17, 18]. Recent methods, especially the 'shotgun' lipidomics work flow allowed the quantification of more than 250 lipid species in a few minutes [Fig. 1.2] [19]. The lipid extraction procedure in the work flow [Fig. 1.2] is different for polar and non-polar lipids because the protocol depends on the lipid class [20]. Some lipids must be chemically modified to facilitate the identification using ESI-MS, example, cholesterol and similar lipids are acetlylated [21].

The ability to quickly identify lipids from biological material is reflected in the increased number of lipidomics publications [Fig. 1.3]. The lipids of Human Immuno deficiency Virus [24], yeast [23,25] and an epithelial cell line during differentiation [26] were characterized with ESI-MS. The high-throughput nature of current lipidomics work flow results in thousands of mass spectra from a single sample run in the ESI-MS instrument [27–29]. The process of identifying the lipids from the *m/z* spectra is automated through software pipelines such as mzMine, LipidXplorer, ALEX [30–35]. The sensitivity of lipid detection from biological samples is currently limited to 10-20 lipid classes and a few hundred lipid species, but these numbers are expected to increase [36].

1.1 Use of Model Organisms in Lipidomics

Model organisms are employed in biological research because of their 1. shorter life cycle facilitating multi-generational experiments and 2. the relative simplicity of their genome that allows targeted gene modifications [37, 38]. For example, Klose *et al.* employed yeast as a model organism to study the physical properties of lipid membranes [39, 40]. Santos *et al.* used yeast as a model to study the function of fatty acid elongase enzymes [25]. The nucleotide sequences of the yeast

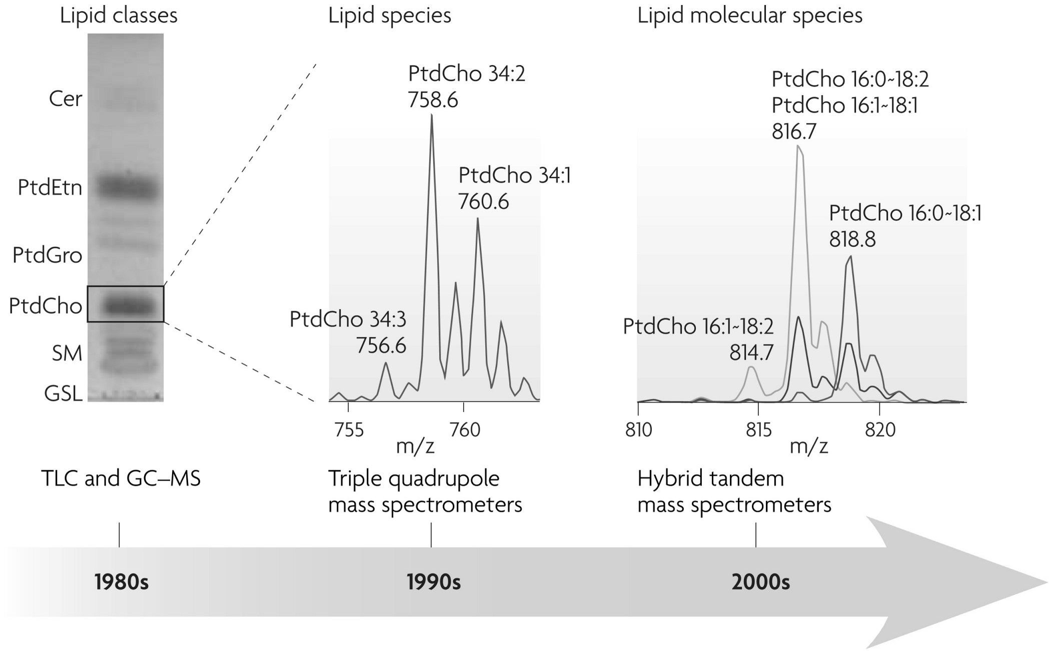


Figure 1.1: Timeline of improvements in lipid identification. In the 1980's, lipid classes were identified with TLC and GC-MS. In the 1990's, lipid-classes were separated to the level of lipid molecular species with the use of triple quadrupole mass spectrometers (for example, PtdCho class is separated to PtdCho 34:1 (34 carbon atoms and 1 double bond), 34:2 and 34:3 species based on the *m/z* ratio spectrum (*m* mass, *z* charge). In the 2000's, it was possible to identify the acyl chain composition for each lipid species with the use of tandem mass spectrometers. TLC - Thin Layer Chromatography; GC-MS - Gas Chromatography - Mass Spectrometry; Cer - Ceramide, PtdEtn (or PE) Phosphatidyl Ethanolamine; PtdGro (or PG) - Phosphatidyl Glycerols; SM - Sphingomyelin; GSL - Glycero-Sphingolipids. Extracted from Shevchenko and Simons [22].

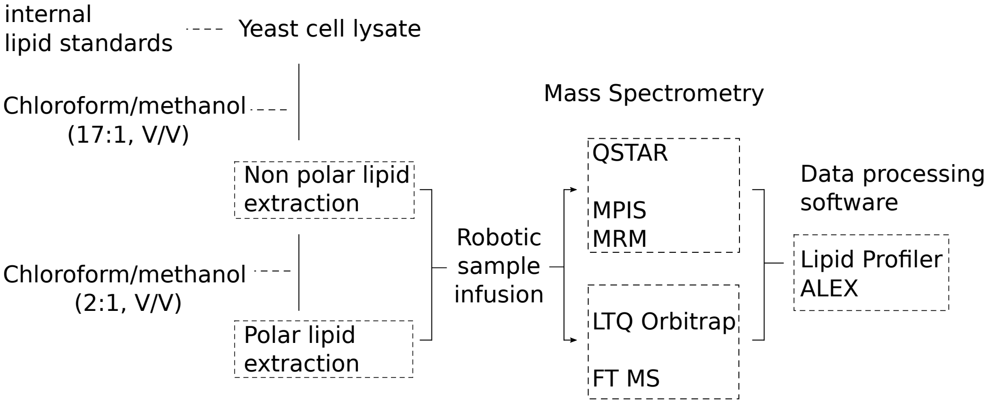


Figure 1.2: Overview of Shotgun lipidomics workflow. Internal lipid standards were added to cell lysate for quantification. QSTAR and LTQ Orbitrap are instrument models. MPIS - Multiple Precursor Ion Scanning; MRM - Multiple Reaction Monitoring; FT MS - Fourier Transform Mass Spectrometry. Modified from Ejsing *et al.* [23].

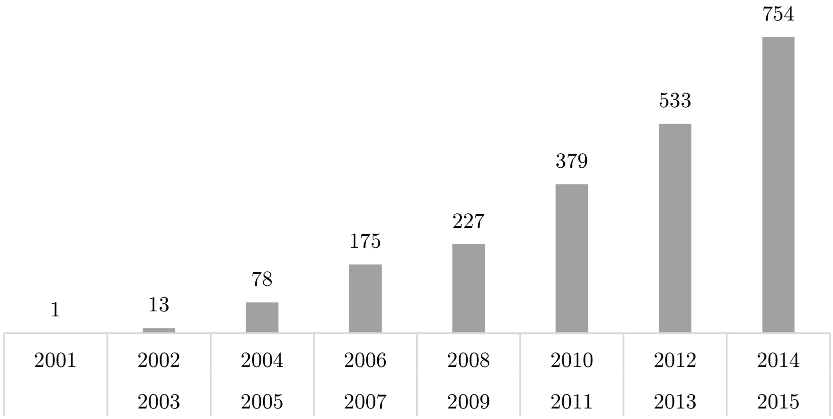


Figure 1.3: Number of publications per year with the word 'Lipidom[e][ics]' in the title or abstract section of a manuscript. Collected from Web of Science database.

fatty acid elongase genes Elo1, Elo2 and Elo3 have sequence similarity with the mammalian genes Cig30, Ssc1, and Ssc2, which makes the knock-out experiments in yeast relevant to mammals [41, 42]. Lipids were studied in pathogens such as *Candida* [43], *Trypanosoma* [44], *Toxoplasma* [45], *Leishmania* [46] and *Mycobacterium tuberculosis* [47] for their role in regulating the disease progression. *Caenorhabditis elegans* (round worm) is used as a model organism to visualize lipid droplets because of its transparent body [48]. Sterols in fruit fly have been studied for their role in maturation from larvae to adult [49]. Tortoriello *et al.* suggested fruit fly as a model to study lipid signaling pathways [50, 51].

Although model organisms like yeast and fruit fly are routinely used in lipidomics, the regulatory enzymes [Fig. 1.4] and lipid structures are different, especially the acyl chain length, the degree of unsaturation and the hydrocarbon branching pattern [Table 1.2] [52–56]. For instance, the highly abundant lipid, cholesterol in mammals is not present as structural component of membranes in yeast and fruit fly, but they have a structurally similar molecule, ergosterol [Fig. 1.5]. The membrane lipid, sphingomyelin (named after the white fatty substance surrounding nerve cell axons) is found in mammals, but a structural analogue ceramide phosphoethanolamine is present in fruit fly [Fig. 1.6] [57]. The long chain base (LCB) of ceramides is an 18 carbon length sphingosine in yeast and mammals, but it is shorter (14 carbon length) in fruit fly [49, 58]. The round worm has a unique branched chain head group with odd-number of carbon atoms [59, 60]. Lipid structures also depend on the habitat temperature, example, thermally acclimatized

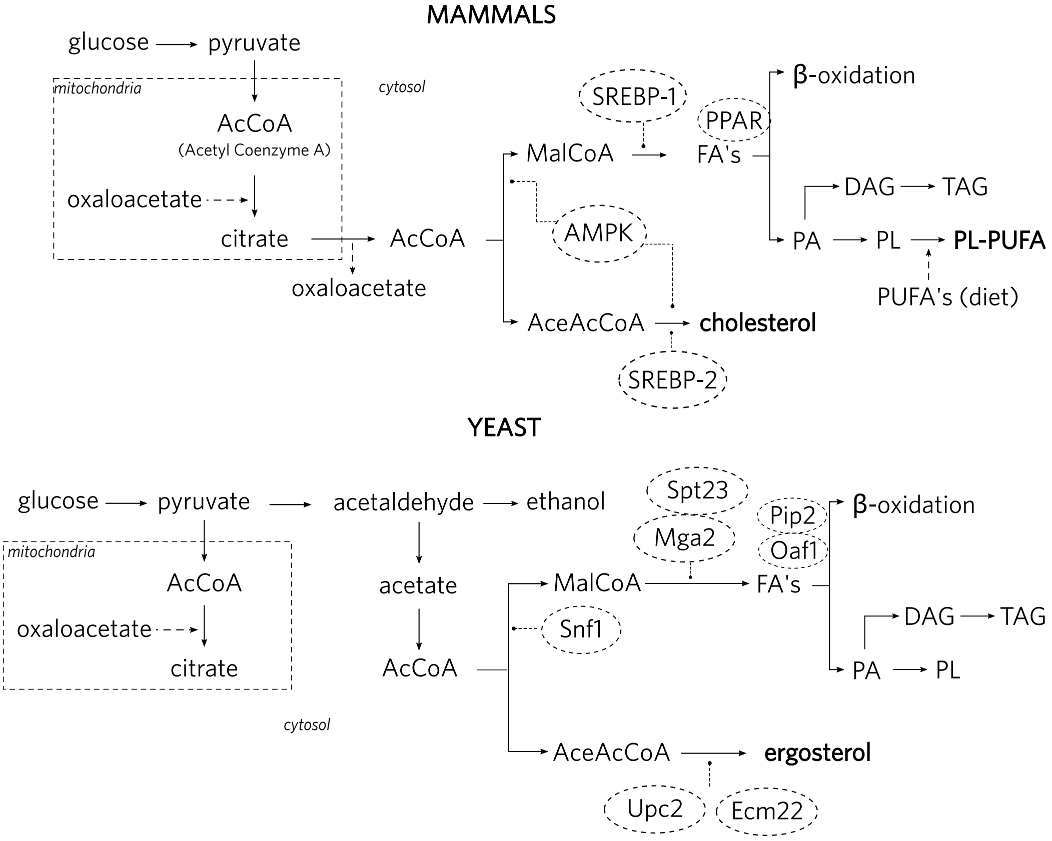


Figure 1.4: Lipid metabolism, regulation in mammals and yeast. The substrates of lipid biosynthesis such as AceAcCoA, MalCoA are common for the two systems. The yeast has more enzymes involved in regulation (Spt23, Mga2, Pip2, Oaf1, Snf1, Upc2, Ecm22) in comparison to mammals (SREBP-1, PPAR, AMPK, SREBP-2). Cholesterol is major sterol in mammals but in yeast, it is ergosterol. Mammals take up fatty acids through diet and these are incorporated into the different lipid pools (here only illustrated to the PL-pool). AceAcCoA - Aceto Acetyl Coenzyme A; MalCoA - Malonyl Coenzyme A; FAs – fatty acids; PA – phosphatidic acid; DAG – diacylglycerols; TAG – triacylglycerols; PL – phospholipids; PL-PUFA – phospholipids containing poly-unsaturated fatty acids. Modified from Nielsen [63].

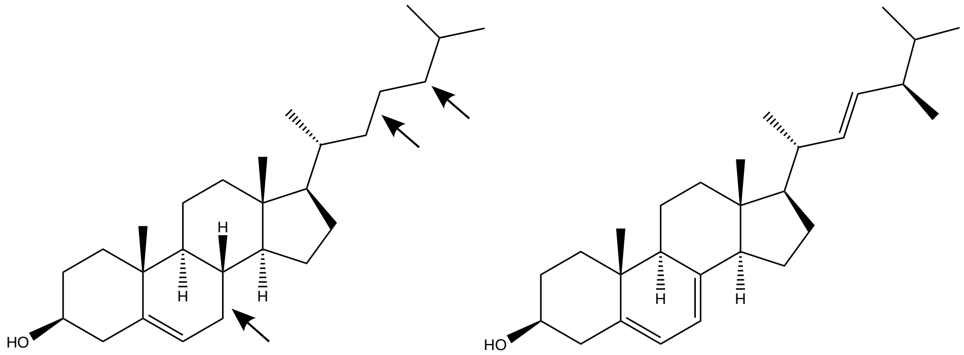
organisms have a higher proportion of ether-linked phospholipids and their lipids are more saturated [61, 62].

Given these variations, the focus of this thesis is to develop methods for systematically measuring the lipid structure differences between organisms. I will consider the problem of comparing lipid structures under two sections a. the different approaches to represent structures and b. algorithms to calculate structure similarity.

Table 1.2: Major lipid classes in selected organisms

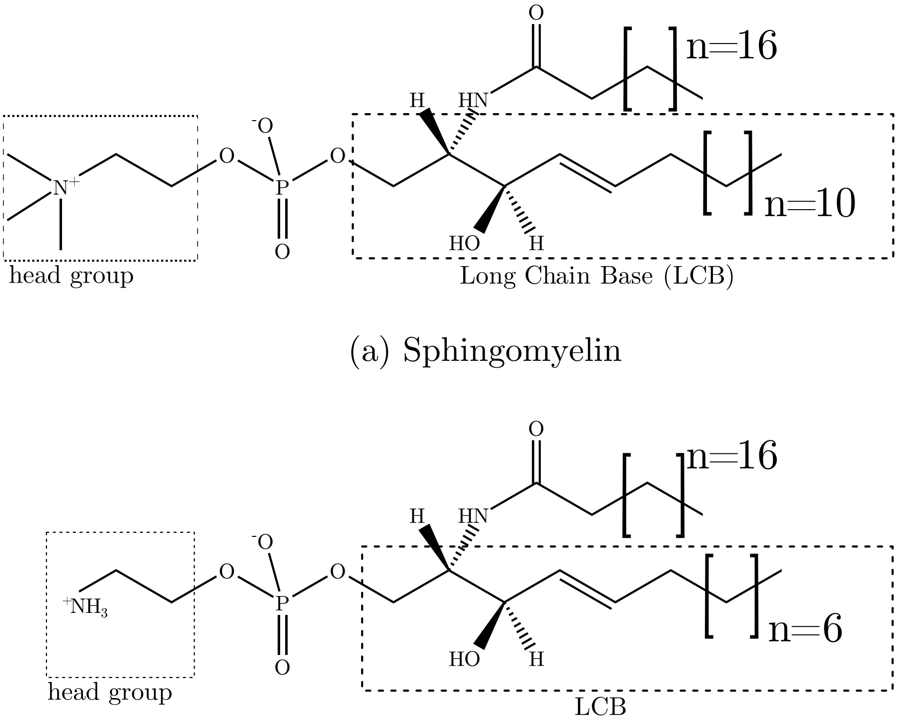
	Organism	Phospholipids	Sterols	Sphingolipids
1	*Saccharomycetes cervicae* (yeast)	PI, PE, PC, PA, PS, PG	Ergosterol, Ergostedienol	IPC, LCB(C18)
2	*Caenorhabditis elegans* (round worm)	PC, PE, PI, PS, PG, PA, high abundance of PUFA		SM, iso-branch LCB(C17)
3	*Drosophila melanogaster* (fruit fly)	PE, PC, PI, PS, PG, PA, lacks PUFA	Ergosterol	CerPE, shorter LCB(C14)
4	Mammals	PC, PE, PI, PS, PG, PA	Cholesterol	SM, LCB(C18)

PI - Phoshatidyl Inositol; PE - Phosphatidyl Ethanolamine; PC - Phosphatidyl Choline; PA - Phosphatidic Acid; PS - Phosphatidyl Serine; PG - Phosphatidyl Glycerol; PUFA - Poly Unsaturated Fatty Acid; IPC - Inositol Phosphoryl Ceramide; LCB - Long Chain Base; SM - Sphingomyelin; C[14][17][18] - number of carbon atoms in LCB.



(a) Cholesterol (b) Ergosterol

Figure 1.5: The structure of major sterol in mammals, Cholesterol (a), in yeast and fruit fly, Ergosterol (b). Three regions of the Cholesterol structure were marked to indicate the difference with Ergosterol.



(a) Sphingomyelin

(b) Ceramide-phosphoethanolamine

Figure 1.6: The structure of important sphingolipid in mammals, Sphingomyelin (a) and its closest structural counterpart in fruit fly, ceramide-phosphoethanolamine (b). The LCB in mammals has 18 carbon atoms (C18) but in fruit fly, it has 14. In ceramide-phosphoethanolamine the head group is ethanolamine but in sphingomyelin it is choline.

1.2 Representation of Lipid Structures as Strings

The text book representation of molecules is a ball and stick model which represents the topology of a molecular structure [Fig. 1.7] [64, 65]. This model is also referred as graph representation, with nodes as atoms, and edges as chemical bonds between them. Numerous flat text file formats are available that provide rules for drawing molecule graphs consistently [66]. Molecular structures can also be represented with linear models (also referred to as line or string notation).

Linear models have a long history dating back to Wiswesser Line Notation in 1949 [67, 68]. The most popular linear representation format currently in use is SMILES (Simplified Molecular Input Line Entry Specification), although many more (ROSDAL, SLN etc.) are available [69–71]. String representations do not contain 3D coordinate information but are very popular in large databases such as the Chemical Abstracts Service (CAS) Registry. InChI (IUPAC International Chemical Identifier) is an extension of the IUPAC nomenclature of molecules [72, 73]. It is an identifier for molecules, similar to CAS Registry number, PubChem ID or LIPIDMAPS ID. Unlike other identifiers, the InChI string also functions like a line notation of a chemical structure [71].

SMILES is a linear chemical notation system to represent structures using plain text characters. The SMILES specification was developed by Weininger [74] for the purpose of database retrieval, substructure searching and property prediction models. SMILES notation was used for calculating the surface property of molecules by Ertl *et al.* [75]. Structure repositories such as ZINC [76], Drug Bank [77], ChEBI [78], PubChem [79] and LMSD [80] provide SMILES notation for molecules.

A chemical structure can have many valid SMILES representations. In Fig. 1.7, a ceramide molecule is shown in 2-Dimensional representation, followed by two valid SMILES representations. Notice that the first SMILES starts from omega carbon of the acyl chain (ω) and second from omega carbon of head group (Ω). To avoid maintaining two copies of the same molecule in a database, the concept of unique SMILES (also referred as canonical SMILES) was introduced by Weininger *et al.* [81, 82]. Programs to generate SMILES and canonical SMILES are made available

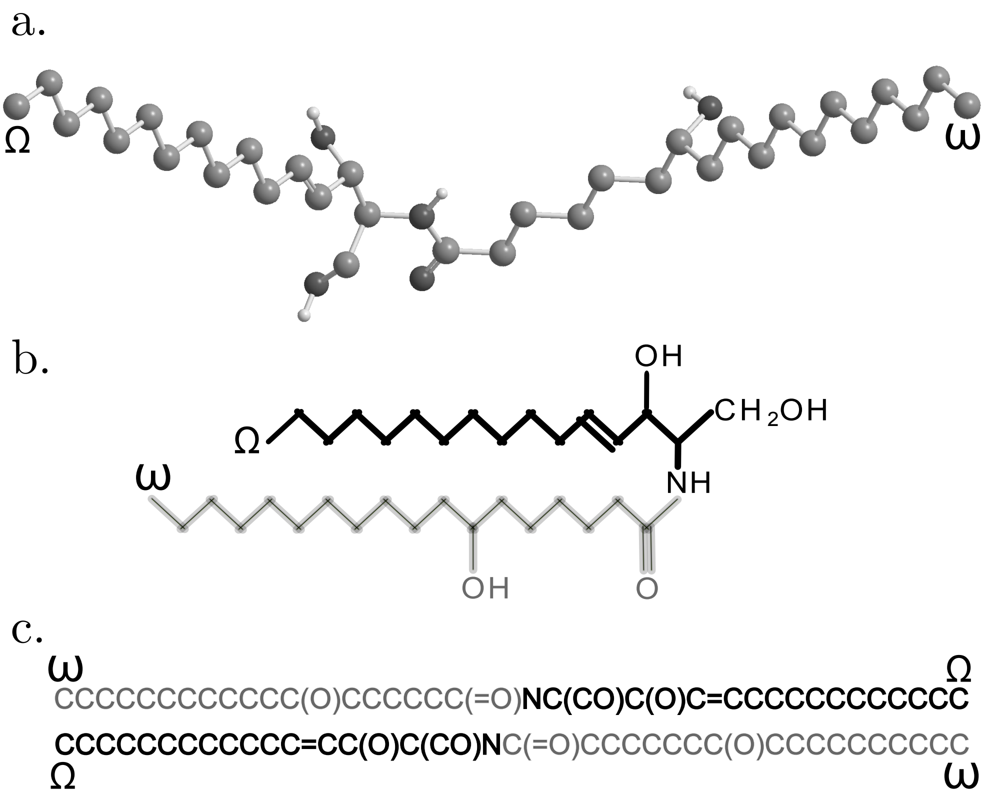

a.
Ω
ω
b.
OH
CH₂OH
Ω
ω
NH
OH
O
c.
ω
Ω
CCCCCCCCCCCC(O)CCCCCC(=O)NC(CO)C(O)C=CCCCCCCCCCCC
CCCCCCCCCCCC=CC(O)C(CO)NC(=O)CCCCCCC(O)CCCCCCCCCC
Ω
ω

Figure 1.7: Representation of a ceramide (Cer 34:1) structure. The molecule has 34 carbon atoms, one double bond and a total of three hydroxyl groups. The structure is shown in three ways. a. Ball and stick model b. Line drawing and c. SMILES. The head group is dark shaded and acyl chain is lighter. The farthest carbon from carboxyl end in acyl chain is indicated with omega symbols. The two SMILES representations in panel c are written 1. starting from acyl chain (ω) and 2. starting from head group (Ω) respectively.

by both proprietary vendors (like Chem3D software program) and by open source community, like CACTVS [83] and Open Babel [84].

Rules for defining SMILES strings have developed further since the first publication in 1988, but this has led to different and occasionally, conflicting proposals. For instance, Open Babel uses open SMILES specification by James *et al.* CACTVS algorithm is based on unique SMILES definition of Weininger *et al.* [81] but it is not up to date with latest specifications. SMILES notation is suitable for representing lipid structures because of its simplicity and readability but, the problem is - how to select the best specification suitable for lipids ?

The broader aim of this work may be the comparison of lipidomes, but the first step is finding an appropriate SMILES representation. This means that intuitively small changes in structure should lead to small changes in the SMILES strings. The first part of the results chapter [section 3.1] compares string from three SMILES-generation methods (1. Template SMILES 2. CACTVS canonical SMILES and 3. Open Babel canonical SMILES) based on a smaller set of lipids. In the second part of the results [section 3.2], only the most appropriate representation was used, but tested on a larger data set.

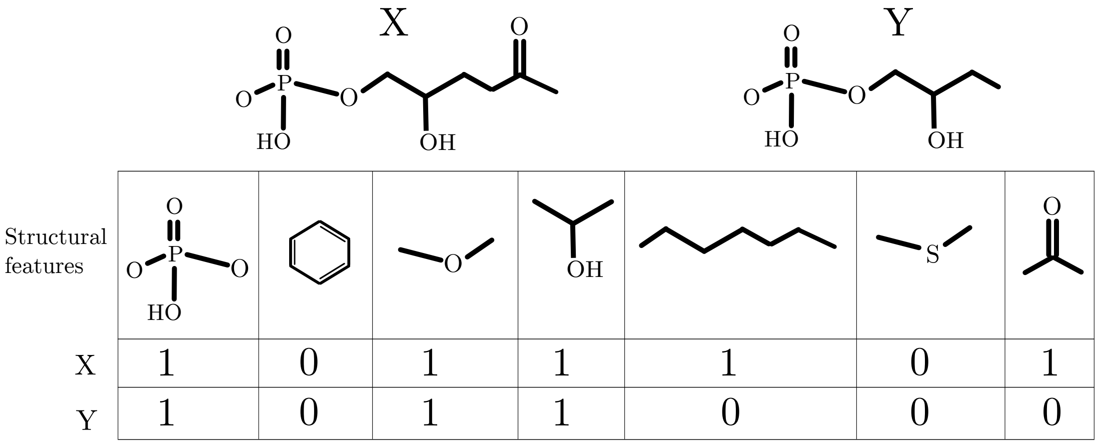
1.3 Algorithms to Measure Structure Similarity from Strings

Methods to determine structure similarity from linear representation of molecules were initially developed for database searches [85]. The process of a database search starts with a query structure and the aim is to retrieve a ranked list of similar molecules. This procedure involves matching the query structure to a database molecule. The class of algorithms that perform this task are referred as sub-structure matching algorithms. However, very large databases use precomputed pairwise similarity matrices to reduce the sub-structure matching time. In the absence of a query, the complete structure of a molecule is used for calculating pairwise similarity scores. Three sub-structure and three complete-structure comparison algorithms were tested in this study.

1.3.1 Molecular Fingerprints

Fingerprints are by definition unique to an individual, like a signature, but that definition is misleading when applied in the context of molecules. Fingerprints are a hashed version of structural features that are commonly used for comparing molecules in pharmacological research [85,86]. Depending on the nature of algorithm used to generate them, a fingerprint can be a physical descriptor (such as molecular weight), an atom coordinate or the connectivity [87,88]. A brief working principle of a fingerprinting procedure is illustrated in Fig. 1.8.

FP2 and FP3 fingerprinting algorithms of Open Babel software library convert SMILES to bit strings of length 1024 [84]. In the FP2 algorithm, molecules are broken to overlapping fragments of length 7. A hash function is used on each fragment, returning a number between 0 and 1023 which is used to set a bit in an



Structural features							
X	1	0	1	1	1	0	1
Y	1	0	1	1	0	0	0

Number of bits set in both (X) and (Y) = 3
Number of bits set in (X), but not in (Y) = 2
Number of bits set in (Y), but not in (X) = 0

Figure 1.8: Illustration of a simple fingerprinting approach. Two molecules (X) and (Y) were compared using a selected list of 7 structure fingerprints. The presence (or absence) of a feature is marked with digits $(1, 0)$ respectively. The summation of the bits in (X) and (Y), can be used to calculate a similarity measure.

bit-vector. MACCS (Molecular ACCess System) fingerprinting procedure assigns a unique number to each 'feature' of the structure [89]. Experiments analyzing the FP2 fingerprinting algorithm on a set of ceramide and phosphotidyl-inositol structures will be described in the results chapter.

LINGO is a type of molecular fingerprint that uses a fixed length sub string of SMILES [Equation 1.1]. The process of LINGO generation is described in Fig. 1.9. LINGO fingerprint was used for virtual screening of drug candidate molecules [90]. Structure similarity is measured from fingerprints with the use of a scoring function [Equation 1.2].

$$N_l = n - (q - 1) \tag{1.1}$$

n is the length of SMILES string, N_l is the number of LINGOs, each of length q.

$$s_l = \frac{\sum_{i=1}^{l} 1 - \frac{|N_{A,i} - N_{B,i}|}{N_{A,i} + N_{B,i}}}{l} \tag{1.2}$$

s_l is the similaity between a pair of SMILES strings A, B. $N_{A,i}$ is the number of LINGOs of type i in molecule A, $N_{B,i}$ is the number of LINGOs of type i in B, and l is the number of LINGOs contained in either molecule A or B [91].

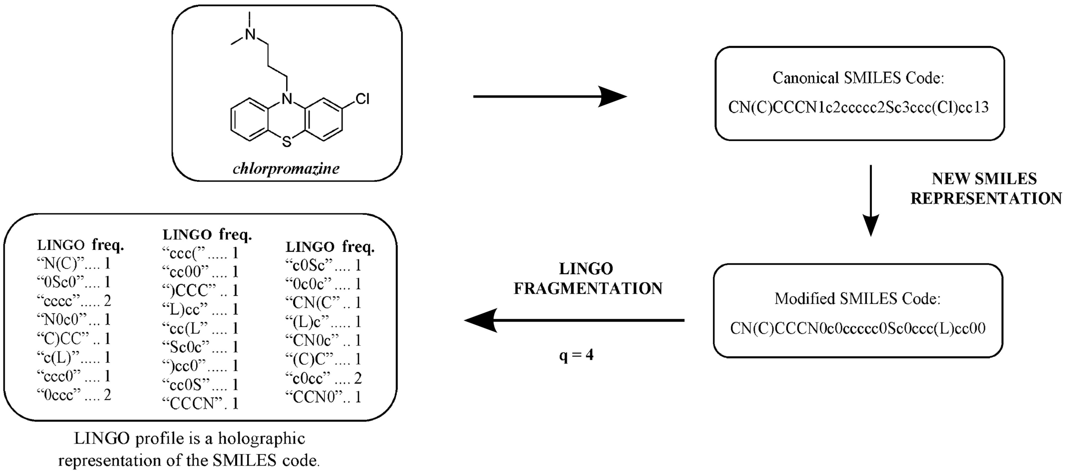


Figure 1.9: LINGO generation work flow. The process of generating LINGOs for chlorpromazine is summarized in 3 steps. First, the canonical SMILES of the molecule is generated, followed by transformation of digits $(1 - 9)$ to 0 and two letter atom (Cl) to a single letter (L). A moving frame of length 4 ($q = 4$) is used to fragment the transformed SMILES (length 31) to 25 LINGOs. The frequency of occurrence of each LINGO is compared with another molecule's LINGO frequency to calculate structure similarity. Modified from Vidal *et al.* [91]

1.3.2 Sequence Alignment

Sequence alignments were used to measure similarity between a pair of amino acid sequences [92, 93]. Smith and Waterman proposed a formal definition for the alignment procedure that can be used to calculate the sequence similarity [94]. Their function is optimized for finding regions with high similarity (called local alignment) but alternative approaches that were optimized for entire sequences (global alignment) are also available [95]. In biology, one is often interested in comparing all sequences of a gene or protein family, called Multiple Sequence Alignment (MSA) [96–98]. Edgar developed an MSA program that is faster and hence useful for larger sets [Fig. 1.10]. I posed the question, as to whether lipid SMILES could be regarded as sequences and compared using these methods.

Levenshtein described a method to detect errors in binary code, often referred as a fuzzy string matching approach [100]. Levenshtein's method is popular in the field of natural language processing to perform spell checks [101–104]. Given a pair of strings and costs for editing, a dynamic programming approach is used to determine the sequence of edits that minimizes the total cost of transforming one string to another [Fig. 1.11] [105, 106]. I investigated whether fuzzy string matching could be

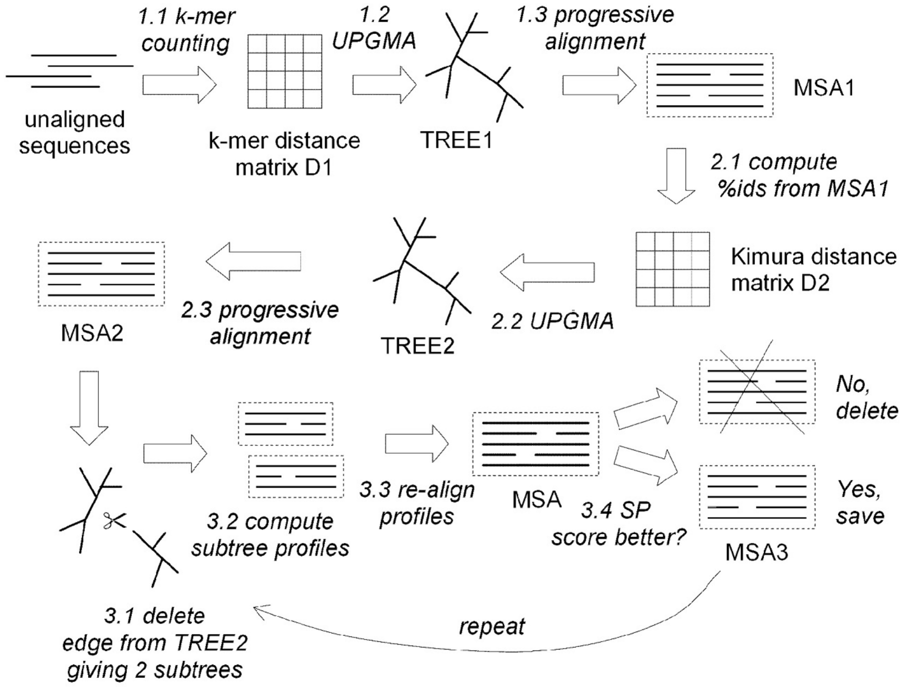

1.1 k-mer counting
1.2 UPGMA
1.3 progressive alignment
unaligned sequences
k-mer distance matrix D1
TREE1
MSA1
2.1 compute %ids from MSA1
Kimura distance matrix D2
2.2 UPGMA
TREE2
2.3 progressive alignment
MSA2
3.1 delete edge from TREE2 giving 2 subtrees
3.2 compute subtree profiles
3.3 re-align profiles
MSA
3.4 SP score better?
No, delete
Yes, save
MSA3
repeat

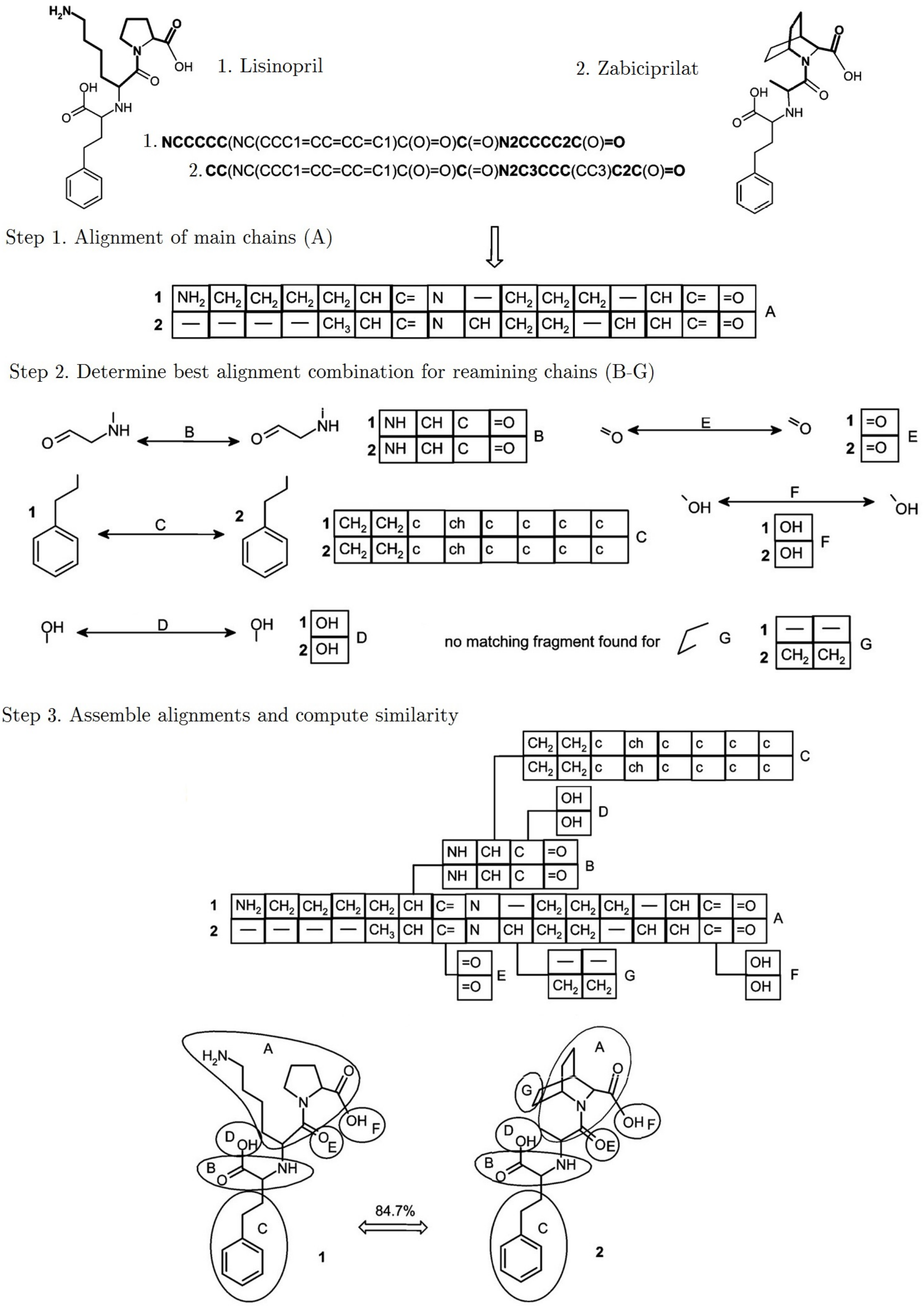
Figure 1.10: Summary of the steps in MUSCLE algorithm. The three main stages are 1. draft progressive alignment 2. improved progressive alignment and 3. refinement. UPGMA - Unweighted Pair Group Method with Arithmetic mean; *Kmer* - a contiguous subsequence of length k; SP - sum of pairwise alignment scores. Extracted from Edgar [99].

used for comparing lipid SMILES.

$$\begin{cases} d_{00} = 0 \\ d_{ij} = \min \begin{cases} d_{i-1,j-1} + \begin{cases} 0, & a_i = b_j \\ c_c, & a_i \neq b_j \end{cases} \\ d_{i-1,j} + c_d \\ d_{i,j-1} + c_i \end{cases} & \text{if } i > 0 \text{ or } j > 0 \end{cases}$$

Figure 1.11: Illustration of a dynamic programming approach. Let $A = a_1 \dots a_m$ and $B = b_1 \dots b_n$ are two strings, c_d, c_i and c_c are costs for deletion, insertion and change, then, $d_{ij} = d(a_1...a_i, b_1...b_j)$ for $0 \leq i \leq m,\ 0 \leq j \leq n$ can be calculated by recursion. Modified from Ukkonen [107].

Bioisosteric method uses SMILES representation and dynamic programming to measure structural similarity [Fig. 1.12]. It was originally developed for virtual screening of drug candidates but tested in this study to compare lipid SMILES.


1. Lisinopril
2. Zabiciprilat
1. NCCCCC(NC(CCC1=CC=CC=C1)C(O)=O)C(=O)N2CCCC2C(O)=O
2. CC(NC(CCC1=CC=CC=C1)C(O)=O)C(=O)N2C3CCC(CC3)C2C(O)=O
Step 1. Alignment of main chains (A)
Step 2. Determine best alignment combination for reamining chains (B-G)
no matching fragment found for
Step 3. Assemble alignments and compute similarity
A
B
C
D
E
F
G
1
2
84.7%

1	NH_2	CH_2	CH_2	CH_2	CH_2	CH	C=	N	—	CH_2	CH_2	CH_2	—	CH	C=	=O
2	—	—	—	—	CH_3	CH	C=	N	CH	CH_2	CH_2	—	CH	CH	C=	=O

1	NH	CH	C	=O
2	NH	CH	C	=O

1	CH_2	CH_2	c	ch	c	c	c	c
2	CH_2	CH_2	c	ch	c	c	c	c

1	OH
2	OH

1	=O
2	=O

1	OH
2	OH

1	—	—
2	CH_2	CH_2

Figure 1.12: Illustration of Bioisosteric similarity calculation procedure. The similarity is calculated for two molecules lisinopril and zabiciprilat. In the first step, the main chain (A) from the CACTVS canonical SMILES representation of the two molecules were aligned. The smaller chains (B-G) were aligned next iterating for best combination. In the last step, the aligned chains were assembled to compute overall similarity between two molecules. Modified from Krier *et al.* [108].

1.3.3 Metric Space for Lipid Structures

A metric space is a pair (X, ρ) of a set X and a metric ρ on X if ρ satisfies the four conditions [Equation 1.3] [109]. Chemical space is a theme in pharmaceutical research, the computational search for new drug compounds often starts from the region nearer to an existing drug molecule in the chemical space [110–112]. Inspired by the chemical space for drug compounds, I asked the question, weather such spaces could be created for lipid structures? Molecular descriptors were used to create chemical spaces of pharmacologically relevant compounds [113, 114]. The structural similarity scores obtained from fingerprints or sequence alignments could be used as a metric for lipids [115]. Metric spaces could be visualized by converting similarity matrices to coordinates with the use of dimensional scaling methods such as Principal Component Analysis, henceforth referred as PCA space [116, 117].

$$\begin{aligned}
\rho(x,y) &\geq 0 \\
\rho(x,y) &= 0 \text{ if and only if } x = y \\
\rho(x,y) &= \rho(y,x) \text{ for all } x, y \in X \\
\rho(x,y) + \rho(y,z) &\geq \rho(x,z) \text{ for all } x, y, z \in X
\end{aligned} \tag{1.3}$$

1.4 Comparative Lipidomics

The volume of experimental data has led to the need for methods to cluster lipids and compare lipidomes [118]. Often, researchers look at comparative lipidomics from the perspective of an increase (or decrease) in lipid abundances [23, 119–122]. Simple difference of the lipid concentration levels is a frequently used approach to compare lipid profiles [Fig. 1.13]. The aim of lipidomics experiments is to quantify as many lipids as possible from samples and then, use a correlation coefficient between lipid levels to compare them [25, 123–125]. Principal Component Analysis (PCA) and hierarchical clustering of lipid profiles were used to find associations between the yeast lipidomes [39, 49].

However, until recently, only the concentration change between individual lipid species (rather tha structures) were used for exploratory data analysis and clustering [118, 126] but the structure differences (example, sterol acyl chain) are an important determinant of phenotype [127, 128]. The experimental set-up to characterize lipidomes were often carried out between the strains of an individual species or between the tissues of a single organism [23, 121, 122]. In such cases, the major changes are noticeable with the lipid concentrations but for comparisons involving multiple species (such as model organisms and humans), often, the same lipid is not present in both profiles [Table 1.2], which means that many unique lipids are left out of abundance-based comparative analysis [Fig. 1.13]. One of the objectives of this study is to use structural similarity as the basis for comparing lipidomes, that makes use of unique lipids.

1.4.1 Metric Space for Lipidomes

The ability to cluster lipid structures in a PCA space opens the possibility to compare lipidomes in novel ways. Hausdorff distance is a measure of the overlap between two sets of data points, commonly used in image comparison [129]. Huttenlocher *et al.* considered six variations of directed Hausdorff distance measures [Equation 1.4] that could be applied for lipidome comparison [129, 130].

$$\begin{aligned}
d_{H_1}(AB) &= \min_{a \in A} \; d(a, B) \\
d_{H_2}(AB) &= {}^{50}K^{th}_{a \in A} \; d(a, B) \\
d_{H_3}(AB) &= {}^{75}K^{th}_{a \in A} \; d(a, B) \\
d_{H_4}(AB) &= {}^{90}K^{th}_{a \in A} \; d(a, B) \\
d_{H_5}(AB) &= \max_{a \in A} \; d(a, B) \\
d_{H_6}(AB) &= \frac{1}{N_a} \sum_{a \in A} \; d(a, B)
\end{aligned} \tag{1.4}$$

where ${}^{x}K^{th}_{a \in A}$ represents the K^{th} ranked distance. ${}^{50}K^{th}_{a \in A}$ corresponds to the median of the distances $d(a, B), \forall a \in A$. The $\min_{a \in A}$ and $\max_{a \in A}$ will capture only the outliers. Although, all six Hausdorff distance measures were tested in this study, only the results from the average of the shortest distances between the sets, $\frac{1}{N_a} \sum_{a \in A}$ are presented in the results chapter [section 3.3].

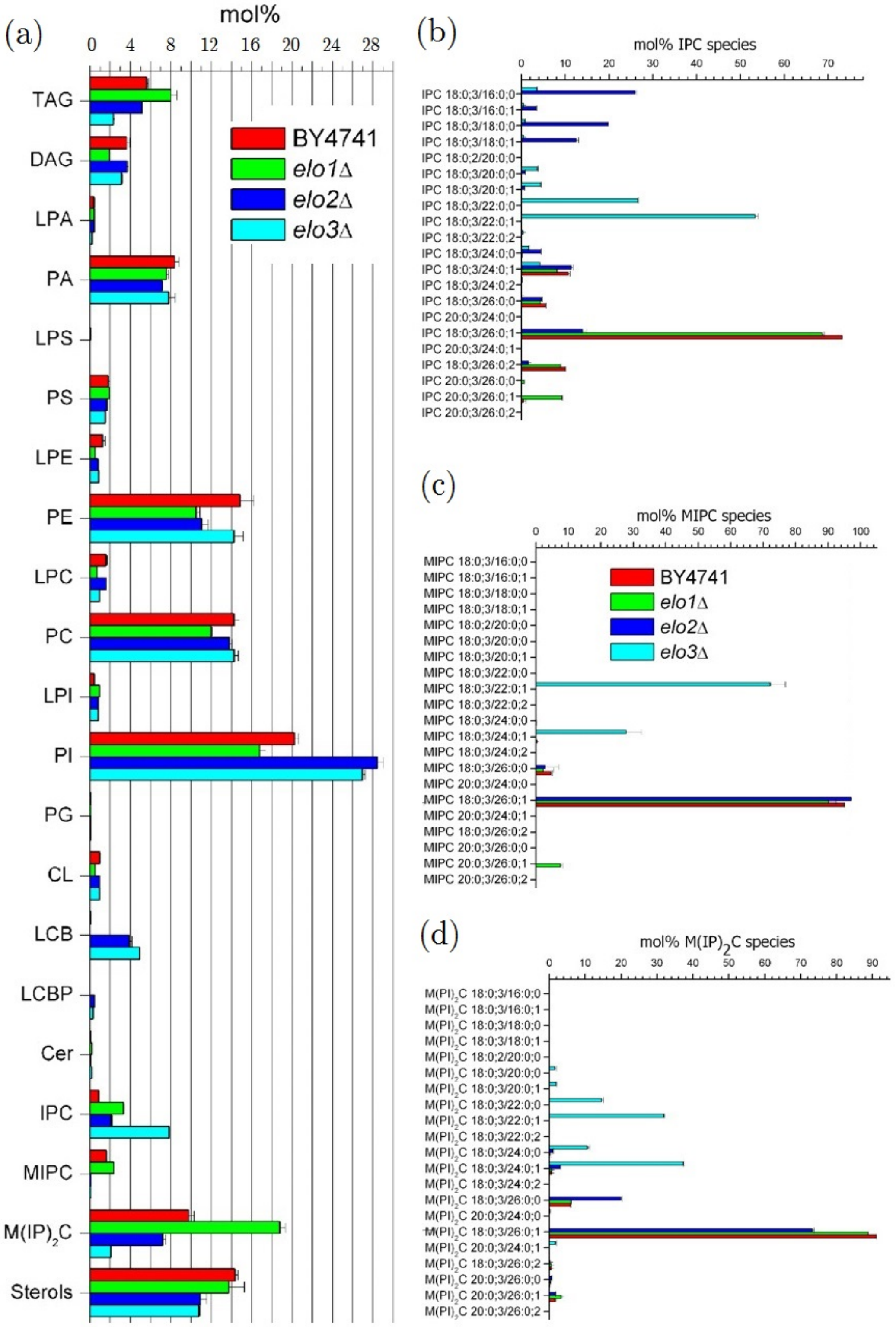


Figure 1.13: Illustration of comparatitive lipidomics based on lipid abundances. 4 yeast strains (BY4741 - control and Elo1, Elo2, Elo3 - mutation in Elongase gene) are compared based on the differnces in lipid classes (a) and lipid species (b-d). The average of all lipid species were used for plotting lipid class abundances (a). Molecular species of IPC class (b), MIPC class (c) and M(IP)$_2$C class (d). IPC - Inositol Phosphoryl Ceramides; MIPC - Mannose-inositolphospho-ceramide; M(IP)$_2$C 18:0;3/20:0;1 - Mannose-bis(inositolphospho)ceramide 18 carbon atoms, 0 double bonds, 3 hydroxyl groups in first acyl chain, 20 carbon atoms, 0 double bonds and 1 hydroxylation in second acyl chain. Modified from Ejsing *et al.* [23].

Hausdorff distance (d_H) between two sets (A, B) is directional [Equation 1.5]

$$d_H\ (AB) \neq d_H\ (BA) \tag{1.5}$$

Dubuisson *et al.* proposed ways to combine the directed Hausdorff distances to make it symmetric [130]. A modified symmetric hausdorff distance was used for comparing lipidomes [section 2.6]. Experiments were performed to validate the metric with yeast and fruit fly lipidomes [23, 121]. The tissue lipidomes of lung cancer patients were analyzed with structure based clustering method that was developed in this study [122].

Material and Methods

2.1 Lipid Structure Datasets

2.1.1 Ceramide and PI datasets

17 ceramide [Fig. 2.1a] and 16 phosphatidyl-inositol structures [Fig. 2.1b], varying in fatty acid chain length and number of double bonds were first drawn using PubChem Sketcher [131] and exported in SDF format [66]. SDF files were converted to template and canonical SMILES as described in section 2.3. The structural similarity between the ceramides and the phosphatidyl-inositols was calculated with six scoring methods [section 2.4].

2.1.2 LIPID MAPS Structure Database

The complete LIPIDMAPS Structure Database (LMSD) comprising 30 150 lipid structures in SDF format was downloaded from their website [132].

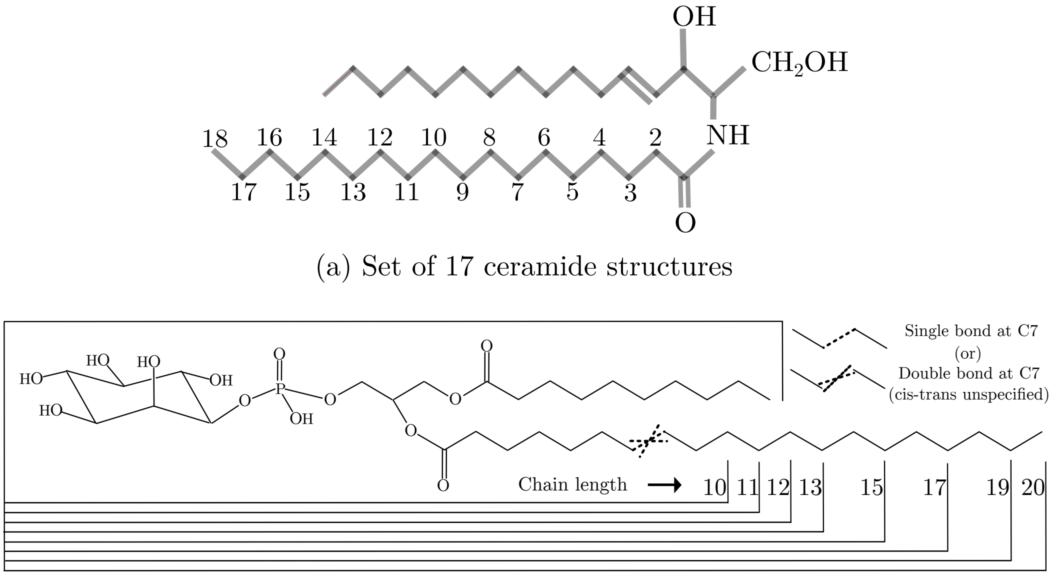
SDF files were converted to template SMILES with Open Babel [section 2.3]. Levenshtein distance was calculated for all pairs of SMILES strings [section 2.4], followed by Principal Component Analysis [section 2.5]. The LIPID MAPS classification [Table 1.1] by Fahy *et al.* [8] was applied for analyzing the PCA space.

2.2 Lipidome Datasets

Yeast, fruit fly and human lung lipidomes were used in this study [23, 121, 122]. The lipid names in these datasets were written in a simple form which is explained below.

2.2.1 Short hand notation of lipid names

Lipid categories were abbreviated as Glycerophospholipids (GP), Diacylglycerols (DAG), Triacylglycerols (TAG), Sphingolipids (SP) and Cholesterol Esters (CE). A compiled list of abbreviations for other lipids were given in Table 1. Lipid species abbreviation is described for three main classes in next page.



(a) Set of 17 ceramide structures

(b) Set of 16 PI structures

Figure 2.1: Sets of ceramide and PI molecules are graphically represented. (a) A set of ceramide molecules with a C-16 sphingoid base, an amide linked acyl chain and a hydroxyl group that is attached to different carbon atoms in the acyl chain. IUPAC numbering of carbon atoms is displayed for acyl chain. A hydroxyl group is sequentially moved from position 2 to 18 in the acyl chain, generating 17 different structures. The hydroxyl group position is simultaneously used for identifying the molecule in the later chapters. For example, the ceramide structure with hydroxyl group at 14 position in acyl chain will be just referred as 14. (b) A set of phospatidylinositol molecules with an acyl chain that has variable length, from 10 to 20. C7-C8 connection in the acyl chain is either a single bond or a double bond. By varying the acyl chain length and saturation level, 16 distinct PI structures were generated. The chain length is used for naming the molecules and * is used to denote the double bond. For example, molecule 17* has 17 carbon atoms in acyl chain and it is unsaturated at C7.

2.2.1.1 GP, DAG and TAG

`<lipid species> <space> <no. of carbons in all fatty acids> : <no. of double bonds in all acyl chains combined>` Example - DAG 40:1. If the sn1 and sn2 position [Fig. 2.2] for the acyl chains is known, they were annotated as

`<lipid class> <space> <no. of carbons of sn1 fatty acid> : <no. of double bonds> / <no. of carbons of sn2 fatty acid> : <no. of double bonds>` Example - DAG 40:1 with 22 carbon atom sn1 acyl chain and an 18 length sn2 acyl chain is written as DAG 22:0/18:1

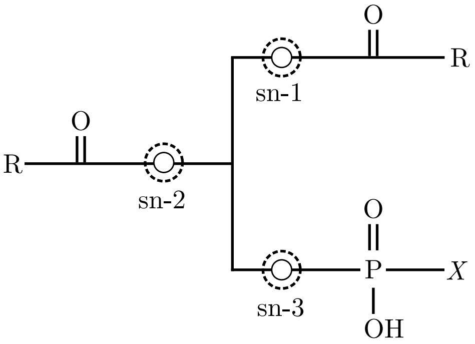


Figure 2.2: Sn1 and Sn2 labeling of alkyl chains [133, 134]

2.2.1.2 SP

`<lipid species> <space> <no. of carbons in the long-chain base and fatty acid moieties> : <no. of double bonds in the long-chain base and fatty acid moieties> ; <no. of hydroxyl groups in the long-chain base and fatty acid moieties>` Example - Cer 32:1;2 but when the head group composition is known (say 18:0;2), then the same molecule is written as Cer 18:0;2/16:0;0

2.2.1.3 CE

`<lipid species> <space> <no. of carbons additional to cholesterol> : <no. of double bonds>: <no. of hydroxyl groups additional to the hydroxyl group at position 3>` Example - CE 24:4;0

2.2.2 Yeast Elongase Mutants

Eight lipidomes, comprising three elongase mutants (Elo1, Elo2, Elo3), and a control strain (BY4741), cultured at two temperatures regimes each (24 and 37 °C), were obtained from a previous study [23]. The number of lipid species identified in each lipidome varied - only 145 lipids were measured in BY4741 cultured at 37 °C, but 176 in BY4741 24 °C. The number of lipids that overlap between the eight lipidomes were summarized in Fig. 2.3. The lipid species in all eight lipidomes were combined, duplicates removed to generate a master list (contains 248 lipids) that is subsequently used in a. pairwise structure similarity calculation (with Levenshtein distance), b. PCA space representation of yeast lipidome(s) and c. LUX Score calculation [section 2.6].

LIPID MAPS structure drawing tools [135] were customized for programmatic generation of structures for all lipid classes, except sterols [136]. The output of structure drawing program in SDF format was converted to template SMILES with Open Babel library [section 2.3]. Ergosterol and ergosta-5,7-dien-3β-ol structures in SDF format were obtained from LMSD separately, converted to template SMILES, and added to the structure list. SMILES for phytosphingosine 1-phosphate was generated manually by editing the SMILES string for phytosphingosine.

Acyl chains with the possible position of double bonds and hydroxylations in yeast were compiled from previous studies [Table 2.1] [137, 138]. This list is used for drawing structures. The sn1, sn2 and sn3 specific acyl chain composition could not be conclusively determined for many lipid species (example TAG 14:1/16:1/22:0 can be TAG 16:1/14:1/22:0 or TAG 14:1/22:0/16:0). In such cases, a list of isomers was generated and a representative structure selected (the isomer with least average Levenshtein distance [section 2.4.6] was chosen as the representative).

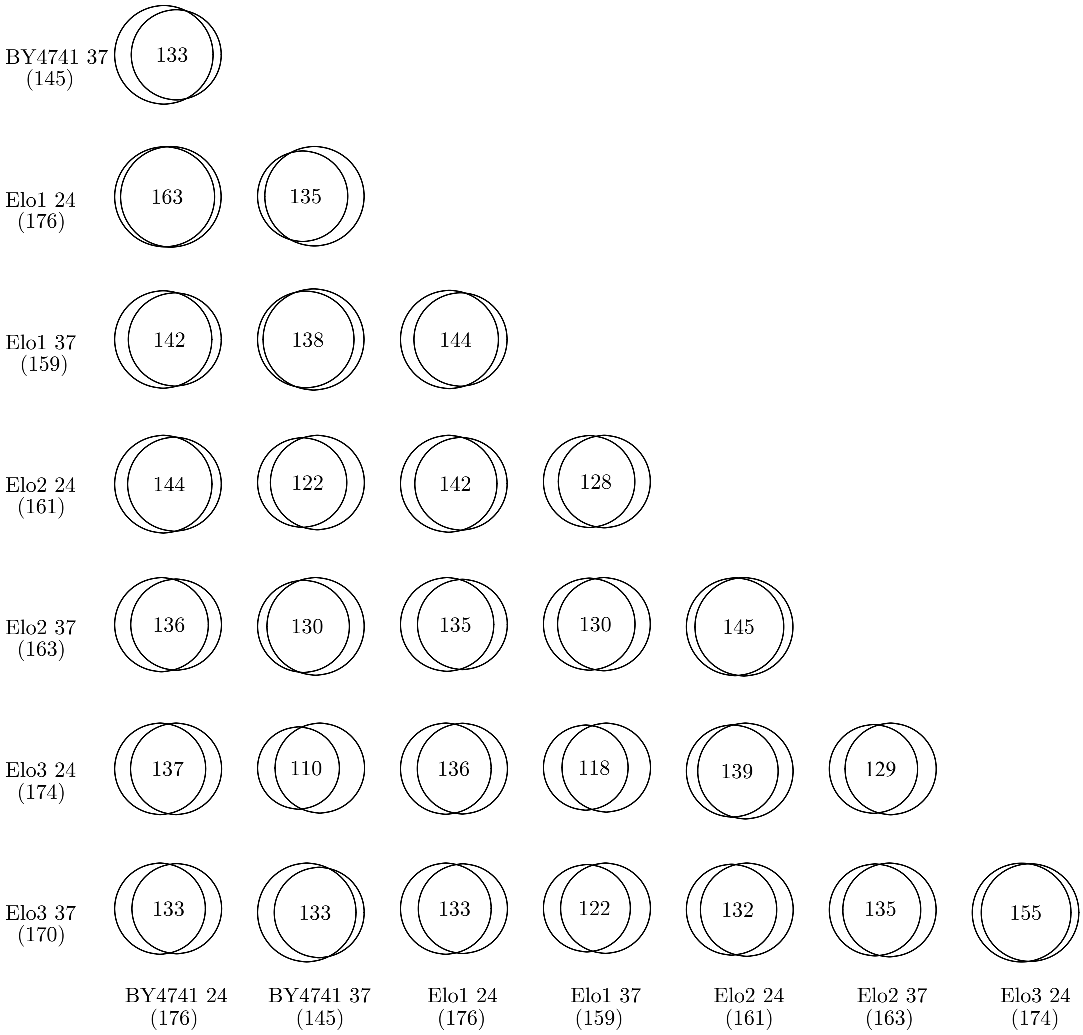


Figure 2.3: Number of lipids that overlap between yeast elongase mutant lipidomes. By4741 is control strain. Elo1, Elo2, and Elo3 are mutants. 24 and 37 in names refer to growth temperature in Celsius. The number of lipids in each lipidome is shown below the name. The area of the circle is proportional to the number of lipids. The pair, BY4741 24 and Elo3 24 has highest number of overlapping lipids (163), BY4741 37 and Elo3 24 pair has the least number of overlapping lipids (110).

Table 2.1: Compiled list of fatty acids in yeast

No. of Carbon atoms	No. of Double bonds	No. of Carbon atoms	No. of Double bonds	Unsaturation position
10	0	24	0	
12	0	26	0	
14	0	12	1	(9Z)
16	0	14	1	(9Z)
18	0	16	1	(9Z)
20	0	16	2	(9Z,12Z)
22	0	18	1	(9Z)
		18	2	(9Z,12Z)

2.2.3 Fruit fly Larva, tissue-specific Lipidomes

356 lipid species from 12 lipidomes of *Drosophina melanoaster* larvae were obtained from a recent study [121]. In that study, the larvae were fed with two diet regimes a. Plant based food (PF) and b. Yeast based food (YF) and 6 tissues were dissected (gut, brain, wing disc, salivary glands, fat body and lipoprotein). The 12 lipidomes are summarized in Table 2.2. The lipid species from 12 lipidomes were combined (and duplicates removed) to create the list of 356 lipid species, however, structures for 10 species could not be drawn (described later).

Fatty acids reported in fruit fly [56] and those that might have been incorporated from the food source or from the larval gut microbiome [139] were compiled to a generate a list of 29 possible acyl chains [Table 2.3]. This list was used as input to LIPID MAPS structure drawing tools for programmatic generation of lipid structures. All lipid structures were drawn programmatically except sphingolipids and sterols. These two classes could not be correctly drawn using LIPID MAPS tools, they were manually curated to ensure correct structure selection.

Eight lipids (DAG 28:4, PC 38:7, PE 40:7, PE 40:8, PE 40:9, PI 38:7, PS 38:7, TAG 55:8) were omitted because they contain an unusually high number of double bonds. One sphingolipid (Cer 39:1;2) and one sterol (ST 14:0) could not be drawn. Cer 39:1:2 could not be drawn because the combination of fatty acid and sphingosine structure could be assigned. SMILES for all sterols were derived from cholesterol as basic structure. But one sterol (ST 1:4:0) was omitted because the structure could not be generated programmatically. In summary, 10 structures (out of 356) could not be drawn. 346 lipids were later used for structure similarity calculation (described later) and for LUX analysis [section 2.6].

There were situations where it was not straight forward to draw structures, broadly for two reasons 1. for some lipid species, the number of hydroxylations and double bonds was known, but their position was not (example, TAG 48:4 and Cer 32:2;2). 2. The sn1, sn2 and sn3 specific acyl chain composition was not available for many lipid species. In these cases, all the isomer possibilities were computationally generated and the isomer with lowest average Levenshtein distance to other isomers

Table 2.2: Overview of fruit fly larval tissue lipidomes [121]

	Tissue	Food	No. of Lipids
1	Gut	Yeast	267
2	Gut	Plant	261
3	Brain	Yeast	198
4	Brain	Plant	208
5	Wing disc	Yeast	209
6	Wing disc	Plant	204
7	Salivary glands	Yeast	205
8	Salivary glands	Plant	196
9	Fat body	Yeast	182
10	Fat body	Plant	162
11	Lipoprotein	Yeast	164
12	Lipoprotein	Plant	165

was selected as a representative structure [section 2.4.6].

Sphingolipids of fruit fly are special and difficult to process with LIPID MAPS tools for two reasons 1. They have a conserved ceramide structure that contains a long chain bases of length 14 (or 16) carbon atoms [49, 140], which was not possible with LIPID MAPS structure drawing tools. To solve this problem, I made a changes to the LIPID MAPS scripts. 2. In Drosophila δ(4,6)-sphingadienes are found, which could not be drawn with the LIPIDMAPS tools. I modified LIPID MAPS structure drawing scripts to place an additional hydroxyl group at the alpha position of the fatty acids.

2.2.4 Human Lung, Cancer versus Non-cancer Tissue

311 lipid species from 43 human lung tissue biopsies were obtained from a recent study [122]. 21 tissues were from cancerous region of the lung and remaining were from the alveolar tissue (tumor-free) [Table 2.5]. 35 fatty acids possibilities were considered to generate lipid structures [Table 2.6]. Modified LIPID MAPS structure drawing tools [135] were used for programmatic generation of lipid structures, similar to the procedure described for yeast and fruit fly. Cholesterol structure was separately obtained from LMSD and added to the SMILES list. For ceramides and sphingolipids, long chain base with 18 carbon atoms (C18) is used to draw structure 18 lipids could not be drawn, hence, excluded from the LUX analysis [Table 2.8].

Table 2.3: Compiled list of fatty acids in fruit fly

Carbon atoms	Double bonds	Unsaturation position	Carbon atoms	Double bonds	Unsaturation position
10	0		18	0	
12	0		18	1	(9Z)
12	1	(9Z)	18	2	(9Z,12Z)
13*	0		18	3	(9Z,12Z,15Z)
13*	1	(9Z)	19*	0	
14	0		19*	1	(9Z)
14	1	(9Z)	20	1	(9Z)
15*	0		20	2	(9Z,12Z)
15*	1	(9Z)	20	3	(9Z,12Z,15Z)
16	0		22	1	(9Z)
16	1	(9Z)	22	2	(9Z,12Z)
16	2	(9Z,12Z)	22	3	(9Z,12Z,15Z)
17*	0		24	1	(9Z)
17*	1	(9Z)	24	2	(9Z,12Z)
			24	3	(9Z,12Z,15Z)

* Fatty acids from food or microbes.

Table 2.5: Overview of Lung Lipidome

Sample	Gender	Age	Cancer Type	Tissue	No. of Lipids*
ID11	Male	49	Adeno	Alveolar	180
ID12	Male	70	Squamous	Alveolar	195
ID12	Male	70	Squamous	Tumor	254
ID15	Male	45	Squamous	Alveolar	169
ID17	Male	69	Squamous	Alveolar	186
ID18	Female	48	Squamous	Alveolar	187
ID18	Female	48	Squamous	Tumor	267
ID19	Female	54		Alveolar	195
ID19	Female	54		Tumor	247
ID2	Male	71	Squamous	Alveolar	193
ID2	Male	71	Squamous	Tumor	236
ID22	Male	57	Squamous	Alveolar	226
ID22	Male	57	Squamous	Tumor	222
ID24	Male	55	Adeno	Alveolar	199
ID24	Male	55	Adeno	Tumor	248
ID29	Male	59	Squamous	Alveolar	223
ID29	Male	59	Squamous	Tumor	267
ID30	Male	53	Adeno	Tumor	209
ID31	Male	68	Squamous	Tumor	212
ID32	Male	46	Adeno	Tumor	191
ID39	Male	60	Squamous	Alveolar	213
ID39	Male	60	Squamous	Tumor	246
ID4	Male	67	Squamous	Alveolar	197
ID4	Male	67	Squamous	Tumor	234
ID43	Male	60		Alveolar	217
ID43	Male	60		Tumor	253
ID50	Male	63	Squamous	Alveolar	201
ID50	Male	63	Squamous	Tumor	264
ID52	Female	44		Alveolar	186
ID53	Male	48	Adeno	Alveolar	166
ID53	Male	48	Adeno	Tumor	233
ID6	Male	71	Squamous	Alveolar	176
ID6	Male	71	Squamous	Tumor	169
ID61	Male	70	Adeno	Alveolar	231
ID61	Male	70	Adeno	Tumor	230
ID64	Female	52	Adeno	Alveolar	196
ID64	Female	52	Adeno	Tumor	253
ID66	Male	63	Adeno	Alveolar	183
ID66	Male	63	Adeno	Tumor	262
ID67	Female	52	Adeno	Alveolar	184
ID67	Female	52	Adeno	Tumor	240
ID71	Female	57	Adeno	Tumor	242
ID72	Male	62	Adeno	Tumor	253

*No. of lipids used as input to the LUX score calculation program

Table 2.6: Putative fatty acids composition for human lung lipdiome

C	db	db position	C	db	db position
14	0		20	3	(9Z,12Z,15Z)
14	1	(9Z)	20	4	(9Z,12Z,15Z,18Z)
15	0		20	5	(6Z,9Z,12Z,15Z,18Z)
16	0		21	1	(9Z)
16	1	(9Z)	22	0	
16	2	(9Z,12Z)	22	1	(9Z)
17	0		22	2	(9Z,15Z)
17	1	(9Z)	22	3	(9Z,12Z,15Z)
17	2	(9Z,12Z)	22	4	(9Z,12Z,15Z,18Z)
18	0		22	5	(6Z,9Z,12Z,15Z,18Z)
18	1	(9Z)	22	6	(6Z,9Z,12Z,15Z,18Z,21Z)
18	2	(9Z,12Z)	24	0	
18	3	(9Z,12Z,15Z)	24	1	(9Z)
19	1	(9Z)	24	2	(9Z,12Z)
20	0		24	3	(9Z,12Z,15Z)
20	1	(9Z)	24	4	(9Z,12Z,15Z,18Z)
20	2	(9Z,12Z)			

C is the number of carbon atoms, db is the number of double bonds

Table 2.8: List of lung lipids excluded from LUX analysis

CE 19:2	SM 41:2;2	TAG 37:3	TAG 45:7
Cer 41:2;2	PE_O 38:9	TAG 37:4	TAG 46:7
DAG 41:9	PS 39:9	TAG 39:5	TAG 46:8
TAG 41:7	SM 41:1;2	TAG 41:6	TAG 48:9
	HexCer 40:3;3	HexCer 37:3;2	

CE - Cholesterol Ester; SM - Sphingomyelin; TAG - Triacylglycerol; HexCer - Hexosyl ceramide; DAG - Diacylglycerol.

2.3 SMILES Conversion

Molecules were drawn either with PubChem Sketcher [131] or LIPID MAPS Structure Drawing tools [141]. The structures were exported as chemical table files in SDF format [66]. The three SMILES representations were derived from SDF files with the following tools and options.

1. Template SMILES

LIPID MAPS Structure Drawing Tools [141] were used to draw structures. Open Babel molecule conversion script is used for generating SMILES [84].

2. Open Babel canonical SMILES

Similar to the template SMILES protocol but with canonical option in Open Babel program.

3. CACTVS canonical SMILES

The molecules were hand drawn in PubChem Sketcher and exported in MDL MOL format. CACTVS SMILES translator web interface was used to bring them to a canonical form. [83].

Template and canonical SMILES for 17 ceramides and 16 PI are provided in Tables. 2.9 to 2.14. LIPID MAPS scripts were modified to generate a wider spectrum of structures for yeast, fruit fly and human lung lipids [135,141]. Characters indicating chirality, cis–trans isomerism and charges were removed.

Table 2.9: Template SMILES for 17 ceramides

2	CCCCCCCCCCCCCCCCC(O)C(=O)NC(CO)C(O)C=CCCCCCCCCCCC
3	CCCCCCCCCCCCCCCC(O)CC(=O)NC(CO)C(O)C=CCCCCCCCCCCC
4	CCCCCCCCCCCCCCC(O)CCC(=O)NC(CO)C(O)C=CCCCCCCCCCCC
5	CCCCCCCCCCCCCC(O)CCCC(=O)NC(CO)C(O)C=CCCCCCCCCCCC
6	CCCCCCCCCCCCC(O)CCCCC(=O)NC(CO)C(O)C=CCCCCCCCCCCC
7	CCCCCCCCCCCC(O)CCCCCC(=O)NC(CO)C(O)C=CCCCCCCCCCCC
8	CCCCCCCCCCC(O)CCCCCCC(=O)NC(CO)C(O)C=CCCCCCCCCCCC
9	CCCCCCCCCC(O)CCCCCCCC(=O)NC(CO)C(O)C=CCCCCCCCCCCC
10	CCCCCCCCC(O)CCCCCCCCC(=O)NC(CO)C(O)C=CCCCCCCCCCCC
11	CCCCCCCC(O)CCCCCCCCCC(=O)NC(CO)C(O)C=CCCCCCCCCCCC
12	CCCCCCC(O)CCCCCCCCCCC(=O)NC(CO)C(O)C=CCCCCCCCCCCC
13	CCCCCC(O)CCCCCCCCCCCC(=O)NC(CO)C(O)C=CCCCCCCCCCCC
14	CCCCC(O)CCCCCCCCCCCCC(=O)NC(CO)C(O)C=CCCCCCCCCCCC
15	CCCC(O)CCCCCCCCCCCCCC(=O)NC(CO)C(O)C=CCCCCCCCCCCC
16	CCC(O)CCCCCCCCCCCCCCC(=O)NC(CO)C(O)C=CCCCCCCCCCCC
17	CC(O)CCCCCCCCCCCCCCCC(=O)NC(CO)C(O)C=CCCCCCCCCCCC
18	C(O)CCCCCCCCCCCCCCCCC(=O)NC(CO)C(O)C=CCCCCCCCCCCC

Table 2.10: CACTVS Canonical SMILES for 17 ceramides

2	CCCCCCCCCCCCCCCCC(O)C(=O)NC(CO)C(O)C=CCCCCCCCCCCC
3	CCCCCCCCCCCCCCCC(O)CC(=O)NC(CO)C(O)C=CCCCCCCCCCCC
4	CCCCCCCCCCCCCCC(O)CCC(=O)NC(CO)C(O)C=CCCCCCCCCCCC
5	CCCCCCCCCCCCCC(O)CCCC(=O)NC(CO)C(O)C=CCCCCCCCCCCC
6	CCCCCCCCCCCCC(O)CCCCC(=O)NC(CO)C(O)C=CCCCCCCCCCCC
7	CCCCCCCCCCCC=CC(O)C(CO)NC(=O)CCCCCC(O)CCCCCCCCCCC
8	CCCCCCCCCCCC=CC(O)C(CO)NC(=O)CCCCCCC(O)CCCCCCCCCC
9	CCCCCCCCCCCC=CC(O)C(CO)NC(=O)CCCCCCCC(O)CCCCCCCCC
10	CCCCCCCCCCCC=CC(O)C(CO)NC(=O)CCCCCCCCC(O)CCCCCCCC
11	CCCCCCCCCCCC=CC(O)C(CO)NC(=O)CCCCCCCCCC(O)CCCCCCC
12	CCCCCCCCCCCC=CC(O)C(CO)NC(=O)CCCCCCCCCCC(O)CCCCCC
13	CCCCCCCCCCCC=CC(O)C(CO)NC(=O)CCCCCCCCCCCC(O)CCCCC
14	CCCCCCCCCCCC=CC(O)C(CO)NC(=O)CCCCCCCCCCCCC(O)CCCC
15	CCCCCCCCCCCC=CC(O)C(CO)NC(=O)CCCCCCCCCCCCCC(O)CCC
16	CCCCCCCCCCCC=CC(O)C(CO)NC(=O)CCCCCCCCCCCCCCC(O)CC
17	CCCCCCCCCCCC=CC(O)C(CO)NC(=O)CCCCCCCCCCCCCCCC(O)O
18	CCCCCCCCCCCC=CC(O)C(CO)NC(=O)CCCCCCCCCCCCCCCCC(O)

Table 2.11: Open Babel Canonical SMILES for 17 ceramides

2	CCCCCCCCCCCCCCCCC(C(=O)NC(C(C=CCCCCCCCCCCC)O)CO)O
3	CCCCCCCCCCCCCCCC(CC(=O)NC(C(C=CCCCCCCCCCCC)O)CO)O
4	CCCCCCCCCCCCCCC(CCC(=O)NC(C(C=CCCCCCCCCCCC)O)CO)O
5	CCCCCCCCCCCCCC(CCCC(=O)NC(C(C=CCCCCCCCCCCC)O)CO)O
6	CCCCCCCCCCCC=CC(C(NC(=O)CCCCC(CCCCCCCCCCCC)O)CO)O
7	CCCCCCCCCCCC=CC(C(NC(=O)CCCCCC(CCCCCCCCCCC)O)CO)O
8	CCCCCCCCCCCC=CC(C(NC(=O)CCCCCCC(CCCCCCCCCC)O)CO)O
9	CCCCCCCCCCCC=CC(C(NC(=O)CCCCCCCC(CCCCCCCCC)O)CO)O
10	CCCCCCCCCCCC=CC(C(NC(=O)CCCCCCCCC(CCCCCCCC)O)CO)O
11	CCCCCCCCCCCC=CC(C(NC(=O)CCCCCCCCCC(CCCCCCC)O)CO)O
12	CCCCCCCCCCCC=CC(C(NC(=O)CCCCCCCCCCC(CCCCCC)O)CO)O
13	CCCCCCCCCCCC=CC(C(NC(=O)CCCCCCCCCCCC(CCCCC)O)CO)O
14	CCCCCCCCCCCC=CC(C(NC(=O)CCCCCCCCCCCCC(CCCC)O)CO)O
15	CCCCCCCCCCCC=CC(C(NC(=O)CCCCCCCCCCCCCC(CCC)O)CO)O
16	CCCCCCCCCCCC=CC(C(NC(=O)CCCCCCCCCCCCCCC(CC)O)CO)O
17	CCCCCCCCCCCC=CC(C(NC(=O)CCCCCCCCCCCCCCCC(O)C)CO)O
18	CCCCCCCCCCCC=CC(C(NC(=O)CCCCCCCCCCCCCCCCCO)CO)O

Table 2.12: CACTVS Canonical SMILES for 16 PI

10	CCCCCCCCCC(=O)OCC(CO[P](O)(=O)OC1C(O)C(O)C(O)C(O)C1O)OC(=O)CCCCCCCCC
10*	CCCCCCCCCC(=O)OCC(CO[P](O)(=O)OC1C(O)C(O)C(O)C(O)C1O)OC(=O)CCCCCC=CCC
11	CCCCCCCCCCC(=O)OC(COC(=O)CCCCCCCCC)CO[P](O)(=O)OC1C(O)C(O)C(O)C(O)C1O
11*	CCCCCCCCCC(=O)OCC(CO[P](O)(=O)OC1C(O)C(O)C(O)C(O)C1O)OC(=O)CCCCCC=CCCC
12	CCCCCCCCCCCC(=O)OC(COC(=O)CCCCCCCCC)CO[P](O)(=O)OC1C(O)C(O)C(O)C(O)C1O
12*	CCCCCCCCCC(=O)OCC(CO[P](O)(=O)OC1C(O)C(O)C(O)C(O)C1O)OC(=O)CCCCCC=CCCCC
13	CCCCCCCCCCCCC(=O)OC(COC(=O)CCCCCCCCC)CO[P](O)(=O)OC1C(O)C(O)C(O)C(O)C1O
13*	CCCCCCCCCC(=O)OCC(CO[P](O)(=O)OC1C(O)C(O)C(O)C(O)C1O)OC(=O)CCCCCC=CCCCCC
15	CCCCCCCCCCCCCCC(=O)OC(COC(=O)CCCCCCCCC)CO[P](O)(=O)OC1C(O)C(O)C(O)C(O)C1O
15*	CCCCCCCCCC(=O)OCC(CO[P](O)(=O)OC1C(O)C(O)C(O)C(O)C1O)OC(=O)CCCCCC=CCCCCCCC
17	CCCCCCCCCCCCCCCCC(=O)OC(COC(=O)CCCCCCCCC)CO[P](O)(=O)OC1C(O)C(O)C(O)C(O)C1O
17*	CCCCCCCCCC=CCCCCCC(=O)OC(COC(=O)CCCCCCCCC)CO[P](O)(=O)OC1C(O)C(O)C(O)C(O)C1O
19	CCCCCCCCCCCCCCCCCCC(=O)OC(COC(=O)CCCCCCCCC)CO[P](O)(=O)OC1C(O)C(O)C(O)C(O)C1O
19*	CCCCCCCCCCCC=CCCCCCC(=O)OC(COC(=O)CCCCCCCCC)CO[P](O)(=O)OC1C(O)C(O)C(O)C(O)C1O
20	CCCCCCCCCCCCCCCCCCCC(=O)OC(COC(=O)CCCCCCCCC)CO[P](O)(=O)OC1C(O)C(O)C(O)C(O)C1O
20*	CCCCCCCCCCCCC=CCCCCCC(=O)OC(COC(=O)CCCCCCCCC)CO[P](O)(=O)OC1C(O)C(O)C(O)C(O)C1O

Table 2.13: Open Babel Canonical SMILES for 16 PI

10	CCCCCCCCCC(=O)OCC(OC(=O)CCCCCCCCC)COP(=O)(OC1C(O)C(O)C(C(C1O)O)O)O
10*	CCCCCCCCCC(=O)OCC(OC(=O)CCCCCC=CCC)COP(=O)(OC1C(O)C(O)C(C(C1O)O)O)O
11	CCCCCCCCCCC(=O)OC(COP(=O)(OC1C(O)C(O)C(C(C1O)O)O)O)COC(=O)CCCCCCCCC
11*	CCCCCCCCCC(=O)OCC(OC(=O)CCCCCC=CCCC)COP(=O)(OC1C(O)C(O)C(C(C1O)O)O)O
12	CCCCCCCCCCCC(=O)OC(COP(=O)(OC1C(O)C(O)C(C(C1O)O)O)O)COC(=O)CCCCCCCCC
12*	CCCCCCCCCC(=O)OCC(OC(=O)CCCCCC=CCCCC)COP(=O)(OC1C(O)C(O)C(C(C1O)O)O)O
13	CCCCCCCCCCCCC(=O)OC(COP(=O)(OC1C(O)C(O)C(C(C1O)O)O)O)COC(=O)CCCCCCCCC
13*	CCCCCCCCCC(=O)OCC(OC(=O)CCCCCC=CCCCCC)COP(=O)(OC1C(O)C(O)C(C(C1O)O)O)O
15	CCCCCCCCCCCCCCC(=O)OC(COP(=O)(OC1C(O)C(O)C(C(C1O)O)O)O)COC(=O)CCCCCCCCC
15*	CCCCCCCCCC(=O)OCC(OC(=O)CCCCCC=CCCCCCCC)COP(=O)(OC1C(O)C(O)C(C(C1O)O)O)O
17	CCCCCCCCCCCCCCCCC(=O)OC(COP(=O)(OC1C(O)C(O)C(C(C1O)O)O)O)COC(=O)CCCCCCCCC
17*	CCCCCCCCCC=CCCCCCC(=O)OC(COP(=O)(OC1C(O)C(O)C(C(C1O)O)O)O)COC(=O)CCCCCCCCC
19	CCCCCCCCCCCCCCCCCCC(=O)OC(COP(=O)(OC1C(O)C(O)C(C(C1O)O)O)O)COC(=O)CCCCCCCCC
19*	CCCCCCCCCCCC=CCCCCCC(=O)OC(COP(=O)(OC1C(O)C(O)C(C(C1O)O)O)O)COC(=O)CCCCCCCCC
20	CCCCCCCCCCCCCCCCCCCC(=O)OC(COP(=O)(OC1C(O)C(O)C(C(C1O)O)O)O)COC(=O)CCCCCCCCC
20*	CCCCCCCCCCCCC=CCCCCCC(=O)OC(COP(=O)(OC1C(O)C(O)C(C(C1O)O)O)O)COC(=O)CCCCCCCCC

Table 2.14: Template SMILES for 16 PI

10	C(COC(=O)CCCCCCCCC)(OC(=O)CCCCCCCCC)COP(=O)(OC1C(C(C(C(C1O)O)O)O)O)O
10*	C(COC(=O)CCCCCCCCC)(OC(=O)CCCCCC=CCC)COP(=O)(OC1C(C(C(C(C1O)O)O)O)O)O
11	C(COC(=O)CCCCCCCCC)(OC(=O)CCCCCCCCCC)COP(=O)(OC1C(C(C(C(C1O)O)O)O)O)O
11*	C(COC(=O)CCCCCCCCC)(OC(=O)CCCCCC=CCCC)COP(=O)(OC1C(C(C(C(C1O)O)O)O)O)O
12	C(COC(=O)CCCCCCCCC)(OC(=O)CCCCCCCCCCC)COP(=O)(OC1C(C(C(C(C1O)O)O)O)O)O
12*	C(COC(=O)CCCCCCCCC)(OC(=O)CCCCCC=CCCCC)COP(=O)(OC1C(C(C(C(C1O)O)O)O)O)O
13	C(COC(=O)CCCCCCCCC)(OC(=O)CCCCCCCCCCCC)COP(=O)(OC1C(C(C(C(C1O)O)O)O)O)O
13*	C(COC(=O)CCCCCCCCC)(OC(=O)CCCCCC=CCCCCC)COP(=O)(OC1C(C(C(C(C1O)O)O)O)O)O
15	C(COC(=O)CCCCCCCCC)(OC(=O)CCCCCCCCCCCCCC)COP(=O)(OC1C(C(C(C(C1O)O)O)O)O)O
15*	C(COC(=O)CCCCCCCCC)(OC(=O)CCCCCC=CCCCCCCC)COP(=O)(OC1C(C(C(C(C1O)O)O)O)O)O
17	C(COC(=O)CCCCCCCCC)(OC(=O)CCCCCCCCCCCCCCCC)COP(=O)(OC1C(C(C(C(C1O)O)O)O)O)O
17*	C(COC(=O)CCCCCCCCC)(OC(=O)CCCCCC=CCCCCCCCCC)COP(=O)(OC1C(C(C(C(C1O)O)O)O)O)O
19	C(COC(=O)CCCCCCCCC)(OC(=O)CCCCCCCCCCCCCCCCCC)COP(=O)(OC1C(C(C(C(C1O)O)O)O)O)O
19*	C(COC(=O)CCCCCCCCC)(OC(=O)CCCCCC=CCCCCCCCCCCC)COP(=O)(OC1C(C(C(C(C1O)O)O)O)O)O
20	C(COC(=O)CCCCCCCCC)(OC(=O)CCCCCCCCCCCCCCCCCCC)COP(=O)(OC1C(C(C(C(C1O)O)O)O)O)O
20*	C(COC(=O)CCCCCCCCC)(OC(=O)CCCCCC=CCCCCCCCCCCCC)COP(=O)(OC1C(C(C(C(C1O)O)O)O)O)O

2.4 Structure Similarity Measures

Similarity s between a pair of lipids was calculated using six methods 1) LINGO 2) OpenBabel FP2 Fingerprint 3) Bioisosteric similarity 4) SMILIGN 5) Smith Waterman Local Alignment 6) Levenshtein distance.

The distance d between a pair of molecules was computed from similarity s [Equation 2.1]. By construction, d is a positive value guaranteed to lie between 0 (identical structures) and 1 (the maximum possible difference).

$$d = 1 - s \tag{2.1}$$

2.4.1 LINGO

The similarity s_l between a pair of SMILES strings A, B is calculated by Equation 1.2. Similarity score s_l was converted to a distance by Equation 2.1.

LINGOs were generated by step-wise linear fragmentation of a SMILES string. The authors noted that these changes 1. improve statistical sampling in QSAR models 2. but prevent reconstruction of unmodified SMILES strings. In this study, I did not apply changes to SMILES strings because my interest is not in QSAR models. The method described in the original paper used a fixed value, $q = 4$ in equation 1.1, I retain the same value.

Vidal *et al.* [91] generated LINGOs from canonical SMILES only. But in this study, LINGO method is tested on Open Babel canonical SMILES, CACTVS canonical SMILES and template SMILES. LINGO distances for ceramides and phosphotidyl inositol molecules were calculated by first drawing structures in PubChem Sketcher, exported to SDF format and converted to SMILES as per the procedure described in section 2.3.

2.4.2 FP2 Fingerprint

FP2 fingerprint similarity score s_f is generated with a set of SMILES as input to the Open Babel library version 2.3.2 [84]. s_f was later converted to distance [Equation 2.1].

Table 2.15: Overview of edits made to SMILES

deleted		*modified*		
symbol	*description*	*symbol*	*replacement*	*description*
C@	carbon chirality	Cl	D	Chlorine
+ , -	charge	=	G	double bond
H	hydrogen	#	G	triple bond
Na	sodium	O	E	oxygen
*, .	wild and join rule	(,)	K, L	branch open, close
1 - 9	cyclic notation	@	R	other chirality
][	atom delimiters	Br	A	Bromine
		/ \	Q, M	cis - trans

2.4.3 Bioisosteric similarity score

The source code package to calculate the similarity score was obtained from Krier *et al.* [108]. The script *querysmiles.pl* is used with CACTVS canonical SMILES as input. The similarity score s_b is converted to distance d_b [Equation 2.1].

2.4.4 SMILIGN

A new method was tested in this work, taking advantage of an existing protein sequence alignment program [99]. First, SMILES strings were mapped into an alphabet of size 20 as given in Table 2.15. This was necessary as the program expects 20 symbols corresponding to the 20 amino acids.

An identity matrix was used for scoring alignments [Equation 2.2]. Gap opening and gap extension were forbidden by assigning high penalty -10000.

$$s(a_i, b_j) = \begin{cases} +1, & a_i = b_j \\ -10000, & a_i \neq b_j \end{cases} \tag{2.2}$$

No limit was set to the number of iterations, so alignments were optimized until they converged. The similarity score s_s was calculated for each pair of aligned SMILES after the final Multiple Sequence Alignment [Equation 2.3] and later converted to distance [Equation 2.1]

$$s_s = \frac{n}{l} \tag{2.3}$$

n is the number of gaps in the alignment, l is the length of the alignment for the specific SMILES pair.

2.4.5 Smith-Waterman Alignment

Given a pair of SMILES, (a, b) of length i and j respectively, alignments were scored with an identity matrix [Equation 2.4]. Gap opening and gap widening penalties were set to -0.5.

$$s(a_i, b_j) = \begin{cases} +1, & a_i = b_j \\ -10, & a_i \neq b_j \end{cases} \tag{2.4}$$

Similarity score s_W is calculated from the number of mis-matches in the alignment n and length l of the longer SMILES $max(i, j)$

$$s_W = \frac{n}{l} \tag{2.5}$$

Smith-Waterman implementation by Forrest Bao[†] is used. Similarity score is converted to distance by Equation 2.1.

2.4.6 Levenshtein distance

Levenshtein algorithm [100, 142] implemented by Martin Schimmels[‡] was used. For a pair of SMILES (a, b) with length m, n respectively, a substitution matrix with a cost to align $a_i \rightarrow b_j$ is set to 0 if $a_i = b_j$ and 1 if $a_i \neq b_j$. Gap opening and extension cost was set to 1. The sum of all edit costs required to completely transform $a \rightarrow b$ is normalized by the length of longer SMILES $max(m, n)$ to generate levenshtein distance d_l.

[†] http://fsbao.net

[‡] http://code.activestate.com

2.5 Principal Component Analysis (PCA)

PCA was performed using the *gdata* library in R [143]. Principal components were plotted using package *scatterplot3d* [144]. Interactive plots were generated using library *RSVGTipsDevice*.

2.6 Lipidome Juxtaposition Score (LUX) Calculation

The LUX score is based on the Hausdorff distance [145, 146] and summarizes the similarity between lipidomes. Levenshtein distance is the chosen metric between lipids. The maximum of the two average Hausdorff distances [Equation 1.4] is used in this study and henceforth it is referred as Lipidome Juxtapostion (LUX) score [Equation 2.6].

$$d_{LUX}(AB) = max\ (d_{H_6}(AB),\ d_{H_6}(BA)) \tag{2.6}$$

LUX score between sets of lipids is a also a metric (similar to Levenshtein distance) and holds the four conditions of a metric-space, i.e, Non-negativity, Identity of indiscernibles, Symmetry and Triangular inequality [Fig. 4.1].

2.7 Hierarchical Cluster Analysis

Complete linkage clustering was performed with R, version 2.14.1, library – ‘stats’ and function ‘hclust’.

For yeast elongase mutants, three pairwise lipidome distance matrices were used as input to the clustering program 1. LUX score(s) 2. Pearson correlation coefficient distances (calculated from lipid abundance values) and 3. normalized number of common-lipids. Only LUX scores were used as distance matrices for fruit fly and human lung lipidomes.

2.8 Error Modeling

An error model for the lipidomes was generated by taking each measured lipid quantity x, and adding Gaussian-distributed noise with a fixed standard deviation.

The detection limit t_{detect} and standard deviation σ were defined so that only low abundant lipids were significantly affected. In R language *rnorm* function gives Gaussian noise.

Error modeling returns a new list of lipid species that has a smaller number of lipids, compared to the input list (for example, one error model returned 238 lipids from an input of 248). The error modeling was repreated 100 times, generating 100 new lists. Pairwise LUX scores were calculated for eight yeast lipidomes after each error model. 100 pairwise LUX score matrices were used to generate the corresponding 100 hierarchical clusters.

The number of times a particular branch (of the master list derived hierarchical cluster) re-appears in the error model list derived cluster was counted using the R library, ape::boot.phylo::prop.part [147]. This number is plotted alongside the branch in yeast lipidome hierarchical clusters [Fig. 3.15b]. The higher values of brach-reoccurance frequency indicate robust association.

Based on the distribution of lipid abundances [Table 2.16], the following three parameter sets were used for yeast lipidome error model:

1. $t_{detect} = 0.003$ mol%, $\sigma = 0.001$
2. $t_{detect} = 0.003$ mol%, $\sigma = 0.002$
3. $t_{detect} = 0.006$ mol%, $\sigma = 0.004$

Based on the distribution of lipid abundances for lung lipidome [Table 2.17], the following two parameter sets were used for error model:

1. $t_{detect} = 0.003$ mol%, $\sigma = 0.002$
2. $t_{detect} = 0.005$ mol%, $\sigma = 0.002$

Table 2.16: Frequency distribution of lipid abundances for yeast lipidome

Bin (mol %)	Frequency
0	0
0.001	12
0.002	25
0.003	21
0.004	19
0.005	19
0.006	21
0.007	13
0.008	13
0.009	17
0.01	18
> 0.01	1150

Table 2.17: Frequency distribution of lipid abundances for human lung lipidome

Bin	Frequency	%	Cumulative %	Cumulative % w/ base 0.001
0	3563	26.64	26.64	
0.001	19	0.14	26.79	0.14
0.002	28	0.21	26.99	0.35
0.003	55	0.41	27.41	0.76
0.004	91	0.68	28.09	1.44
0.005	89	0.67	28.75	2.11
0.006	110	0.82	29.57	2.93
0.007	130	0.97	30.55	3.90
0.008	166	1.24	31.79	5.14
0.009	130	0.97	32.76	6.12
0.01	137	1.02	33.78	7.14
> 0.01	8855		100.00	
Number of lipids		311		
Number of samples		43		
Number of measurements		9720		

Results

3.1 Evaluation of SMILES Representation by Measuring Structure Similarity

In this section, 3 SMILES specifications (1. CACTVS canonical 2. Open Babel canonical and 3. Template) were evaluated, whether they are suitable for use (together with distance scoring methods) to measure changes in double-bond and hydroxyl group position of selected lipid structures. For each SMILES specification, I constructed a pairwise distance matrix of 17 ceramides and 16 PI structures [section 2.1.1], with 6 distance scoring methods [section 2.4]. In each case, I looked at two properties. The first is straight forward, if two molecules are not identical, the distance between them must be non-zero [section 2.4, Equation 2.1]. The second is less formal, for a series of pairs, does the distance increase gradually with the intuitive difference between the molecules? For example, does the distance increase as the position of the double bond or hydroxylation is changed?

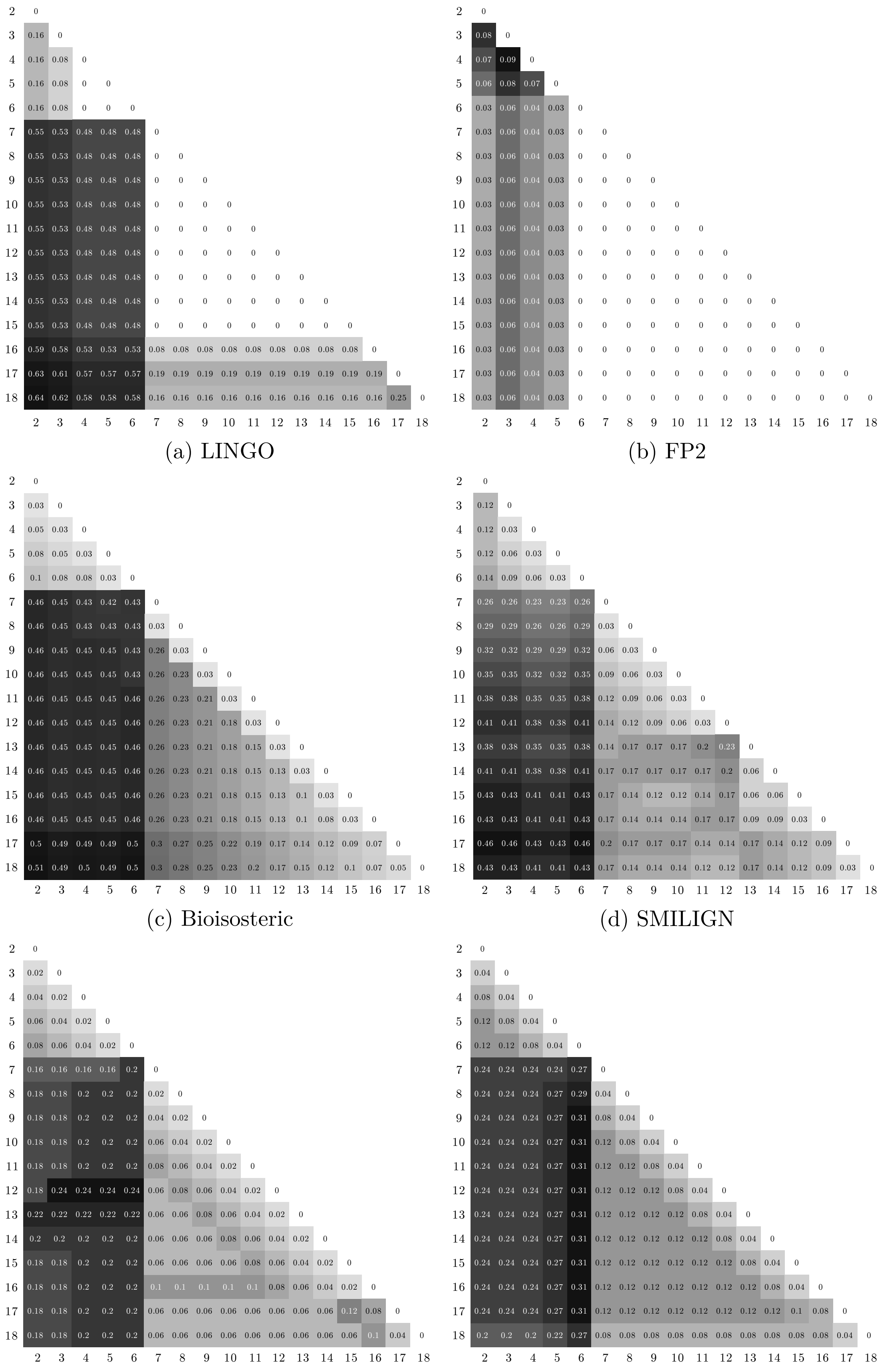
3.1.1 CACTVS canonical SMILES

First, the CACTVS canonical SMILES representation was evaluated by measuring the structural similarity between 17 ceramides with six distance scoring methods. The results with each distance measure are described below.

39 pairs of ceramide structures have zero LINGO distance, which is not expected because their structures are obviously different [Figs. 3.1a, 2.1a]. The highest LINGO distance is between 2 and 18, henceforth written as {2-18}[†], is expected because the separation of hydroxyl group is farthest between these two molecules [section 2.1.1]. Lower values were expected for molecule pairs with closer positioning of hydroxyl group, such as {2-3}, {3-4} but the molecule 16 paired with {7--15}[‡] and {3-4, 3-5, 3-6} have the lowest LINGO distance value [lighter shade

[†] Hyphen symbol separating two molecule names indicates a pair. Curly brackets indicate a set of molecules or molecule pairs.

[‡] Double hyphenation inside the curly brackets indicates a range of molecules or molecule pairs.



	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
2	0																
3	0.16	0															
4	0.16	0.08	0														
5	0.16	0.08	0	0													
6	0.16	0.08	0	0	0												
7	0.55	0.53	0.48	0.48	0.48	0											
8	0.55	0.53	0.48	0.48	0.48	0	0										
9	0.55	0.53	0.48	0.48	0.48	0	0	0									
10	0.55	0.53	0.48	0.48	0.48	0	0	0	0								
11	0.55	0.53	0.48	0.48	0.48	0	0	0	0	0							
12	0.55	0.53	0.48	0.48	0.48	0	0	0	0	0	0						
13	0.55	0.53	0.48	0.48	0.48	0	0	0	0	0	0	0					
14	0.55	0.53	0.48	0.48	0.48	0	0	0	0	0	0	0	0				
15	0.55	0.53	0.48	0.48	0.48	0	0	0	0	0	0	0	0	0			
16	0.59	0.58	0.53	0.53	0.53	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0		
17	0.63	0.61	0.57	0.57	0.57	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0	
18	0.64	0.62	0.58	0.58	0.58	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.25	0

(a) LINGO

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
2	0																
3	0.08	0															
4	0.07	0.09	0														
5	0.06	0.08	0.07	0													
6	0.03	0.06	0.04	0.03	0												
7	0.03	0.06	0.04	0.03	0	0											
8	0.03	0.06	0.04	0.03	0	0	0										
9	0.03	0.06	0.04	0.03	0	0	0	0									
10	0.03	0.06	0.04	0.03	0	0	0	0	0								
11	0.03	0.06	0.04	0.03	0	0	0	0	0	0							
12	0.03	0.06	0.04	0.03	0	0	0	0	0	0	0						
13	0.03	0.06	0.04	0.03	0	0	0	0	0	0	0	0					
14	0.03	0.06	0.04	0.03	0	0	0	0	0	0	0	0	0				
15	0.03	0.06	0.04	0.03	0	0	0	0	0	0	0	0	0	0			
16	0.03	0.06	0.04	0.03	0	0	0	0	0	0	0	0	0	0	0		
17	0.03	0.06	0.04	0.03	0	0	0	0	0	0	0	0	0	0	0	0	
18	0.03	0.06	0.04	0.03	0	0	0	0	0	0	0	0	0	0	0	0	0

(b) FP2

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
2	0																
3	0.03	0															
4	0.05	0.03	0														
5	0.08	0.05	0.03	0													
6	0.1	0.08	0.08	0.03	0												
7	0.46	0.45	0.43	0.42	0.43	0											
8	0.46	0.45	0.43	0.43	0.43	0.03	0										
9	0.46	0.45	0.45	0.45	0.43	0.26	0.03	0									
10	0.46	0.45	0.45	0.45	0.43	0.26	0.23	0.03	0								
11	0.46	0.45	0.45	0.45	0.46	0.26	0.23	0.21	0.03	0							
12	0.46	0.45	0.45	0.45	0.46	0.26	0.23	0.21	0.18	0.03	0						
13	0.46	0.45	0.45	0.45	0.46	0.26	0.23	0.21	0.18	0.15	0.03	0					
14	0.46	0.45	0.45	0.45	0.46	0.26	0.23	0.21	0.18	0.15	0.13	0.03	0				
15	0.46	0.45	0.45	0.45	0.46	0.26	0.23	0.21	0.18	0.15	0.13	0.1	0.03	0			
16	0.46	0.45	0.45	0.45	0.46	0.26	0.23	0.21	0.18	0.15	0.13	0.1	0.08	0.03	0		
17	0.5	0.49	0.49	0.49	0.5	0.3	0.27	0.25	0.22	0.19	0.17	0.14	0.12	0.09	0.07	0	
18	0.51	0.49	0.5	0.49	0.5	0.3	0.28	0.25	0.23	0.2	0.17	0.15	0.12	0.1	0.07	0.05	0

(c) Bioisosteric

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
2	0																
3	0.12	0															
4	0.12	0.03	0														
5	0.12	0.06	0.03	0													
6	0.14	0.09	0.06	0.03	0												
7	0.26	0.26	0.23	0.23	0.26	0											
8	0.29	0.29	0.26	0.26	0.29	0.03	0										
9	0.32	0.32	0.29	0.29	0.32	0.06	0.03	0									
10	0.35	0.35	0.32	0.32	0.35	0.09	0.06	0.03	0								
11	0.38	0.38	0.35	0.35	0.38	0.12	0.09	0.06	0.03	0							
12	0.41	0.41	0.38	0.38	0.41	0.14	0.12	0.09	0.06	0.03	0						
13	0.38	0.38	0.35	0.35	0.38	0.14	0.17	0.17	0.17	0.2	0.23	0					
14	0.41	0.41	0.38	0.38	0.41	0.17	0.17	0.17	0.17	0.17	0.2	0.06	0				
15	0.43	0.43	0.41	0.41	0.43	0.17	0.14	0.12	0.12	0.14	0.17	0.06	0.06	0			
16	0.43	0.43	0.41	0.41	0.43	0.17	0.14	0.14	0.14	0.17	0.17	0.09	0.09	0.03	0		
17	0.46	0.46	0.43	0.43	0.46	0.2	0.17	0.17	0.17	0.14	0.14	0.17	0.14	0.12	0.09	0	
18	0.43	0.43	0.41	0.41	0.43	0.17	0.14	0.14	0.14	0.12	0.12	0.17	0.14	0.12	0.09	0.03	0

(d) SMILIGN

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
2	0																
3	0.02	0															
4	0.04	0.02	0														
5	0.06	0.04	0.02	0													
6	0.08	0.06	0.04	0.02	0												
7	0.16	0.16	0.16	0.16	0.2	0											
8	0.18	0.18	0.2	0.2	0.2	0.02	0										
9	0.18	0.18	0.2	0.2	0.2	0.04	0.02	0									
10	0.18	0.18	0.2	0.2	0.2	0.06	0.04	0.02	0								
11	0.18	0.18	0.2	0.2	0.2	0.08	0.06	0.04	0.02	0							
12	0.18	0.24	0.24	0.24	0.24	0.06	0.08	0.06	0.04	0.02	0						
13	0.22	0.22	0.22	0.22	0.22	0.06	0.06	0.08	0.06	0.04	0.02	0					
14	0.2	0.2	0.2	0.2	0.2	0.06	0.06	0.06	0.08	0.06	0.04	0.02	0				
15	0.18	0.18	0.2	0.2	0.2	0.06	0.06	0.06	0.06	0.08	0.06	0.04	0.02	0			
16	0.18	0.18	0.2	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.08	0.06	0.04	0.02	0		
17	0.18	0.18	0.2	0.2	0.2	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.12	0.08	0	
18	0.18	0.18	0.2	0.2	0.2	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.1	0.04	0

(e) Smith-Waterman

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
2	0																
3	0.04	0															
4	0.08	0.04	0														
5	0.12	0.08	0.04	0													
6	0.12	0.12	0.08	0.04	0												
7	0.24	0.24	0.24	0.24	0.27	0											
8	0.24	0.24	0.24	0.27	0.29	0.04	0										
9	0.24	0.24	0.24	0.27	0.31	0.08	0.04	0									
10	0.24	0.24	0.24	0.27	0.31	0.12	0.08	0.04	0								
11	0.24	0.24	0.24	0.27	0.31	0.12	0.12	0.08	0.04	0							
12	0.24	0.24	0.24	0.27	0.31	0.12	0.12	0.12	0.08	0.04	0						
13	0.24	0.24	0.24	0.27	0.31	0.12	0.12	0.12	0.12	0.08	0.04	0					
14	0.24	0.24	0.24	0.27	0.31	0.12	0.12	0.12	0.12	0.12	0.08	0.04	0				
15	0.24	0.24	0.24	0.27	0.31	0.12	0.12	0.12	0.12	0.12	0.12	0.08	0.04	0			
16	0.24	0.24	0.24	0.27	0.31	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.08	0.04	0		
17	0.24	0.24	0.24	0.27	0.31	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.1	0.08	0	
18	0.2	0.2	0.2	0.22	0.27	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.04	0

(f) Levenshtein

Figure 3.1: Pairwise distances for the CACTVS canonical SMILES representation of 17 ceramides. Distance values in the range of 0 to 1, were calculated with six methods (a) LINGO (b) FP2 (c) Bioisosteric similarity (d) SMILIGN (e) Smith-Waterman and (f) Levenshtein distance [section 2.4]. Rows and columns were numbered from 2 to 18. These numbers correspond to the 17 ceramide molecules that were named, also from 2 to 18 [section 2.1.1]. The distance between a pair of ceramides can be inferred from the matrix by looking up the corresponding row and column (for example, the Levenshtien distance between ceramides, 2 and 18 is displayed in matrix (f) column 2 - row 18). Only the lower triangular representation of distances were shown because the values are symmetrical (i.e distance from 2 to 18 is same as the distance from 18 to 2). Grey scaling was applied as the background for distance values. The intensity is proportional to the distance - white background for zero values and darker shade for higher values. Grey scaling is applied separately for each distance matrix (for example, the Biosisosteric distance value 0.2 has different intensity (c) as compared to the Levenshtein distance of the same value (f)).

values, Fig. 3.1a]. 17 ceramides could be divided in two groups based on the LINGO distance matrix 1. {2--6} and 2. {7--18}. The distance values were higher for pairs that combine a molecule from group 1 and the second from group 2 (for example {2-7}, {2-18}, {6-7} and {6-18}) [darker shaded values in Fig. 3.1a].

The FP2 distance is zero for 78 pairs of ceramides [Fig. 3.1b]. More zero values were observed in the FP2 matrix as compared to the LINGO. The highest FP2 distance value is for the pair {3-4} but the expectation is to see the highest value for {2-18}, because, as noted earlier, {2-18} has the farthest separation of hydroxyl group and hence it the most dissimilar pair in matrix [Figs. 3.1b, 2.1a]. The second and third highest FP2 distances were between {2-3} and {4-5} respectively, which was also not expected because the hydroxyl group is closer in these pairs [Figs. 3.1b, 2.1a].

The Bioisoteric distance is non-zero for all non-identical pairs [Fig. 3.1c]. The lowest distance is between {2-3}, which is expected because the hydroxyl group is located just one carbon atom apart in this pair and the largest distance is between {2-18}, again as expected, because of the farthest separation of hydroxyl group [Figs. 3.1c, 2.1a]. Except {6-7, 16-17 and 17-18}, all pairs with one carbon atom separation of hydroxyl position (example, {2-3, 3-4, 4-5 etc.}) have 0.03 distance [Fig. 3.1c]. A gradual decrease in the distance values was observed for the pairs {18-7 -- 18-17}

For example, 16 paired with {7--15} refers to 16-7, 16-8 and so on until 16-15 [rows 9 to 18, columns 7-17, Fig. 3.1c]. The distance matrix indicates that hydroxyl position at 6 appears to be a threshold; after position 6, the pairwise distances decrease gradually [rows 9 to 18, columns 7-17, Fig. 3.1c].

SMILIGN distances were greater than zero for all non-identical pairs of ceramides [Fig. 3.1d]. {12-13} has higher distance value in comparison to the pairs with similar one-carbon separation in hydroxyl position such as {13-14, 14-15}. Based on the distances calculated with SMILIGN method, the 17 ceramides could be split in two groups 1. {2--6} and 2. {7--18}. The molecule pairs formed by the group 1 ceramides have lower distances [rows 2--6, Fig. 3.1d]. The pairs formed by group 2 molecules also have lower distance values [columns 7--18, Fig. 3.1d] but the pairs that have one molecule from group 1, and second from group 2, have higher distances [values with darker shade as background, Fig. 3.1d].

The Smith-Waterman distance is non-zero for all non-identical ceramide pairs [Fig. 3.1e]. Two distinct triangular clusters are noticeable in the distance matrix. In the first cluster, the molecules {2--6} have lower distances between them [lighter shade values at the top left corner, Fig. 3.1e]. Second cluster is made of molecules {7--18}. The distances for pairs between the two clusters are higher [darker shade values, Fig. 3.1e]. The highest distance in the set is between the pairs {12-2, 12-3, 12-4, 12-5} but the expectation is to see the value for the most structurally different pair, {2-18}.

The Levenshtein distances for all non-identical ceramides are non-zero [Fig. 3.1f]. Based on the pairwise Levenshtein distance matrix, the ceramides could be split in two groups 1. {2--6} and 2. {7--18} (similar to the Smith-Waterman distance matrix). The pairwise distances within the first set, and within the second set are lower than the average [lighter shaded values, Fig. 3.1f] but the distances for pairs between two groups are higher [darker shaded values, Fig. 3.1f].

3.1.2 Open Babel canonical SMILES

16 PI molecules were used for evaluating Open Babel canonical SMILES. LINGO distance for 34 non-identical pairs of PI structures is zero [Fig. 3.2a] which is not not expected because they do not have identical structures [Fig. 2.1b]. The highest

distance value is between {10-20*}, was expected, because, their structures are the most dissimilar [Figs. 3.2a, 2.1b]. However, the highest distance value was observed for additional 37 pairs [dark shaded values, Fig. 3.2a]. The lowest value is always with molecules paired with 10 [first column, Fig. 3.2a]. The distance value is zero for the pairs when both partners have saturated acyl chains (example, {11-20, 12-19, 13-15}), except when one partner is molecule PI 10 [first column, Fig. 3.2a].

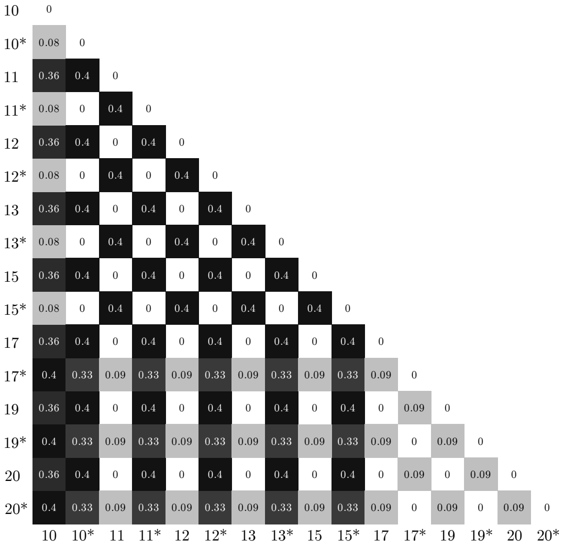
FP2 distance is zero for 56 pairs of non-identical PI structures [Fig. 3.2b]. Zero distance value is observed for pairs when both partners have saturated acyl chains (example, {11-20, 12-19, 13-15}) but the value is non-zero when at least one partner has unsaturated acyl chain (example, {11*-20, 12*-19, 13*-15}).

Bioisosteric algorithm [91] specifies CACTVS canonical SMILES as the only valid input. Because of this limitation, the Bioisosteric distance could not be calculated for Open Babel canonical SMILES representation of 16 PI molecules.

The pairwise distances of 16 PI's with SMILIGN, Smith-Waterman and Levenshtein methods were similar [Figs. 3.2c, 3.2d and 3.2e respectively]. Based on these three distance matrices, PI molecules could be split in two groups 1. {11--15*} and 2. {17--20*}. In the first group, the distances between pairs when both partners have saturated acyl chain (example, {11-12, 12-15}) and when both parters have unsaturated acyl chain (example, {11*-12*, 13*-15*}), have lower distances, as compared to the pairs when one partner has saturated acyl chain and the second has unsaturated acyl chain (example, {11-11*, 13*-15}), resulting in a "checker-board" like pattern [Figs. 3.2c, 3.2d and 3.2e]. The distance values were lower when both partners of a pair were from group 2 [light shaded values in bottom-right region, Fig. 3.2c]. The distance for pairs when one partner is from group 1 and second from group 2, are interesting because the values are consistently lower when the group 1 partner has saturated acyl chain {11, 12, 13, 15}, and higher when the group 1 partner has unsaturated acyl chain {11*, 12*, 13*, 15*}.

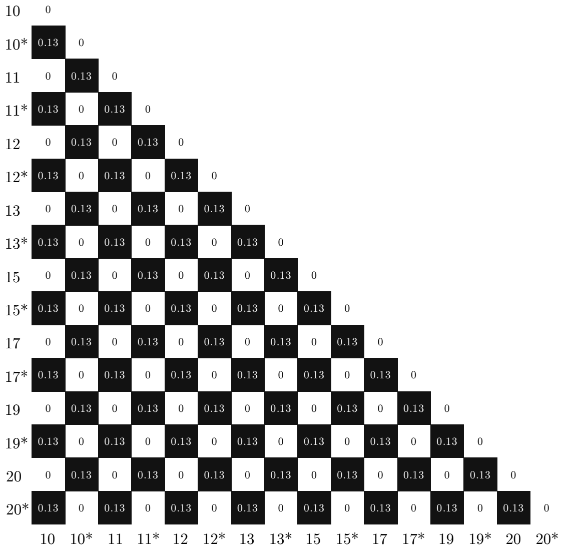
3.1.3 Template SMILES

The LINGO distance between 47 pairs of ceramides is non-zero and it is zero for 108 pairs [Fig. 3.4a]. LINGO distances calculated from the template SMILES of



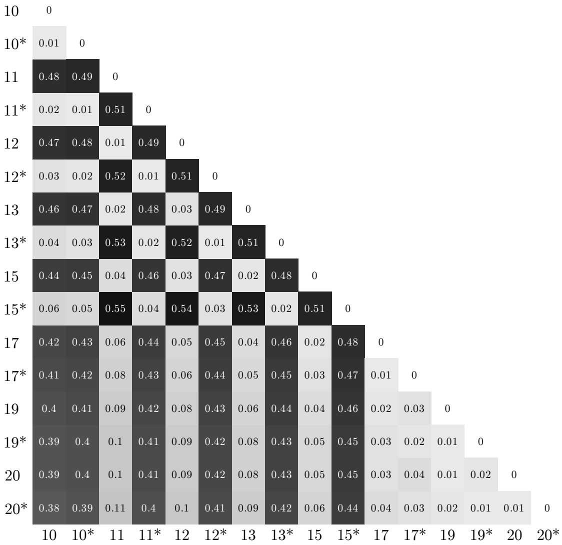
	10	10*	11	11*	12	12*	13	13*	15	15*	17	17*	19	19*	20	20*
10	0															
10*	0.08	0														
11	0.36	0.4	0													
11*	0.08	0	0.4	0												
12	0.36	0.4	0	0.4	0											
12*	0.08	0	0.4	0	0.4	0										
13	0.36	0.4	0	0.4	0	0.4	0									
13*	0.08	0	0.4	0	0.4	0	0.4	0								
15	0.36	0.4	0	0.4	0	0.4	0	0.4	0							
15*	0.08	0	0.4	0	0.4	0	0.4	0	0.4	0						
17	0.36	0.4	0	0.4	0	0.4	0	0.4	0	0.4	0					
17*	0.4	0.33	0.09	0.33	0.09	0.33	0.09	0.33	0.09	0.33	0.09	0				
19	0.36	0.4	0	0.4	0	0.4	0	0.4	0	0.4	0	0.09	0			
19*	0.4	0.33	0.09	0.33	0.09	0.33	0.09	0.33	0.09	0.33	0.09	0	0.09	0		
20	0.36	0.4	0	0.4	0	0.4	0	0.4	0	0.4	0	0.09	0	0.09	0	
20*	0.4	0.33	0.09	0.33	0.09	0.33	0.09	0.33	0.09	0.33	0.09	0	0.09	0	0.09	0

(a) LINGO



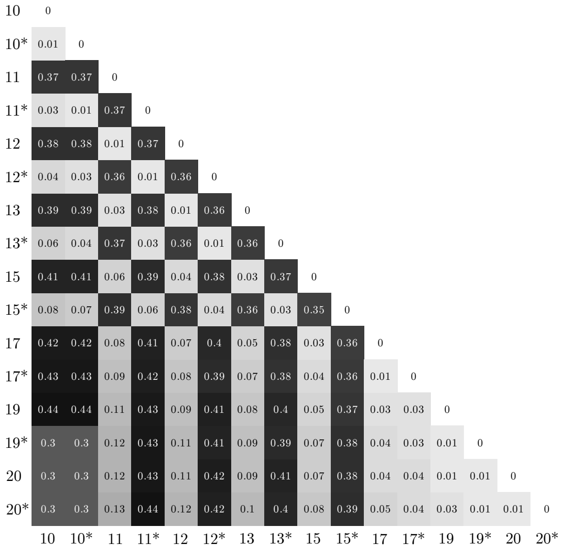
	10	10*	11	11*	12	12*	13	13*	15	15*	17	17*	19	19*	20	20*
10	0															
10*	0.13	0														
11	0	0.13	0													
11*	0.13	0	0.13	0												
12	0	0.13	0	0.13	0											
12*	0.13	0	0.13	0	0.13	0										
13	0	0.13	0	0.13	0	0.13	0									
13*	0.13	0	0.13	0	0.13	0	0.13	0								
15	0	0.13	0	0.13	0	0.13	0	0.13	0							
15*	0.13	0	0.13	0	0.13	0	0.13	0	0.13	0						
17	0	0.13	0	0.13	0	0.13	0	0.13	0	0.13	0					
17*	0.13	0	0.13	0	0.13	0	0.13	0	0.13	0	0.13	0				
19	0	0.13	0	0.13	0	0.13	0	0.13	0	0.13	0	0.13	0			
19*	0.13	0	0.13	0	0.13	0	0.13	0	0.13	0	0.13	0	0.13	0		
20	0	0.13	0	0.13	0	0.13	0	0.13	0	0.13	0	0.13	0	0.13	0	
20*	0.13	0	0.13	0	0.13	0	0.13	0	0.13	0	0.13	0	0.13	0	0.13	0

(b) FP2



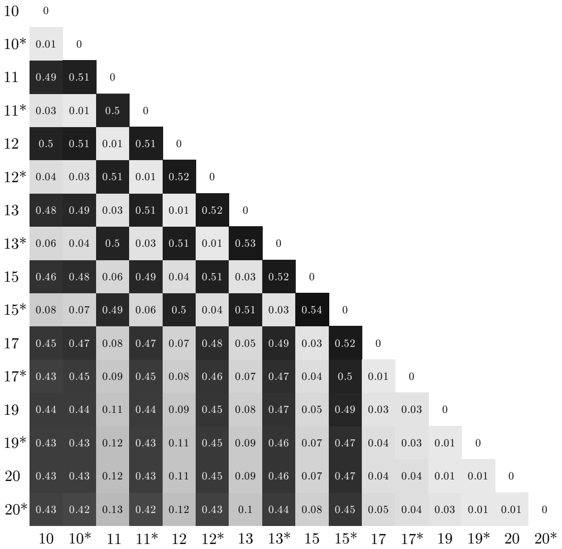
	10	10*	11	11*	12	12*	13	13*	15	15*	17	17*	19	19*	20	20*
10	0															
10*	0.01	0														
11	0.48	0.49	0													
11*	0.02	0.01	0.51	0												
12	0.47	0.48	0.01	0.49	0											
12*	0.03	0.02	0.52	0.01	0.51	0										
13	0.46	0.47	0.02	0.48	0.03	0.49	0									
13*	0.04	0.03	0.53	0.02	0.52	0.01	0.51	0								
15	0.44	0.45	0.04	0.46	0.03	0.47	0.02	0.48	0							
15*	0.06	0.05	0.55	0.04	0.54	0.03	0.53	0.02	0.51	0						
17	0.42	0.43	0.06	0.44	0.05	0.45	0.04	0.46	0.02	0.48	0					
17*	0.41	0.42	0.08	0.43	0.06	0.44	0.05	0.45	0.03	0.47	0.01	0				
19	0.4	0.41	0.09	0.42	0.08	0.43	0.06	0.44	0.04	0.46	0.02	0.03	0			
19*	0.39	0.4	0.1	0.41	0.09	0.42	0.08	0.43	0.05	0.45	0.03	0.02	0.01	0		
20	0.39	0.4	0.1	0.41	0.09	0.42	0.08	0.43	0.05	0.45	0.03	0.04	0.01	0.02	0	
20*	0.38	0.39	0.11	0.4	0.1	0.41	0.09	0.42	0.06	0.44	0.04	0.03	0.02	0.01	0.01	0

(c) SMILIGN



	10	10*	11	11*	12	12*	13	13*	15	15*	17	17*	19	19*	20	20*
10	0															
10*	0.01	0														
11	0.37	0.37	0													
11*	0.03	0.01	0.37	0												
12	0.38	0.38	0.01	0.37	0											
12*	0.04	0.03	0.36	0.01	0.36	0										
13	0.39	0.39	0.03	0.38	0.01	0.36	0									
13*	0.06	0.04	0.37	0.03	0.36	0.01	0.36	0								
15	0.41	0.41	0.06	0.39	0.04	0.38	0.03	0.37	0							
15*	0.08	0.07	0.39	0.06	0.38	0.04	0.36	0.03	0.35	0						
17	0.42	0.42	0.08	0.41	0.07	0.4	0.05	0.38	0.03	0.36	0					
17*	0.43	0.43	0.09	0.42	0.08	0.39	0.07	0.38	0.04	0.36	0.01	0				
19	0.44	0.44	0.11	0.43	0.09	0.41	0.08	0.4	0.05	0.37	0.03	0.03	0			
19*	0.3	0.3	0.12	0.43	0.11	0.41	0.09	0.39	0.07	0.38	0.04	0.03	0.01	0		
20	0.3	0.3	0.12	0.43	0.11	0.42	0.09	0.41	0.07	0.38	0.04	0.04	0.01	0.01	0	
20*	0.3	0.3	0.13	0.44	0.12	0.42	0.1	0.4	0.08	0.39	0.05	0.04	0.03	0.01	0.01	0

(d) Smith-Waterman



	10	10*	11	11*	12	12*	13	13*	15	15*	17	17*	19	19*	20	20*
10	0															
10*	0.01	0														
11	0.49	0.51	0													
11*	0.03	0.01	0.5	0												
12	0.5	0.51	0.01	0.51	0											
12*	0.04	0.03	0.51	0.01	0.52	0										
13	0.48	0.49	0.03	0.51	0.01	0.52	0									
13*	0.06	0.04	0.5	0.03	0.51	0.01	0.53	0								
15	0.46	0.48	0.06	0.49	0.04	0.51	0.03	0.52	0							
15*	0.08	0.07	0.49	0.06	0.5	0.04	0.51	0.03	0.54	0						
17	0.45	0.47	0.08	0.47	0.07	0.48	0.05	0.49	0.03	0.52	0					
17*	0.43	0.45	0.09	0.45	0.08	0.46	0.07	0.47	0.04	0.5	0.01	0				
19	0.44	0.44	0.11	0.44	0.09	0.45	0.08	0.47	0.05	0.49	0.03	0.03	0			
19*	0.43	0.43	0.12	0.43	0.11	0.45	0.09	0.46	0.07	0.47	0.04	0.03	0.01	0		
20	0.43	0.43	0.12	0.43	0.11	0.45	0.09	0.46	0.07	0.47	0.04	0.04	0.01	0.01	0	
20*	0.43	0.42	0.13	0.42	0.12	0.43	0.1	0.44	0.08	0.45	0.05	0.04	0.03	0.01	0.01	0

(e) Levenshtein

Figure 3.2: Pairwise distances for the Open Babel canonical SMILES representation of 16 phosophatidyl inositol structures. Distance values in the range of 0 to 1, were calculated with five methods (a) LINGO (b) FP2 (c) SMILIGN (d) Smith-Waterman and (e) Levenshtein distance [section 2.4]. Rows and columns were numbered from 10 to 20*. These numbers correspond to the 16 PI molecules that were named, also from 10 to 20* [section 2.1.1]. See figure 3.1 explanation for triangular matrix, grey scaling and procedure to read the distance value for a specific pair.

17 ceramides have more zero values, in comparison to the distances with canonical SMILES representation of same molecules described earlier [Figs. 3.4a, 3.1a]. In comparison, the FP2 distance is zero for 78 non-identical pairs of ceramide structures [Fig. 3.4b].

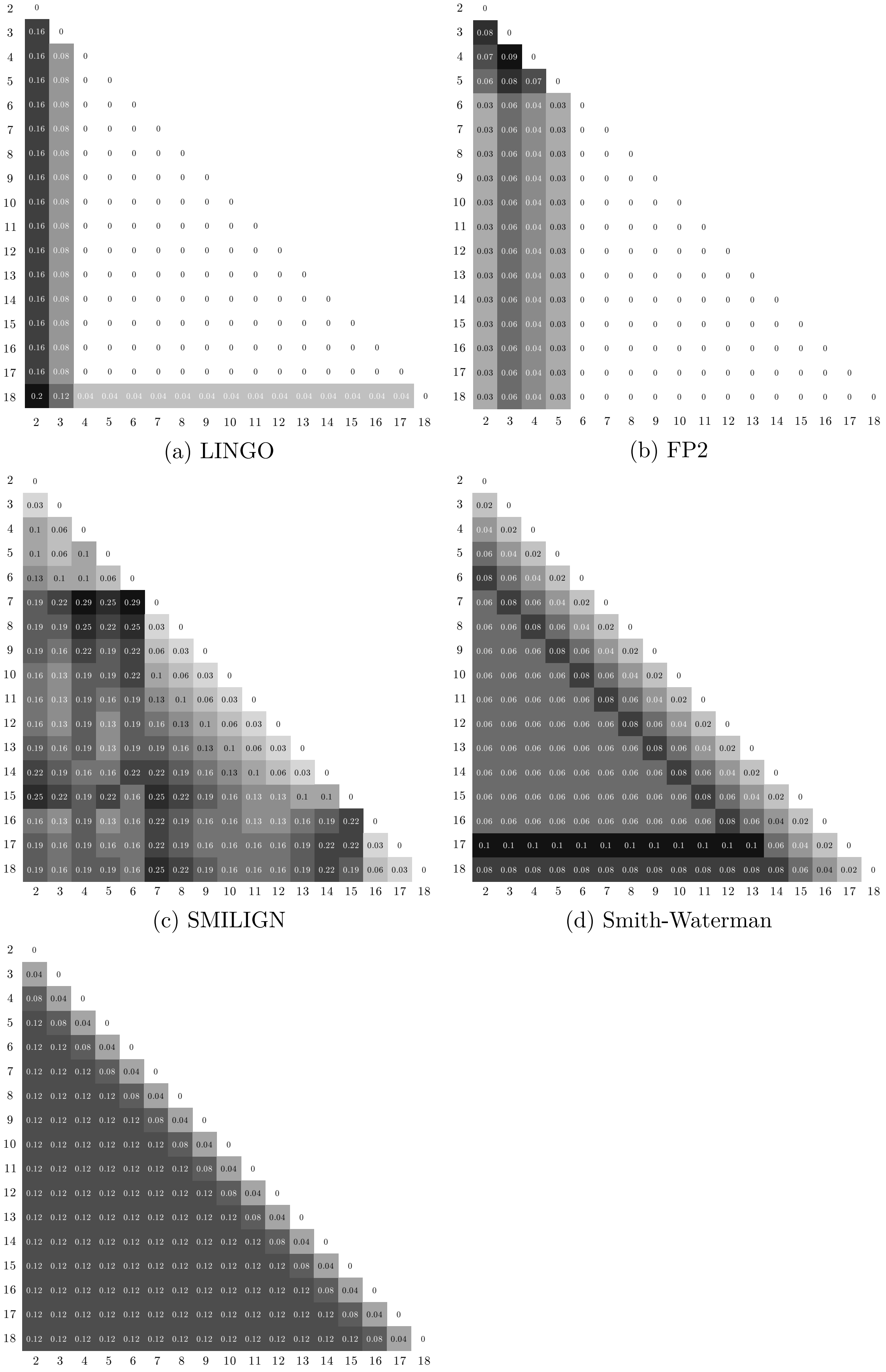
The SMILIGN distances are non-zero for all non-identical pairs of ceramides [Fig. 3.4c]. The ceramide pairs with the hydroxyl group position separated by one methanediyl ($-CH_2-$) group (example, {2-3, 3-4, 6-7}) have varying distances [Fig. 3.4c]. SMILIGN distances do not increase gradually with the increasing separation of hydroxyl group between the pairs, for example, the ceramides paired with molecule 2 ({2-3, 2-4, 2-5 -- 2-17}) show increase in distance with the separation of hydroxyl group until position 15, but beyond that, the distances do not increase [first column, Fig. 3.4c].

The distances associated with ceramide 7 [row and column 7, Fig. 3.4c] were interesting because the value when paired with ceramide 6 is 0.29 but with ceramide 8 is 0.03, although there is one carbon atom separation of hydroxyl group in both pairs {7-6} and {7-8}. The multiple sequence alignments that were generated prior to SMILIGN distance calculation were probed to understand the values observed for ceramide 7 [Fig. 3.3]. Molecules {7--13} share the same alignment position for the hydroxyl group, but {14--18} have a different position [Fig. 3.3]. 17 ceramides can be separated to three groups based on the alignment position of hydroxyl group 1. {2--6} 2. {7--13} and 3. {14--18}. The separation is noticeable in the distance matrix, where the values were lower within- and higher between- the three groups [Fig. 3.3].

The Smith-Waterman distance is non-zero for all non-identical pairs [Fig. 3.4d]. The ceramide pairs with one carbon atom separation of hydroxyl group have the identical 0.02 distance value, and the pairs with two carbon separation have two

Name	Aligned SMILES
2	-CCCC-C---CCCCC---CC-C-C-CCC(O)----C(=O)NC(CO)C(O)C=CCCCCCCCCCC
3	--CCC-C---CCCCC---CC-C-C-CCC(O)--C-C(=O)NC(CO)C(O)C=CCCCCCCCCCC
4	--C-C-C---CCCCC---CC-C-C-CCC(O)C-C-C(=O)NC(CO)C(O)C=CCCCCCCCCCC
5	--CCC-C---CC-CC---CC-C-C-C-C(O)CCC-C(=O)NC(CO)C(O)C=CCCCCCCCCCC
6	--C-C-C---CC-CC---CC-C-C-C-C(O)CCCCC(=O)NC(CO)C(O)C=CCCCCCCCCCC
7	CCCCCCC---CCCCC(O)-C-C-C---C-----C-C(=O)NC(CO)C(O)C=CCCCCCCCCCC
8	CCCCC-C---CCCCC(O)-C-C-C-C-C-----C-C(=O)NC(CO)C(O)C=CCCCCCCCCCC
9	-CCCC-C---CCCCC(O)-C-C-C-C-C---C-C-C(=O)NC(CO)C(O)C=CCCCCCCCCCC
10	-CC-C-C---CCCCC(O)CC-C-C-C-C---C-C-C(=O)NC(CO)C(O)C=CCCCCCCCCCC
11	-CC-C-C---CC-CC(O)CC-C-C-CCC---C-C-C(=O)NC(CO)C(O)C=CCCCCCCCCCC
12	--C-C-C---CC-CC(O)CC-C-C-CCC---CCC-C(=O)NC(CO)C(O)C=CCCCCCCCCCC
13	------C---CCCCC(O)CCCCCCCCCC-----C-C(=O)NC(CO)C(O)C=CCCCCCCCCCC
14	-CCCC-C(O)CC-CC---CC-C-C-C-C---C-C-C(=O)NC(CO)C(O)C=CCCCCCCCCCC
15	-CC-C-C(O)CCCCC---CC-C-C-CCC-----C-C(=O)NC(CO)C(O)C=CCCCCCCCCCC
16	--C-C-C(O)CCCCC---CC-C-C-CCC---C-C-C(=O)NC(CO)C(O)C=CCCCCCCCCCC
17	----C-C(O)CCCCC---CC-C-C-CCC---CCC-C(=O)NC(CO)C(O)C=CCCCCCCCCCC
18	------C(O)CCCCC---CC-CCCCCCC----CC-C(=O)NC(CO)C(O)C=CCCCCCCCCCC

Figure 3.3: Multiple Sequence Alignments of 17 ceramides in template SMILES representation. The 17 ceramides in template SMILES representation were used as input to MUSCLE Multiple Sequence Alignment Program (MSA), as described in methods chapter [section 2.4.4]. The amino acid characters were converted back to SMILES to generate the above list. This back-conversion is not part of the SMILIGN distance calculation workflow, but performed here to improve the readability of MSA output. The MSA symbol for a gap (-) remains unchanged. The hydroxyl group in the acyl chain of ceramides is marked for all 17 aligned SMILES.



(a) LINGO

(b) FP2

(c) SMILIGN

(d) Smith-Waterman

(e) Levenshtein

Figure 3.4: Pairwise distances for the template SMILES representation of 17 ceramides. Distance values in the range of 0 to 1, were calculated with five methods (a) LINGO (b) FP2 (c) SMILIGN (d) Smith-Waterman and (e) Levenshtein distance [section 2.4]. Rows and columns were numbered from 2 to 18. These numbers correspond to the 17 ceramides structures that were named, also from 2 to 18 [section 2.1.1]. See figure 3.1 explanation for triangular matrix, grey scaling and procedure to read the distance value for a specific pair.

times that value. This trend continues, with three and four times the distance values for ceramide pairs with 3 and 4 carbon atom separation of hydroxyl group; with the exception of {13-17} [Fig. 3.4d]. This trend does not continue for pairs with 5 or more carbons separation.

The distance between pairs when one partner is either ceramide 17 or 18, have higher values. For example, the ceramide 2 when paired with {15,16} has the same value 0.06, but when paired with {17,18}, the value is higher. This pattern is true for many ceramides {2--10}. The Smith-Waterman alignments of {2-15,2-16,2-17,2-18} were probed further to identify this discrepancy [Fig. 3.5].

The distance values between ceramides paired with molecule 17 are higher, in comparison to the values when paired with 18 [rows 17 and 18 respectively, 3.4d] which is surprising because the hydroxyl group is nearer between {16-17} rather than {16-18} [Fig. 2.1a].

Levenshtein distance values are non-zero for all non-identical pairs of ceramides [Fig. 3.4e]. One carbon atom separation of hydroxyl group between a pair ({2-3,3-4 and so on}) resulted in the levenshtein distance value 0.04 [distances running parallel to the diagonal]. The distance is two times that value for two carbon atom separation [distances running parallel to the diagonal, Fig. 3.4e]. Three or more carbon atom separation of the hydroxyl group between a pair, resulted in a fixed distance value 0.12.

3.1.4 Section Summary

Three SMILES representations were compared by calculating the pairwise distances of 16 PI and 17 ceramide structures. The first criterion to compare SMILES is to see non-zero distance values for all structurally different pair of molecules. All the three SMILES resulted in non-zero values with the four alignment-based methods

	Name	SMILES (before and after alignment)
Hydroxyl group is retained	16	CCC(O)CCCCCCCCCCCCCCC(=O)NC(CO)C(O)C=CCCCCCCCCCCC
	2	CCCCCCCCCCCCCCCCC(O)C(=O)NC(CO)C(O)C=CCCCCCCCCCCC
		↓
	16	CCC(O)CCCCCCCCCCCCCC---C(=O)NC(CO)C(O)C=CCCCCCCCCCCC
	2	CCC---CCCCCCCCCCCCCC(O)C(=O)NC(CO)C(O)C=CCCCCCCCCCCC
	15	CCCC(O)CCCCCCCCCCCCCC(=O)NC(CO)C(O)C=CCCCCCCCCCCC
	2	CCCCCCCCCCCCCCCCC(O)C(=O)NC(CO)C(O)C=CCCCCCCCCCCC
		↓
	15	CCCC(O)CCCCCCCCCCCCC---C(=O)NC(CO)C(O)C=CCCCCCCCCCCC
	2	CCCC---CCCCCCCCCCCCC(O)C(=O)NC(CO)C(O)C=CCCCCCCCCCCC
Hydroxyl group is NOT retained	17	CC(O)CCCCCCCCCCCCCCCC(=O)NC(CO)C(O)C=CCCCCCCCCCCC
	2	CCCCCCCCCCCCCCCCC(O)C(=O)NC(CO)C(O)C=CCCCCCCCCCCC
		↓
	17	CCCCCCCCCCCCCCC---C(=O)NC(CO)C(O)C=CCCCCCCCCCCC
	2	CCCCCCCCCCCCCCC(O)C(=O)NC(CO)C(O)C=CCCCCCCCCCCC
	18	C(O)CCCCCCCCCCCCCCCCC(=O)NC(CO)C(O)C=CCCCCCCCCCCC
	2	CCCCCCCCCCCCCCCCC(O)C(=O)NC(CO)C(O)C=CCCCCCCCCCCC
		↓
	18	CCCCCCCCCCCCCCCC---C(=O)NC(CO)C(O)C=CCCCCCCCCCCC
	2	CCCCCCCCCCCCCCCC(O)C(=O)NC(CO)C(O)C=CCCCCCCCCCCC

Figure 3.5: Comparison of four SMILES pairs, before and after Smith-Waterman alignment procedure. Ceramide structures in template SMILES representation were used. The arrow mark separates the pre- and post alignment SMILES for each pair. An horizontal line is drawn between {16-2, 15-2} and {17-2, 18-2}, to draw attention to the hydroxyl group (marked by a box frame) that is retained in the first set, but not retained in the later set. Each pair of SMILES is separated by a shorter horizontal line for clarity.

[Table 3.1] but they failed (resulted in at least one zero value) with LINGO and FP2 distances.

Table 3.1: Evaluation of SMILES writing approaches to distinguish non-identical ceramide and PI structures by measuring distance with 6 methods.

	LIN	FP2	Bio	SMI	S-W	Lev
CACTVS canonical SMILES	×	×	✓	✓	✓	✓
Open Babel canonical SMILES	×	×		✓	✓	✓
Template SMILES	×	×		✓	✓	✓

The statement "All non-identical structures have non-zero distance"is True ✓or False ×. LIN - LINGO, Bio - Bioisosterism, SMI - SMILIGN, S-W Smith-Waterman, Lev - Levenshtein

The second criterion is to have distance values that correlate with position of hydroxyl group. To test this, 17 ceramides structures were used. The two canonical SMILES representations resulted in inconsistent distance values for a linear (stepwise) consistent change in hydroxyl group position [Table 3.2]. Template SMILES resulted in inconsistent values with LINGO, FP2 and SMILIGN methods, and it is only partially consistent with Smith-Waterman and Levenshtein.

The final criterion is to have monotonically increasing distance values for a corresponding increase in acyl chain length. This criterion is relevant in the real-world application because the acyl chain length is modified by the elongase enzymes [148], which are involved in the biosynthesis of many lipid classes [149]. The two canonical SMILES resulted in inconsistent distances for all the six methods [Table 3.3]. LINGO and FP2 methods failed to provide consistent values for template SMILES, but the alignment based methods, SMILIGN, Smith-Waterman and Levenshtein distances resulted in monotonously increasing values for a corresponding acyl chain length

Table 3.2: Evaluation of SMILES writing approaches to distinguish hydroxyl group position by measuring distance with 6 methods.

	LIN	FP2	Bio	SMI	S-W	Lev
CACTVS canonical SMILES	×	×	×	×	×	×
Open Babel canonical SMILES	×	×		×	×	×
Template SMILES	×	×		×	(✓)	(✓)

The statement "Distance correlates with the change in hydroxyl group position" is True ✓, False × or Partially true (✓). Distance method abbreviations from previous table increase.

Table 3.3: Evaluation of SMILES writing approaches to distinguish the acyl chain length by measuring distance with 6 methods.

	LIN	FP2	Bio	SMI	S-W	Lev
CACTVS canonical SMILES	×	×	×	×	×	×
Open Babel canonical SMILES	×	×		×	×	×
Template SMILES	×	×		✓	✓	✓

The statement “Distance correlates with the acyl chain length”is True ✓ or False ×. Distance method abbreviations from previous table.

Additional experiments, 1. CACTVS canonical SMILES representation of 16 PI structures [Supl. Fig. 4.2], Open Babel canonical SMILES of 17 ceramides [Supl. Fig. 4.3] and template SMILES of 16 PI [Supl. Fig. 4.4] support the observation that template SMILES is better suited for measuring distances.

The template SMILES representation in combination with Levenshtein distance resulted in the most consistent values for both ceramides and phostphatidyl inositol structures. The template SMILES in combination with Smith-Waterman distance comes second in providing consistent values. In the next sections, these two methods will be tested with additional criteria.

3.2 Determining Coordinates from Distance Matrices with PCA

I have two objectives in this section 1. show that structure based comparison of lipids is valid in a two- and three- dimensional PCA space 2. asses the performance of template SMILES + Levenshtein distance on a larger set of lipids. In the first part of this section, I will compare the distribution of ceramides and phosphatidyl inositols in two- and three- dimensional PCA space. In the second part, I will analyze the distribution of 30 150 lipids in three dimensional PCA space.

3.2.1 Euclidean Distance in 3D PCA Space is Consistent with Levenshtein Distance

A PCA space for a set of lipid structures is generated by performing PCA on their distance matrices and plotting the components that have higher variance [section 2.5]. In the previous section, it was shown that pairwise Levenshtein and Smith-Waterman distances from the template SMILES of 17 ceramides correlated with change in hydroxyl group position [Table 3.3]. Following up with that result, the Levenshtein distances are converted to 2D and 3D coordinates by performing PCA [section 2.5].

3.2.1.1 Distribution of 17 Ceramides in 2D and 3D PCA Space

In the 2D PCA space, the distribution of ceramides show two features 1. they are sequential in arrangement and 2. they have an interesting cardioid curve like pattern [Fig. 3.6]. Sequential arrangement is a positive result because the position of the hydroxyl group in acyl chain is also sequentially changed [Fig. 2.1a]. The combination of 1. template SMILES 2. Levenshtein distance and 3. PCA, correctly portray the structure differences between ceramides in a 2D PCA space [Fig. 3.6].

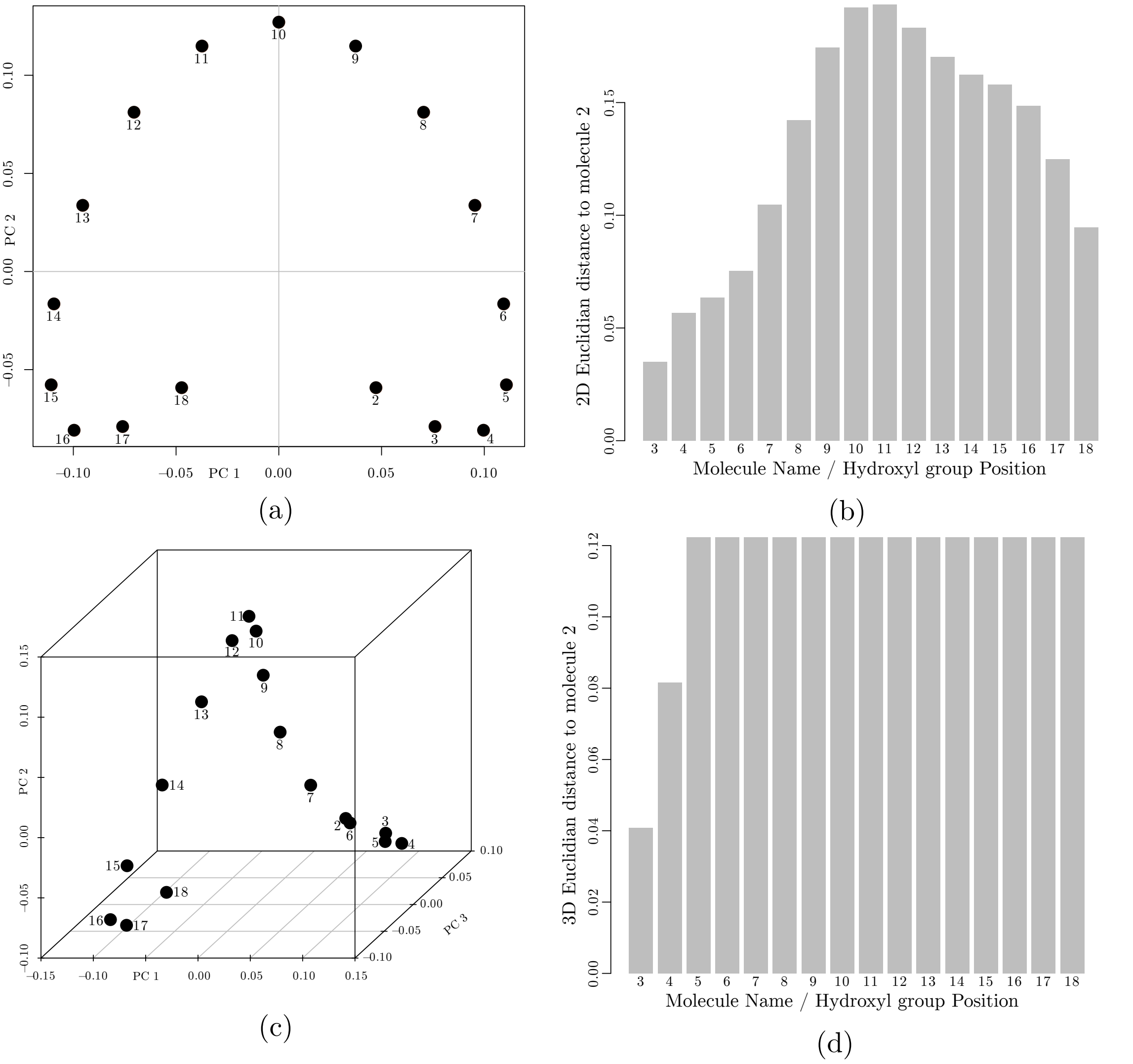
The cardoid-curve like pattern is analyzed by measuring the Euclidean distance between molecule 2 and {3--18} (calculated from PC1 and PC2 coordinates), and generating an bar graph [Fig. 3.6b]. It shows a slight left-skewed distribution, although, the expectation is to see more left-skewness because of the sequential movement of hydroxyl group in the acyl chain [Fig. 2.1a]. Reason for limited left-skewness is that PC1 and PC2 cover only ~55% of the total variance which means the 2D view limits the visualization of true inter-molecular Levenshtein distance. The distribution of ceramides in 3D PCA space shows better separation of molecules 2 and 18 in the z plane, which was not visible in 2D view [Fig. 3.6c]. A bar graph of the Euclidean distances calculated from 3D coordinates follow the pattern of Levenshtein distance [Fig. 3.6d]. The three principal components cover ~67% of the total variance.

In contrast, the 2D PCA space generated from the Smith-Waterman distances do not show sequential arrangement [Suppl. Fig. 4.5]. The bar graph of Euclidean distance has two peaks (expectation is to see linear increase with a sigle peak at right-end) and is not consistent with the sequential change in hydroxyl group position [Fig. 2.1a]. Smith-Waterman distances for the ceramide-pairs associated with molecule 17 were peculiar which could be the reason for two peaks [section 3.1.3].

3.2.1.2 Distribution of 16 PI's in 2D and 3D PCA Space

The 2D PCA space of 16 PI's has two features 1. arrangement of PI's has an inverted U shape 2. coordinate points have a sequential arrangement in the right to left direction [Fig. 3.7]. PI's with saturated acyl chains alternate with unsaturated acyl chain molecules, while maintaining the acyl chain length sequence. The PI's with an two carbon, ethanediyl ($-CH_2 - CH_2-$) difference in chain length were farther separated, in comparison to methanediyl ($-CH_2-$) difference.

The 3D coordinate view shows a better separation of unsaturated molecules in the third principal component axis [Fig. 3.7]. The bar graph of Euclidean distance between molecule 10 and {10*--20*} is a left-skewed distribution, which is a positive result because the Euclidean distances correctly increase with the acyl chain length. The 2D, 3D and the bar graph of Euclidean distances generated from pairwise Smith-Waterman distance matrix closely resemble the results from Levenshtein distances [Suppl. Fig. 4.6]. Molecules {10*, 11} have identical Euclidean distance with molecule 10, although, they are structurally different.



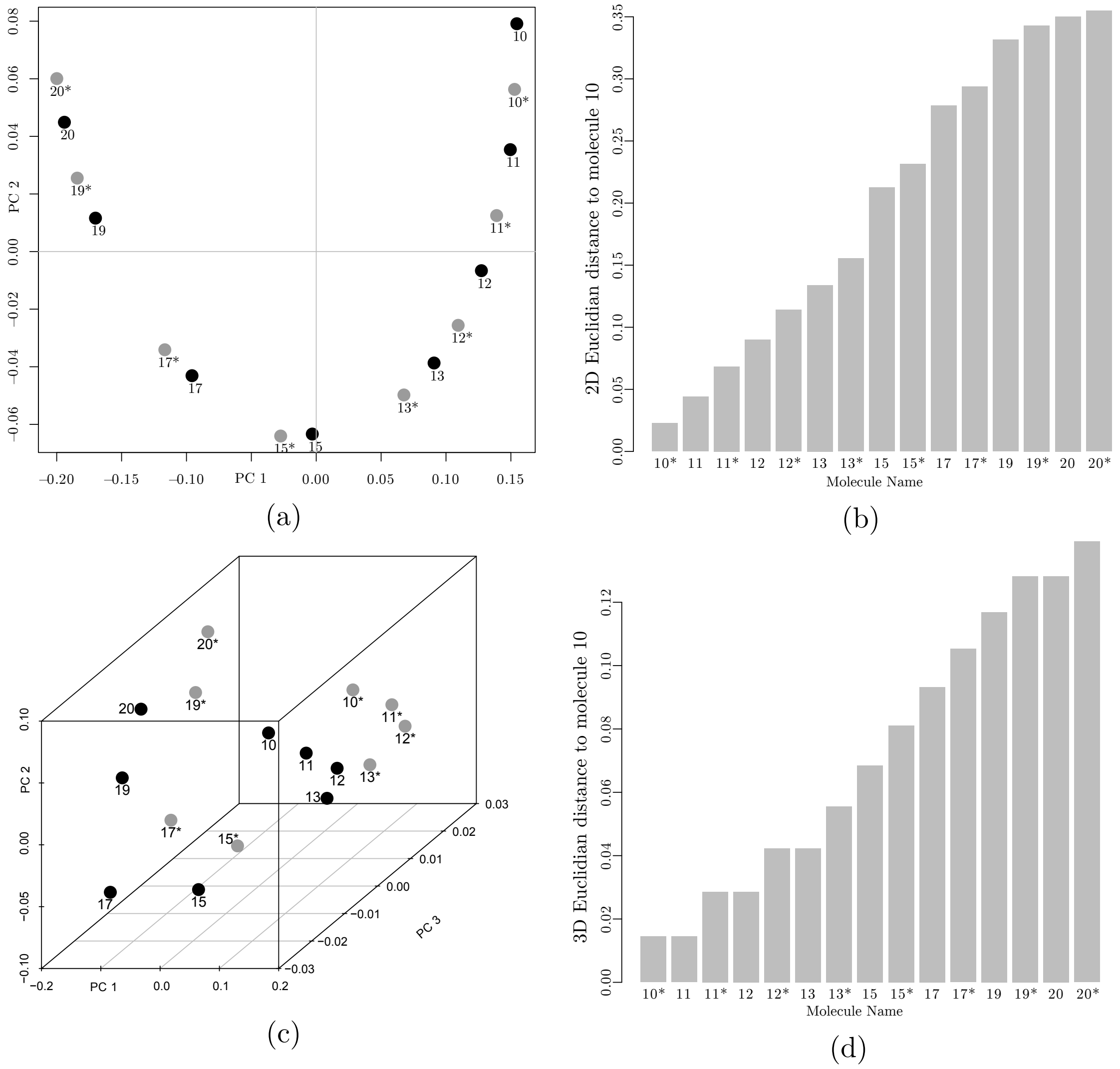
(a)

(b)

(c)

(d)

Figure 3.6: Distribution of 17 Ceramides in 2D and 3D PCA space. Pairwise Levenshtein distances were converted to coordinates with PCA. (a) Prinipcal component 1 (PC1) in x axis and 2 (PC2) in y axis. (b) Bar graph of Euclidean distances between molecule 2 to {3--17}, calculated from PC1 and PC2 coordinates (c) Principal components 1, 2 and 3 in x, y, and z axes. (d) Bar graph of Euclidean distances between molecule 2 to {3--17}, calculated from PC1, PC2 and PC3 coordinates.



(a)

(b)

(c)

(d)

Figure 3.7: Distribution of 16 PI molecules in 2D and 3D PCA space. Pairwise Levenshtein distances were converted to coordinates with PCA. (a) Prinipcal component 1 (PC1) in x axis and 2 (PC2) in y axis. (b) Bar graph of Euclidean distances between molecule 10 to {10*--20*}, calculated from PC1 and PC2 coordinates (c) Principal components 1, 2 and 3 in x, y, and z axes. (d) Bar graph of Euclidean distances between molecule 10 to {10*--20*}, calculated from PC1, PC2 and PC3 coordinates. PI's with unsaturated acyl chains were shown in gray in plots (a) and (c).

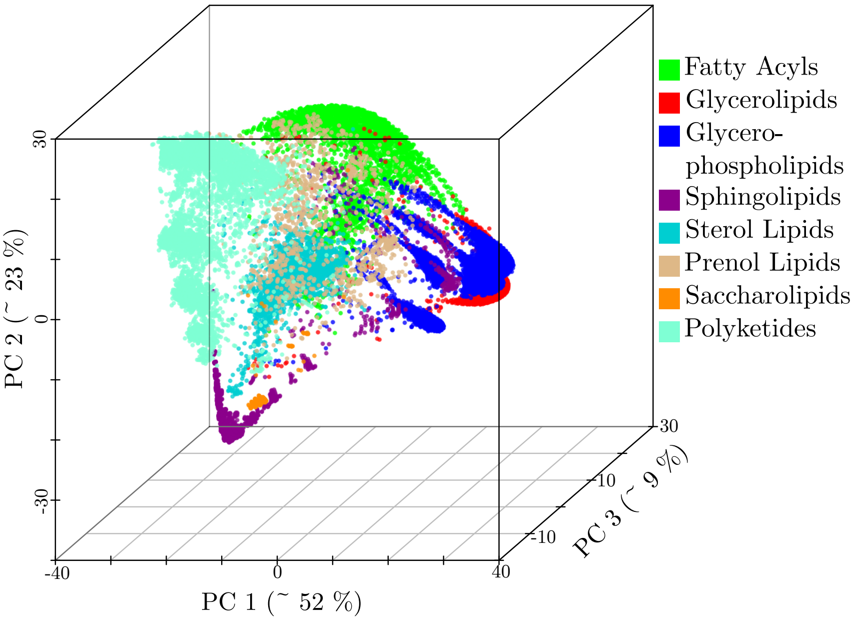


Figure 3.8: 8 lipid categories form noticeable clusters in 3D PCA space. All lipids from LMSD were converted to template SMILES. PCA coordinates were calculated from the pairwise Levenshtein distance matrix. Lipid classification was described earlier [subsection 2.1.2]. The *x, y, z* axes are principal components 1, 2 and 3.

3.2.2 Arrangement of 30 150 structures in 3D PCA Space is Consistent with the Lipid Classification

3D PCA space from the Levenshtein distances has separated lipid classes to noticeable clusters [Fig. 3.8]. The top three principal components capture approximately 84% of the total variance in the dataset.

The sphingolipids category in LMSD contains 3510 structures. The arrangement of sphingolipids in PCA space resembles a pear-shaped curve [Fig. 3.9]. The sphingoid-bases class is a collection of sub-structures with reduced structure overlap [8]. The diverse sphingoid-bases class clustered farther from the highly similar structures in glycosphingolipids sub-class [Fig. 3.9]. Molecules in ceramides sub-class are closely related to sphingoid-bases but with varying acyl chains and no sugar substitutions which is correctly reflected in the PCA space because ceramides sub-class clustered nearer to sphingoid bases and farther to complex glycosphingolipids [Fig. 3.9]. Neutral and acidic glycosphingolipid sub-classes cluster together, because of the identical sub-structures.

The distribution of molecules in the 3D PCA space coincides with the sugar

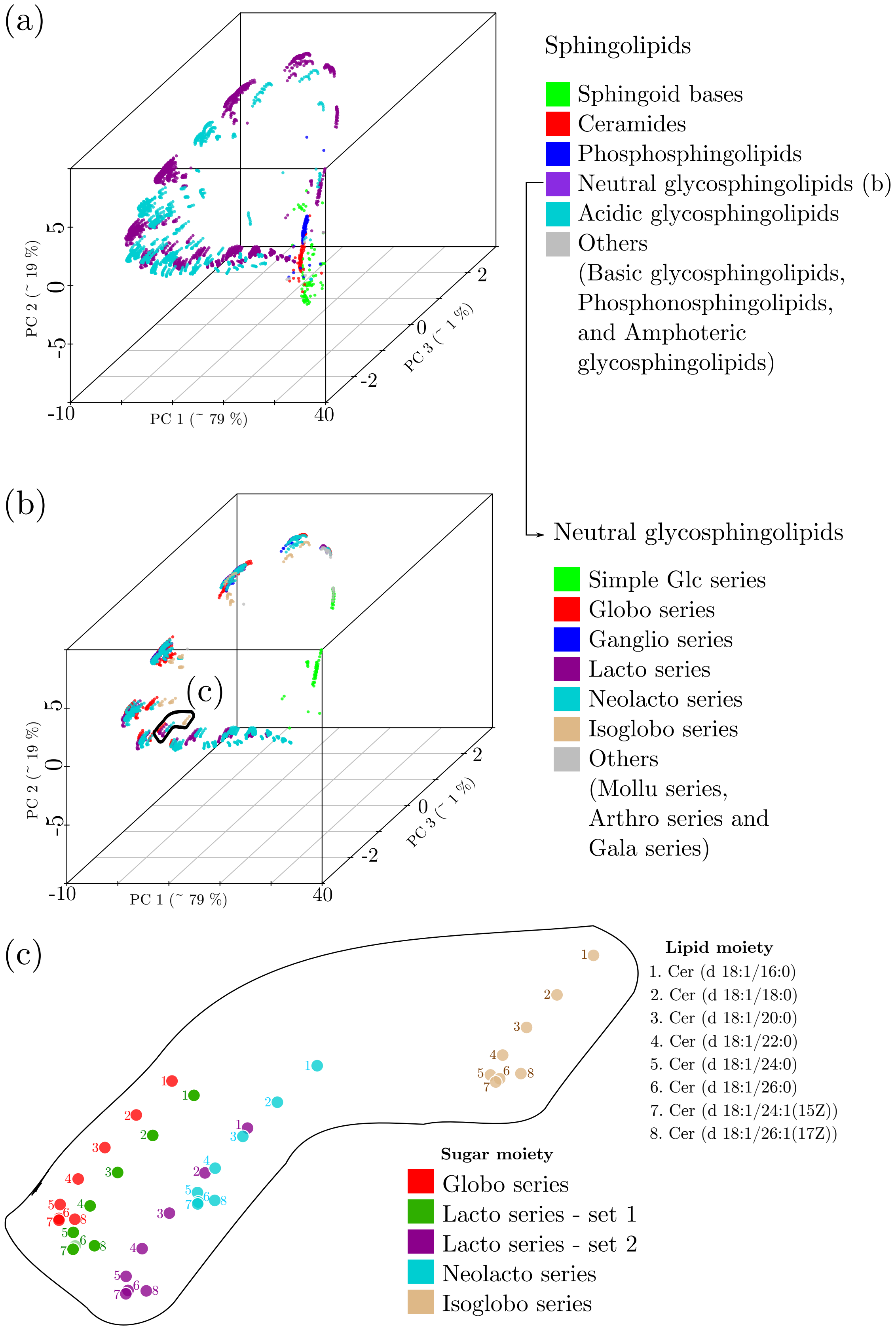
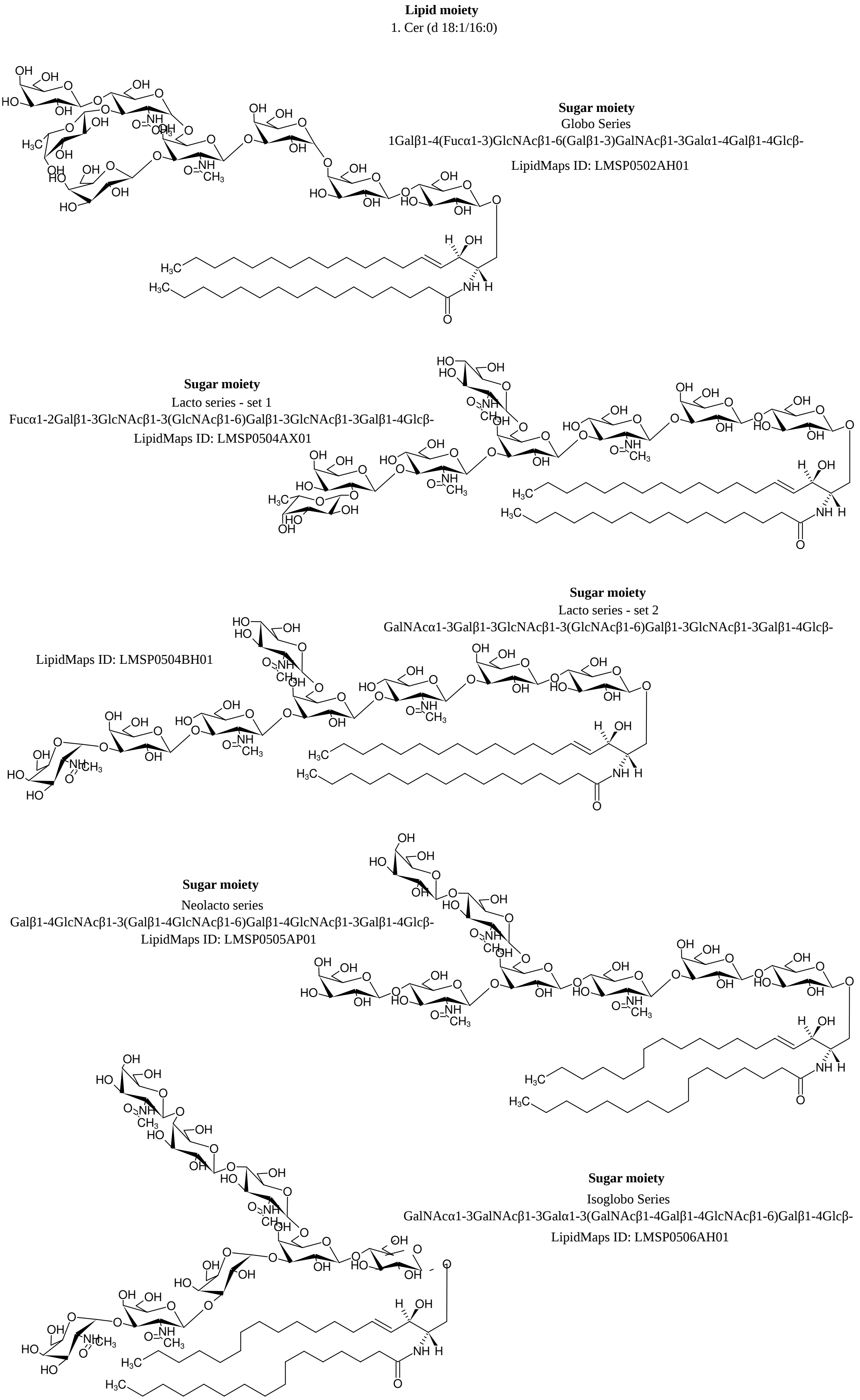


Figure 3.9: Spatial distribution LMSD Sphingolipids. PCA coordinates of 3510 Sphingolipids were calculated from the pairwise Levenshtein distances of the template SMILES representation of structures. Lipid classification and sugar moiety nomenclature from LIPID MAPS.



Lipid moiety

1. Cer (d 18:1/16:0)

Sugar moiety

Globo Series

1Galβ1-4(Fucα1-3)GlcNAcβ1-6(Galβ1-3)GalNAcβ1-3Galα1-4Galβ1-4Glcβ-

LipidMaps ID: LMSP0502AH01

Sugar moiety

Lacto series - set 1

Fucα1-2Galβ1-3GlcNAcβ1-3(GlcNAcβ1-6)Galβ1-3GlcNAcβ1-3Galβ1-4Glcβ-

LipidMaps ID: LMSP0504AX01

Sugar moiety

Lacto series - set 2

GalNAcα1-3Galβ1-3GlcNAcβ1-3(GlcNAcβ1-6)Galβ1-3GlcNAcβ1-3Galβ1-4Glcβ-

LipidMaps ID: LMSP0504BH01

Sugar moiety

Neolacto series

Galβ1-4GlcNAcβ1-3(Galβ1-4GlcNAcβ1-6)Galβ1-4GlcNAcβ1-3Galβ1-4Glcβ-

LipidMaps ID: LMSP0505AP01

Sugar moiety

Isoglobo Series

GalNAcα1-3GalNAcβ1-3Galα1-3(GalNAcβ1-4Galβ1-4GlcNAcβ1-6)Galβ1-4Glcβ-

LipidMaps ID: LMSP0506AH01

Figure 3.10: Selected sphingolipid structures for panel (c) in Fig. 3.9

moiety and lipid acyl chain [panel (c) Fig. 3.9]. In the set of lipids with the same sugar moiety, the arrangement in 3D PCA space conforms with the lipid moiety. The set of lipids with the same sugar moiety form an twisted 'L' shape formation. The molecules with longer acyl chain {4--8} cluster together, farther from the molecules with shorter acyl chain {1, 2, 3}.

The five highlighted clusters in the neutral glycosphingolipid region have the same sum sugar composition, but vary in the arrangement of sugars [Fig. 3.10]. Four clusters (Globo, Lacto 1, Lacto 2 and Neo-lacto) branch out at the fourth sugar, and the fifth cluster (Iso-globo) branches out at the second sugar [Fig. 3.10]. The spatial arrangement of highlighted clusters follows the sugar moiety branching pattern, i.e, four clusters are nearer and the fifth is farther [panel (c) Fig. 3.9].

3.2.3 The Background Ensemble has Low Effect on the Spatial Arrangement in PCA Space

A set of 14 Phosphatidyl Choline (PC) molecules with one varying acyl chain, length {12--26}, and one uniform acyl chain, length {12}, are projected in PCA space, under two conditions 1. in the presence and 2. in the absence of background LMSD ensemble [Fig. 3.11]. The distribution in the presence of LMSD is an twisted 'L' shape in the 2D PCA space. The bar graph of 2D Euclidean distance between PC molecule 12 and {12--26} is an left-skewed distribution, as expected, because of the sequential increase in the acyl chain length [Fig. 3.11]. The distribution of PC molecules in the absence of LMSD is in the shape of a parabola in 2D PCA space [Fig. 3.11]. The parabola shaped arrangement has sequential order (from left to right), coinciding with their acyl chain length. The 2D Euclidean distance derived from parabola is left-skewed, which is a positive result because of the gradual increase in acyl chain length {12--26} [Fig. 3.11]. The comparison of the two bar graphs (presence and absence of LMSD) indicates that the background ensemble has minimal effect on the 2D Euclidean distances [panels c, d in Fig. 3.11].

The arrangement of nine Cholesterol Esters (CE) in the LMSD background ensemble is consistent with the acyl chain length [Fig. 3.12]. Molecules with the odd versus the even number of carbon atoms in the acyl chain are clearly separated. Similarly, the arrangement of ten triacylglycerol (TAG/TG) structures is also consistent with acyl chain length and unsaturation [Fig. 3.12]. Five unsaturated TAG's are distinct from the saturated molecules [Fig. 3.12].

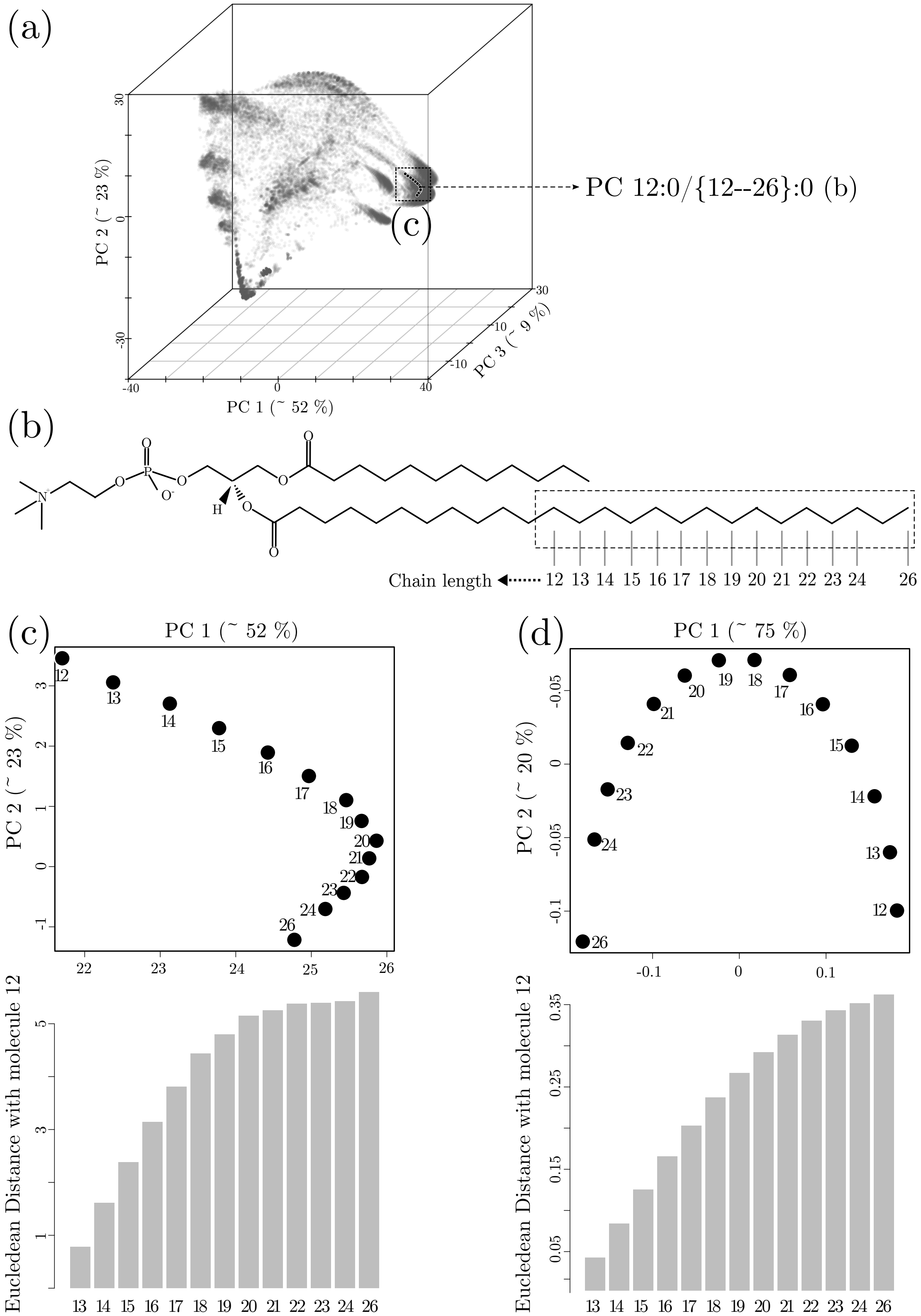


Figure 3.11: Spatial distribution 14 PC molecules varying in acyl chain length (a, b). PCA of LMSD but only 14 PC molecules shown (c) PCA of 14 PC only, i.e, no LMSD background ensemble (d).

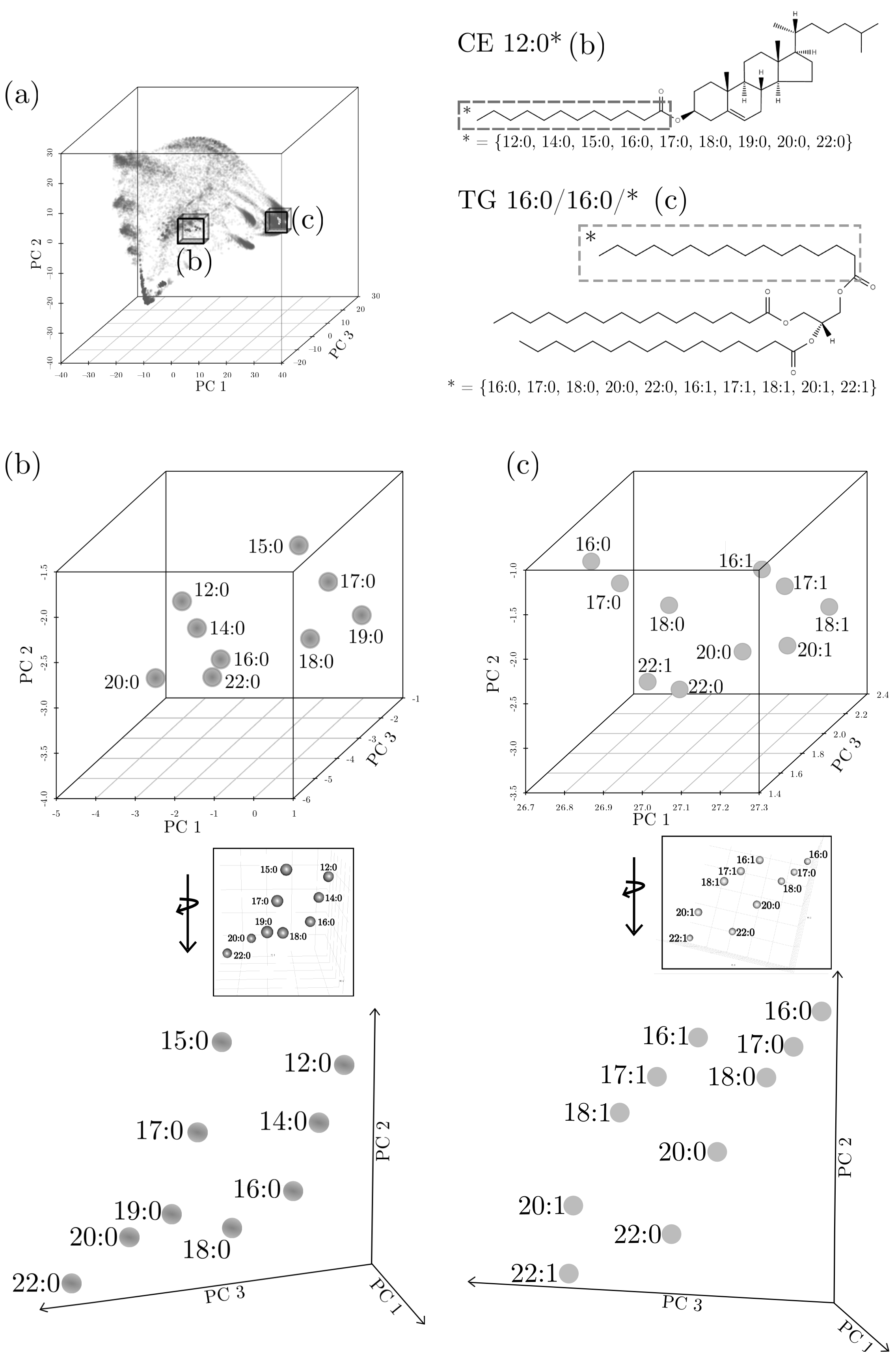


Figure 3.12: Distribution of 9 Cholesterol Esters (CE) (b) and 9 TAGs (c) varying in acyl chain length, in the background of LMSD ensemble. First shown in the LMSD background (a) and later shown (b,c) separately.

3.3 Lipidome Juxtaposition

In this section, I will apply the methods developed in earlier sections (template SMILES, Levenshtein distance and PCA Space) to compare lipidomes of yeast, fruit fly and human lung [section 2.2]. Later, I will apply LUX Score [section 2.6] for hierarchical clustering of lipidomes. I will end the chapter with a structure-based lipidome comparion workflow that will give an overview of all the computational methods developed in this study.

3.3.1 Hierarchical Clustering of Yeast Lipidomes with LUX score concurs with Phenotype

The eight yeast elongase gene mutant lipidomes published by Ejsing et. al [23] are summarized in Table 3.4. Lipids pooled from eight lipidomes, i.e., two each from four yeast strains, are projected in two dimensional PCA space [Fig. 3.13]. The differences between the mutant strains could be visualized by plotting lipidomes separately [Fig. 3.14].

Table 3.4: Overview of yeast elongase mutant lipidomes

		Phenotype		No. of Lipids	
Strain	Gene function	Bud Morphology	Budding	24°C	37°C
BY4741		Normal	Normal	176	145
Elo1	FA elongase	–	–	176	159
Elo2	FA elongase	Abnormal	Abnormal	161	163
Elo3	FA elongase	–	Abnormal	174	170

Brachmann *et al.* [150–153]

The number of TAGs are reduced in all four strains at 37°C, but it is not noticeable in PCA space because the region in Q2-Q3 also contains PC and PE [Fig. 3.14]. The lower Levenshtein distance groups TAGs with phospholipids, which suggests that yeast produces many lipids with overlapping structural features. The IPC in Q1 region of PCA space contains 5--8 molecules for BY4741 and Elo1 strains, but 12--15 for Elo2 and Elo3. IPC structures do not overlap with other lipids which indicates a unique feature of yeast elongase lipidomes [Fig. 3.14]. The number of PS molecules are similar in all mutants but there is more diversity in structures for

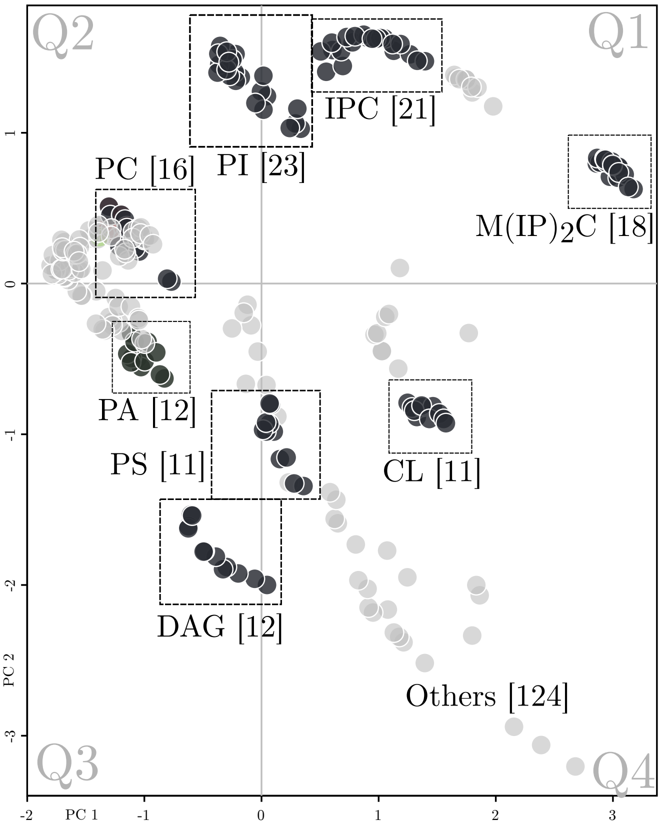


Figure 3.13: Map of 248 lipids pooled from 8 yeast lipidomes [Table 3.4]. Pairwise Levenshtein distances were calculated and the first two pricipal components of the matrix were plotted in x and y axis respectively. Lipids classes that formed distinguishable clusteres were dark shaded. The number of lipid species in each class were shown in brackets. MIPC, M(IP)$_2$C and IPC are located in Q1. A subset of PI, with shorter acyl chains are located near to IPC in Q1. Q2 is composed of TAG and PC. Other phospholipids, such as PA and PE clustered together with TAGs, but in Q3. DAGs are located in Q3. Ceramides, PS and CL are in Q4. MIPC - Mannose-inositolphospho-ceramide(s); M(IP)$_2$C - Mannose-bis(inositolphospho)ceramide(s); IPC - Inositol phosphorylceramide(s); TAG - Triacylglycerol(s); PC - Phosphatidylcholine(s); PA - Phosphatidic acid(s); PE - Phosphatidylethanolamine(s); DAG - Diacylglycerol(s); PS - Phosphatidylserine(s); CL - Cardioplipin(s).

lipidomes at 37°C, which indicates a feature specific to higher growth temperature [Fig. 3.14].

The lipid structure differences between yeast lipidomes were collected to generate pairwise LUX scores [[Equation 2.6], Fig. 3.15a]. 93% of the lipids are common in the control strain BY4741 and Elo1, cultured at 24°C [Fig. 2.3], which is also reflected in the lowest LUX score, 0.003 [Fig. 3.15a]. The most dissimilar lipidomes are Elo1 37°C and Elo3 24°C but the distance between them is only 0.019 (theoretical maximum is 1), which is an indication that the yeast elongase lipids have high structure similarity [Fig. 3.15a].

Elo1, Elo2 and Elo3 have mutation in the fatty acid elongase gene of the sphingolipid metabolism pathway but Elo1 has no reported effect on the yeast bud morphology, while Elo2 and Elo3 caused visible changes [Table 3.4]. Hierarchical clustering based on LUX score groups BY4741 and Elo1 together, Elo2 and Elo3 together (abnormal budding) concurring with their biological phenotype [Fig. 3.15b].

3.3.1.1 Error modeling of yeast lipidome

The aim of error modeling experiment is to test robustness of the tree generated by hierarchical clustering of the LUX scores [section 2.8]. What happens if we exclude low abundant lipids from analysis ? Does it perturb clustering? If yes, to what extent?

With the threshold of detection limited to 0.003 mol% and standard deviation of 0.001, there is no change in the tree branching [Fig. 3.15b]. Threshold of detection limited to 0.003 mol% and standard deviation of 0.002, caused only minor changes and the maximum change is effected by increasing the detection limit further. Clustering frequency at the top level branch comprising BY4741 and Elo1 reduced by 15%, the branch with Elo2 and Elo3 reduced by 21%. More lipids were exluded from the LUX score at $t_{detect} = 0.003\ mol\%$ and $\sigma = 0.002$, but the braching pattern is not effected, especially at the lower level (each branch reoccurs 97, 77, 92 and 99 times of out 100), indicating that the hierarchical clustering is robust.

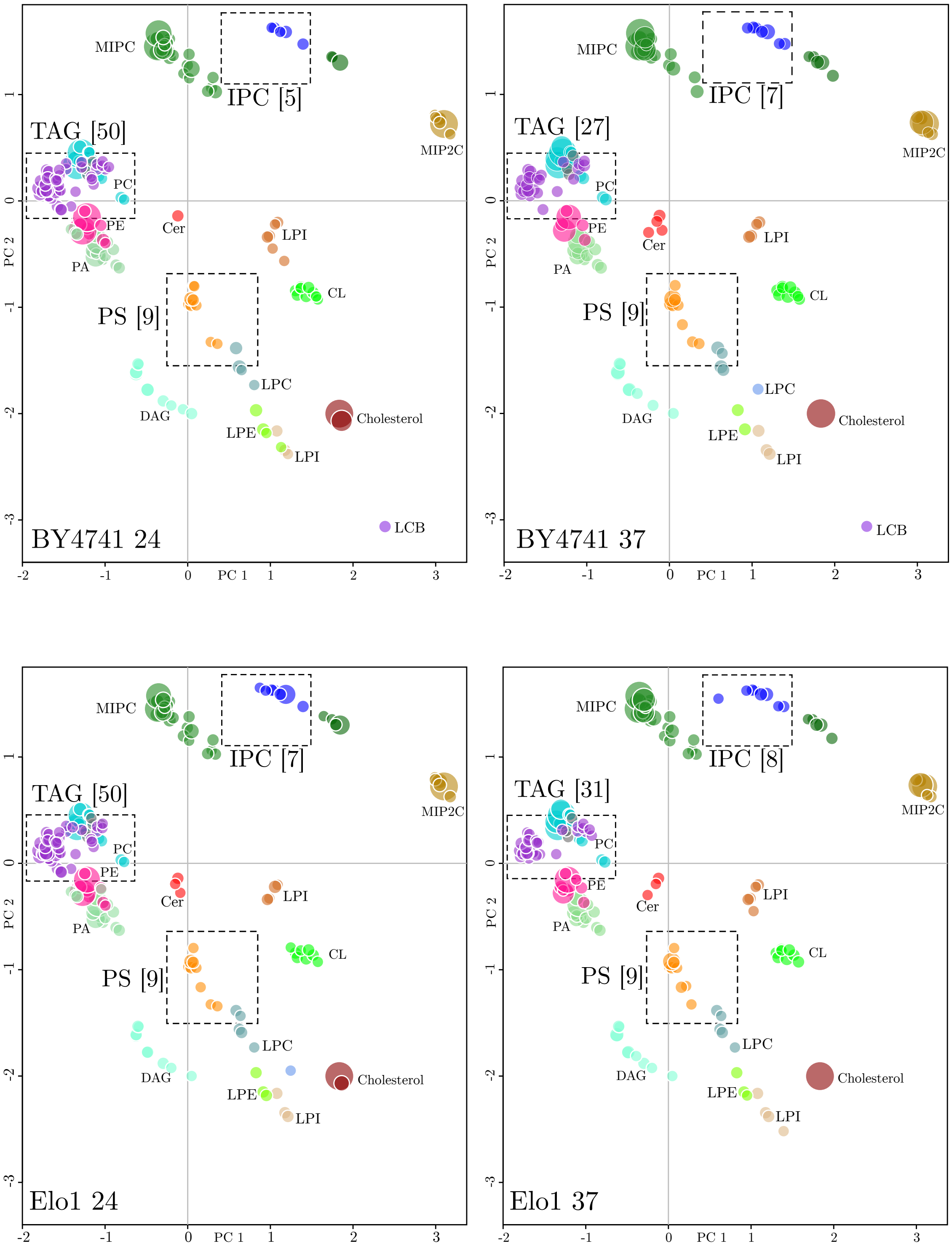


Figure 3.14: Distribution of lipids in 8 yeast strains. Dot size for each lipid is scaled to the concentration. Number of molecules for selected lipid class indicated inside square brackets. MIPC - Mannose-inositolphospho-ceramide(s); $M(IP)_2C$ - Mannose-bis(inositolphospho)ceramide(s); IPC - Inositol phosphorylceramide(s); Cer - Ceramide(s).

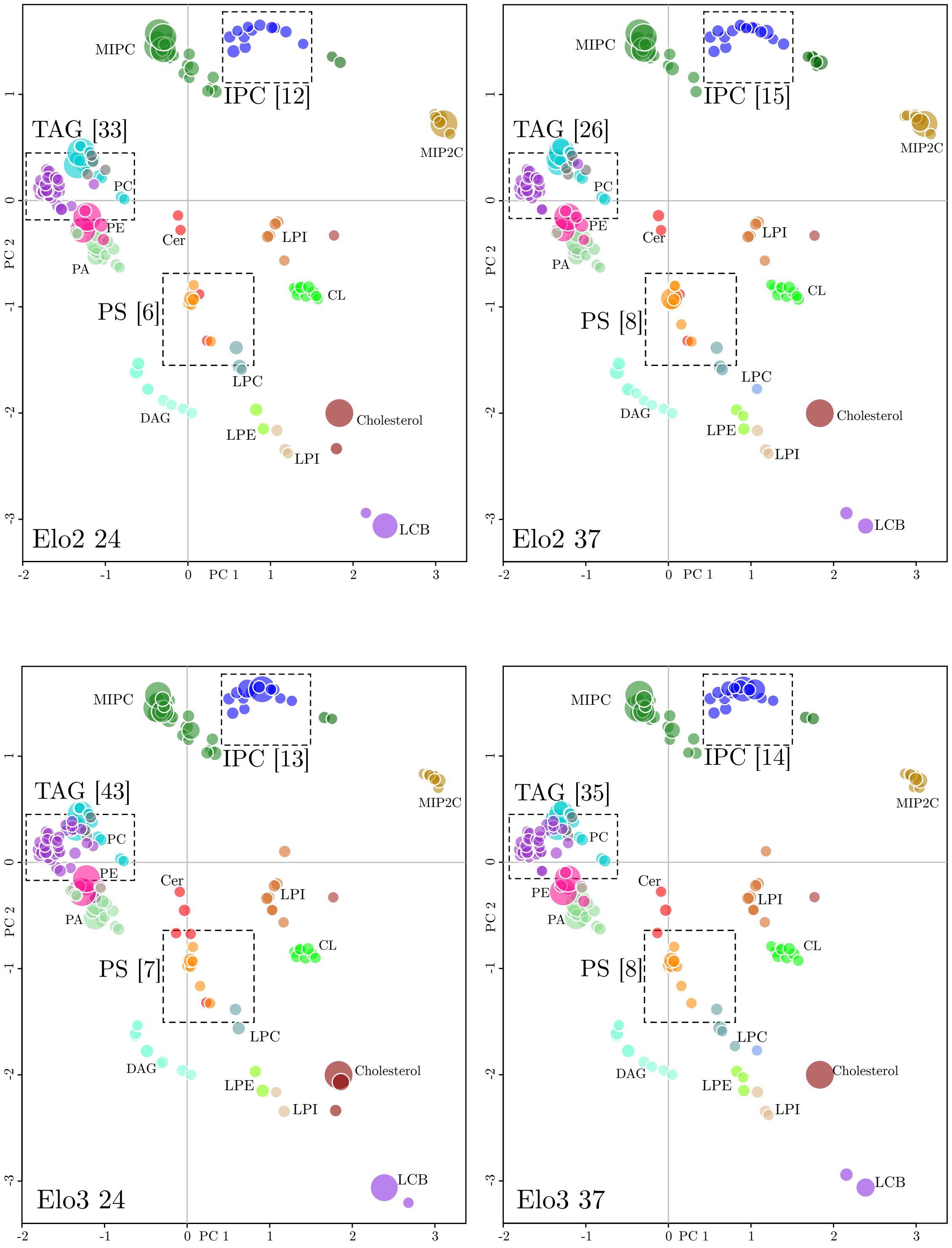
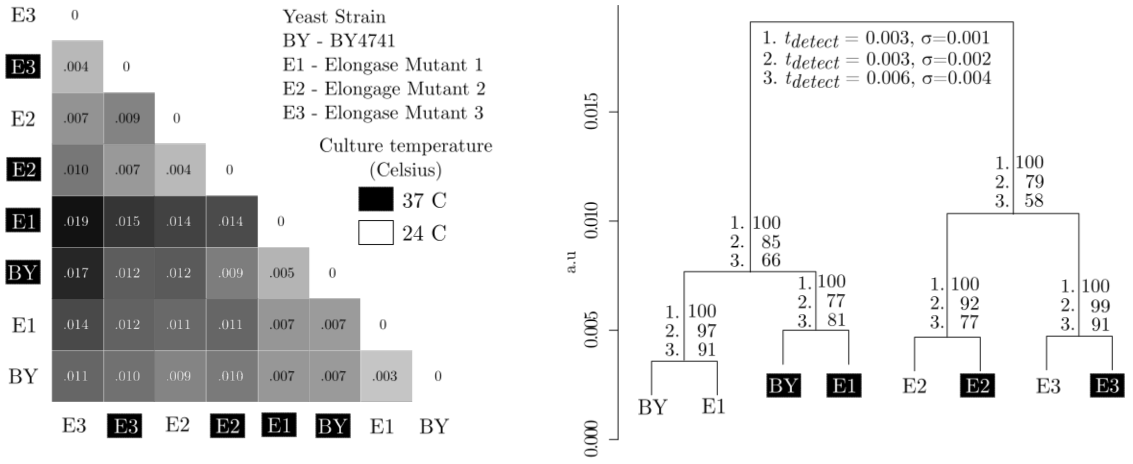


Figure 3.14 continued. TAG - Triacylglycerol(s); PC - Phosphatidylcholine(s); PA - Phosphatidic acid(s); PE - Phosphatidylethanolamine(s); DAG - Diacylglycerol(s); PS - Phosphatidylserine(s); CL - Cardioplipin(s); PI - Phoshatidyl Inositol(s); PS - Phosphatidyl Serine(s); L[PI][PE][PC] - Lyso [PI][PC][PC]; LCB - Long Chain Base(s).



(a) LUX scores

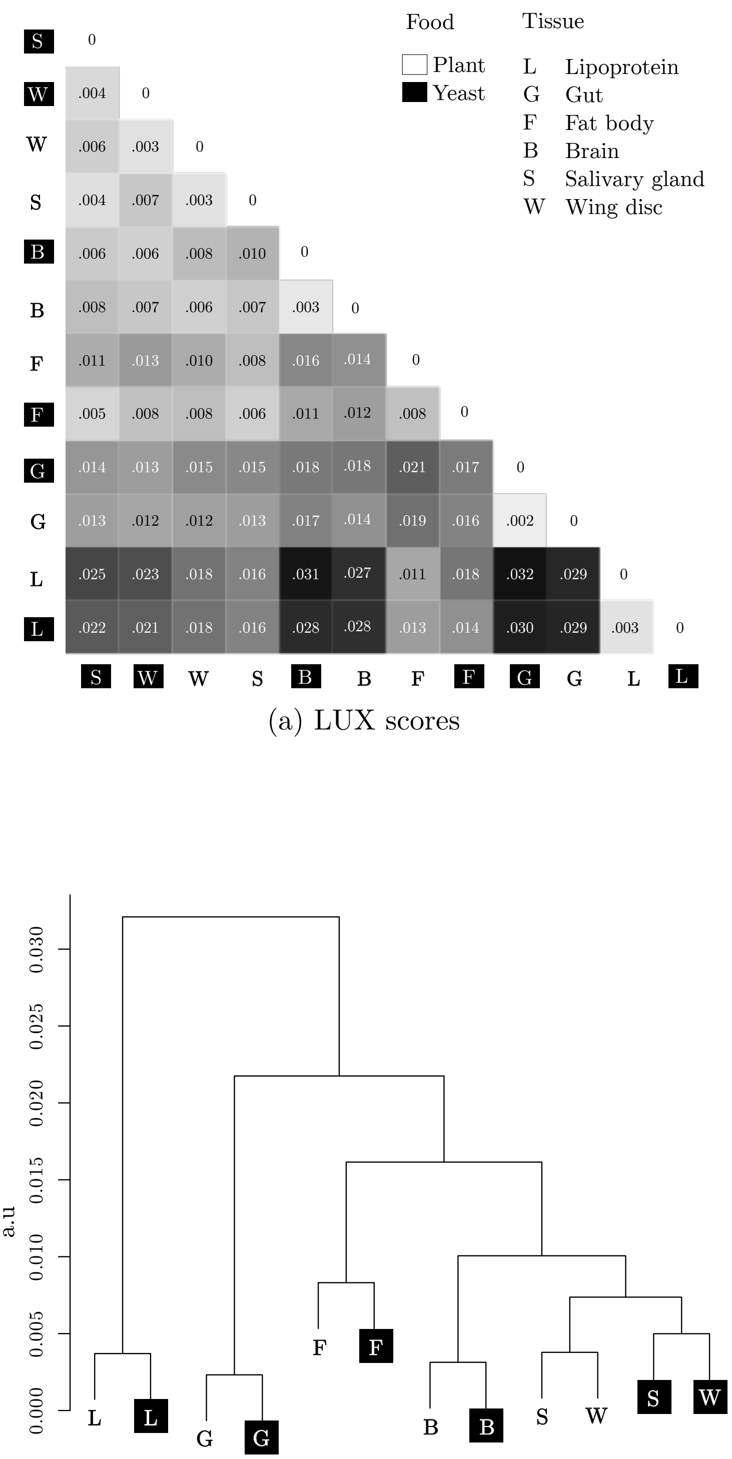
(b) Hierarchical clustering

Figure 3.15: Pairwise LUX Scores for 8 yeast lipidomes (a). Higher values are dark shaded. Clustering of lipidomes based on LUX Scores (b). Branch frequency [section 2.8] displayed at tree nodes.

3.3.2 Juxtaposition of Fruit fly Larva Tissue Lipidomes

12 lipidomes from *Drosophila melanogaster* larvae [section 2.2.3] were used to calculate LUX scores [section 2.6]. The LUX scores are in the range of 0.002 and 0.003, indicating that overall, the lipidomes are highly similar [Fig. 3.16a]. However, the values also point to the subtle differences between the tissues. For instance, the highest value is between brain and salivary gland. The lowest value is between lipoprotein lipidomes collected after feeding the larvae with two diets, indicating that food regime has less effect on the lipid structure diversity of lipoprotein tissue.

The lowest level branches of the hierarchical clustering of LUX scores put lipidomes of the same tissue type together [Fig. 3.16b]. In the higher level branching, brain, salivary gland and wing disc lipidomes cluster together. The fat body, gut and lipoprotein lipidomes are farther in that order. Brain, salivary gland and wing disc are important for specific functions and expected to accumulate specific lipids. In contrast, the gut, lipoprotein, fat body tissues are important for collection and storage of food, which are expected to accumulate a wider variety of lipids. The hierarchical clustering of lipidomes based on LUX scores correctly seperates these two sets of tissues. [Fig. 3.16b].



	■S	■W	W	S	■B	B	F	■F	■G	G	L	■L
■S	0											
■W	.004	0										
W	.006	.003	0									
S	.004	.007	.003	0								
■B	.006	.006	.008	.010	0							
B	.008	.007	.006	.007	.003	0						
F	.011	.013	.010	.008	.016	.014	0					
■F	.005	.008	.008	.006	.011	.012	.008	0				
■G	.014	.013	.015	.015	.018	.018	.021	.017	0			
G	.013	.012	.012	.013	.017	.014	.019	.016	.002	0		
L	.025	.023	.018	.016	.031	.027	.011	.018	.032	.029	0	
■L	.022	.021	.018	.016	.028	.028	.013	.014	.030	.029	.003	0

(a) LUX scores

(b) Hierarchical clustering

Figure 3.16: Pairwise LUX scores for 12 fruit fly lipidomes (a). Higher values are dark shaded. Lipoprotein lipidomes have higher LUX scores. The plant diet salivary gland lipidome has lower LUX scores, followed by fat body lipidome under plant based diet. Clustering of 12 lipidomes based on LUX Scores (b). The diet plays an important role defining the lipid diversity for all tissues except salivary gland and wing disc.

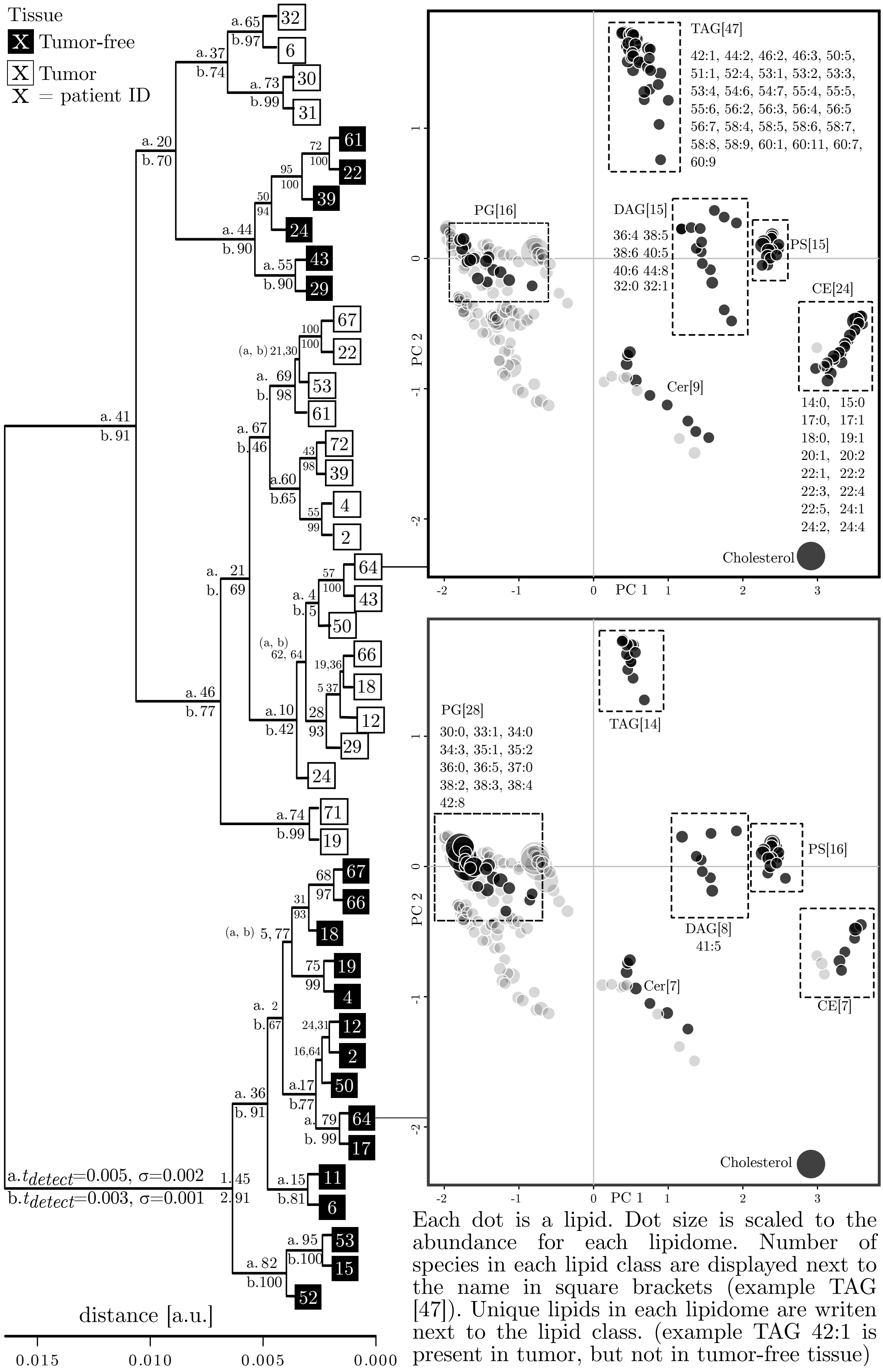
3.3.3 LUX score Separates Tumor from Tumor-free Human Lung Tissue

Lipidomes of 21 tissue samples from alveolar region of lung (non-cancerous/tumor-free) and 23 samples from tumor tissue, were obtained from 26 human lung cancer patients in a recent study [section 2.2.4]. Hierarchical clustering of 43 tissues based on LUX score clearly separated tumor from tumor-free tissue [Fig. 3.17]. 4 distinct clusters emerged from the hierarchical clustering – two corresponding to alveolar tissues and the other two to the tumor tissues. Tumor-free tissues with higher inflammation score (ID22, ID24, ID29, ID43) are clustered with tumor tissues. Error modeling is performed to test the robustness of clustering to the change in lipid detection threshold. In spite of the stringent threshold, LUX score based hierarchical clustering separates tumor from tumor-free tissues on most occasions [Fig. 3.17].

The LUX score between tumor tissues ID43 and ID64 is 0.0014. It is the lowest LUX score in the dataset reflecting the similar lipid structure composition. 239 (94%) lipids in ID43 are present also in ID64. Only 14 lipids in ID43 have no identical counterpart in ID64. But all 14 structures have close neighbors in ID64 [Table 3.5]. The highest LUX score in the dataset is between tumor tissue ID19 and alveolar tissue (tumor-free) ID15 is 0.0165. 122 (49%) lipids in ID19 have no identical counterpart in ID15.

The LUX score between tumor and tumor-free tissue of ID64 is 0.005. This patient is chosen as an example to highlight the lipids that present in tumor tissue but absent in tumor-free and vice-e-versa. TAG molecules present in tumor tissue but absent in tumor-free are marked in the lipidome map and easily visualized by three-fold increase the number of the data points in Q1 [Fig. 3.17]. Tumor-free tissue has no unique TAG molecules. The new TAG molecules of tumor tissue are located farther from the lipids of same class which reflects a lipid strucutral difference between the tumor and tumor-free tissues [Fig. 3.17].

The PCA space of DAG molecules has more data points in tumor tissue. Short acyl chain DAG molecules (32:0 and 32:1) are located at the bottom, slightly farther from other lipids of the same class. The bottom-right quadrant is marked by



Each dot is a lipid. Dot size is scaled to the abundance for each lipidome. Number of species in each lipid class are displayed next to the name in square brackets (example TAG [47]). Unique lipids in each lipidome are writen next to the lipid class. (example TAG 42:1 is present in tumor, but not in tumor-free tissue)

Figure 3.17: Hierarchical clustering of human lung lipidomes with LUX score. Branch frequency [section 2.8] shown at tree node. ID64 tumor and tumor-free lipidomes were displayed for comparison. Lipid classes with significant differences highlighted.

Table 3.5: Comparison of selected lipids in two tumor tissues

Unique lipids in ID43	Close structure counterpart(s) in ID64		
Cer 39:1;2	Cer 38:1;2	Cer 40:1;2	
DAG 35:1	DAG 34:1	DAG 36:1	
DAG 39:5	DAG 38:5	DAG 40:5	
DAG 40:4	DAG 38:4	DAG 40:5	DAG 40:6
DAG 41:5	DAG 40:5		
HexCer 42:2;3	HexCer 42:2;2		
LPC 18:1	LPC 16:0		
PC 31:1	PC 30:1	PC 31:0	
PC 38:3	PC 38:2		
PG 35:1	PG 35:0	PG 36:1	
PG 36:0	PG 35:0	PG 36:1	
PG 38:2	PG 38:5	PG 38:6	PG 39:0
PI 34:0	PI 34:1		
PI 36:3	PI 36:2	PI 36:4	

a three times increase in the point cloud made of CE lipids in tumor tissue. The coordinates of new CE molecules in tumor tissue are continuous . Except PG class, all phospholipids share similar PCA space in tumor and tumor-free tissues. However, 13 PG molecules present in tumor-free tissue are absent in tumor tissue [Fig. 3.17].

3.3.4 Lipidome Juxtaposition Work flow

The computational methods developed in this study, for the comparing lipidomes are summarized as a work flow [Fig. 3.18]. First, the lipid names from multiple experiments/samples are pooled to create a combined list. The duplicate entries are removed from the list, which is important, especially for lipidome datasets of a single tissue but different time points will yield lipid names that will be obviously listed in all samples. In contrast, for instance, lipidomes of different organisms will likely have less number of duplicates [section 2.2].

Next, the structure drawing and SMILES generation is performed. The number of isomeric structures that will be generated for a given lipid species depend on the structural information gathered with mass spectrometry. For example, fatty acid composition of the yeast lipids was available in one study [23], but that information could not be obtained for fruit fly lipidome [121]. In this case, more isomeric possibilities were enumerated for fruit fly lipids because of the limited fatty acyl

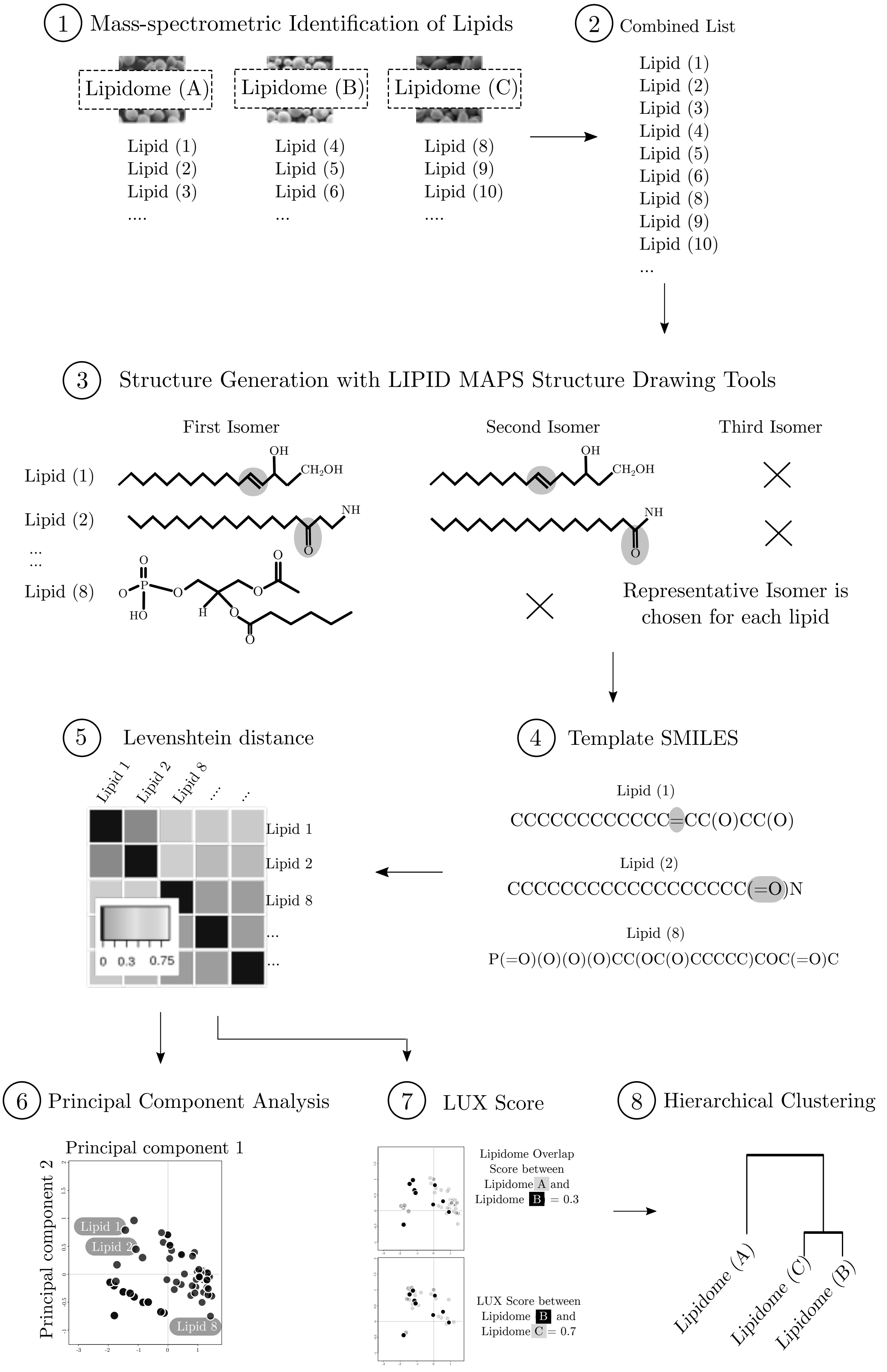


Figure 3.18: Illustration of Lipidome Juxtaposition Work flow.

information [section 2.2.3].

Levenshtein distance was selected for calculating distance matrices, although in principle, other similarity measures could be applied. Principal Component Analysis was performed for visualizing the lipidomes. The pairwise LUX scores between lipidomes was calculated and used as a metric for hierarchical clustering.

Discussion

There were three main themes in this study 1. how to represent lipid structures as strings? 2. how to compare lipid structures? and 3. how to compare lipidomes? In hindsight, the first two questions [section 3.1] were more challenging than third [section 3.3], which was relatively straight-forward after establishing methods to address questions 1 and 2.

4.1 How to represent lipid structures ?

A structure is easy to draw if all the atoms and their connections are well defined. In the case of lipidomics data, neither the atoms nor connections are completely defined, which makes lipid structure drawing a difficult problem.

4.1.1 Mining structure information from biochemistry literature

The output of a typical lipidomics work flow [Fig. 1.2] is a list of lipids but with limited structure details [Table 2.8]. For example, Cer 41:2;2 is has four structure details 1. the lipid class (Cer - ceramide) 2. the number of carbon atoms (42) 3. the number of double bonds (2) and 4. the number of hydroxyl groups (2) but it does not specify 1. the number of carbon atoms in long chain base 2. number of carbon atoms in amide linked acyl chain 3. the position of double bond (in long chain base ? or acyl chain ?) 4. position of hydroxyl group (in head group ? or acyl chain ?). To construct structure from lipid name, the missing details were filled by knowing biochemistry of the tissue such as a. enzymes that carry out hydroxylation reactions in that tissue and b. the typical acyl chains in other lipid-classes of the same organism.

A method was developed in this study to select one representative structure from a list of isomer possibilities. Although, the results with the use of this approach reflect the biological phenotype, follow up studies are required that will improve the selection of representative structure of isomers.

4.1.2 The consistency in lipid nomenclature

The lipid nomenclature used in publications is a concern because it effects the parsing of lipidomics data. Although, there is some consensus with using a common short-hand notation of lipid names [154], there are also differences, for example, a triacylglycerol is referred as TAG or TG [23,121,122]. The number of hydroxylations are sometimes mentioned but could be left out of the publications. There are two reasons for not explicitly indicating them in publications 1. it is known from previous work on the biochemistry of that tissue/organism 2. it could not be determined from the lipidomics work flow. Either way, this missing information poses a problem for generating parsing rules.

In the case of yeast, fruit fly and human lung lipids used in this study, all the data was obtained from a similar lipidomics work flow, which made the parsing less troublesome. To apply the computational methods developed in this study to lipidomics data obtained from other work flows, it is important to have a consistent nomenclature and a constant update of the parsing rules.

4.1.3 1D, 2D and 3D are points of view

A problem that is inherent in drawing any molecular structure is to select the best representation. Protein structures are represented as sequences (1D) or space-filling (3D) models, sugar molecules as line drawings (2D), DNA and RNA as sequences (1D). A major challenge in this study is to select suitable representation to draw lipid structures. 3D coordinates are not relevant because, unlike crystal structures of proteins, the atom coordinates were rarely obtained for lipids. Moreover, 3D models are appropriate for molecules that are measured in solid state (such as crystal structures of proteins), but not for lipids. So, for lipids, the problem is to choose between graph representation (2D) or a sequence (1D). This study focuses entirely on sequences, which is practical for its simplicity, but it is nevertheless incomplete.

A graph (2D) has no left-right orientation but a sequence has start-point and an end-point (culturally, left is start and right is end, but that is a different problem!). The choice to use sequence for lipids restricted a structure to left-right orientation, which had unintended consequences with alignment [section 4.2.6].

SMILES notation was the choice but there are other sequence formats such as InChI, ROSDAL that could be applied [69, 70]. InChI notation is gaining popularity and a hashed version, InChI Key, was proposed as search string for mining relevant literature in Internet [72, 73, 155].

4.1.4 Customize canonicalization for lipids

Canonical SMILES failed in the test cases but only two types were tested in this study [156]. Morgan rule of uniquely numbering atoms [82] could be modified to take advantage of lipid specific features. For example, the central atom in glycerol for TAGs, the amide-linked atom in long chain base for ceramides could be used as a starting atom. Acyl chains could be numbered following the sn1 and sn2 nomenclature.

4.1.5 Programmatic structure generation tools need improvements

Programmatic generation of structures from the lipid names was important task in this study. Each lipid species could have more than one possible structure, depending on the level of structural details gathered by the lipidomics work flow [Fig. 3.18]. Enumerating all the possible isomers and selecting one of them as a model structure, for that particular lipid, is a challenge.

In this study, the first lipid structures were drawn with a simple web-tool [131]. However, it was not realistic to apply the same process to datasets with 300+ lipid species (each with 300+ isomers). The programmatic structure drawing tools [141] were definitely of assistance but there was a problem. These tools were developed to draw lipids specific to mammals, with the focus on humans, which is not sufficient for this study. So, the scripts were modified to draw yeast and fruit fly lipids. It is a clear recommendation from this study that structure drawing tools must support more organisms.

4.2 How to compare lipid structures ?

This is the central problem of this study. There were no reports of comparing lipid structures until now, the only option was to search for methods from related subjects. Pharmacological molecules, proteins, DNA sequences are chemically or biologically related to lipids, naturally, the search for structure comparison methods had a bias in selecting the ones that were popular with those molecules [89]. In hindsight, this was a bad idea, because, the methods that were established for proteins and DNA could not be applied for lipids. The search was not exhaustive, so obviously there will be numerous other methods that could be adapted for lipids [101]. This section focuses on the lessons learned from the six methods and will end with suggestions for future structure comparison studies.

4.2.1 Fingerprints - they are everywhere

Fingerprints are suitable for pharmacological molecules because often, only a specific feature (or a set of features) of interest is compared [157]. In this study, one fingerprint type was tested, FP2, but others like FP3 and MACCS (Molecular ACCess System) could also be used [84]. It was observed that FP2 fingerprints were not suited for measuring the change in hydroxyl position of ceramides [section 3.1]. There are two possible reasons for this 1. the unique structure of ceramides *i.e* a long hydrocarbon chain that might have resulted in highly similar fingerprints 2. the metric used for comparing fingerprints may not be suitable for lipids as they were developed for pharmacological molecules. In future, a. the hash function to generate fingerprint and b. the metric for comparison should be optimized for lipids. The structural features of each lipid class (and sub-class) must be taken into consideration in designing the hash function.

Drug molecules were often compared with a combination of fingerprints [158]. It will be interesting to apply similar approach for lipids.

4.2.2 Sub-strings - put on more weight

The word frequency method to compare SMILES strings was in principle, a straight forward way to score differences, but it failed for ceramide and PI datasets [section 3.1]. Perhaps, LINGO method could be improved in three ways.

1. The scoring function [section 2.4] uses all LINGOs (hence, normalized by the total number of LINGOs), which could be improved by using a weighted measure. Similar to the substitution matrices use for protein sequence alignment, LINGOs could be weighted based on their frequency of occurrence.

2. Only one LINGO length was tested ($q = 4$). That value was reported in an earlier study [91] but with pharmacological molecules. In the future, it could be optimized for lipids.

3. The LINGOs were linear sub-strings, which affected certain functional groups. For example, a SMILES string *CCC(0)C* will contain a LINGO *CC(0* (one of the 4 possible LINGOs with $q = 4$). Although, it is a valid sub-string, the LINGO is chemically invalid because of the missing round closing bracket. Perhaps, LINGOs could be made non-linear by separating the branched atoms.

4.2.3 Bioisosterism - concept versus implementation

Bioisosterism is a useful concept to find chemical substitutes for specific functional groups, which is a common theme in pharmaceutical research [159]. Computational methods to find bioisosteres were developed [108], that will match a query molecule with structures from a database. I asked the question, whether the Biosisosterism concept could be applied for lipids ? As biomolecules are present in all living organisms that perform similar function but have different structures, lipids are right candidates for bioisosteric comparison.

Bioisosteric algorithm was designed for use with CACTVS canonical SMILES. Although the program did not display errors when Open Babel canonical and template SMILES were given as input, those results were not presented, because it is hard to interpret the results, when clearly, the input specifications were not met. In the future, the method could be adapted for use with template SMILES.

The central idea of separating the main chain from secondary chains, used in bioisosteric algorithm [Fig. 1.12] is relevant for lipids. Biosynthesis of ceramides takes place by the addition of acyl chains to head group, so it is appropriate to separate sn1, sn2 chains and compare them separately. It will be interesting to test if word-frequency method performs better with separated sn1 and sn2 chains.

4.2.4 SMIles multiple sequence aLIGNment (SMILIGN) - or not smiling ?

SMILIGN is a logical progression from Bioisosteric similarity in the search for a suitable method to compare SMILES. LINGO and Biosisosteric similarity were the only methods available in literature that specifically use SMILES as input to calculate similarity. These two methods (that were developed for pharmaceutical molecules) failed, so the next step was to try out protein sequence comparison methods for lipids [96, 97].

The first limitation in converting SMILES to amino-acid sequences is the character limit. Only 20 (one for each amino acid) letters could be used, so many SMILES were suitably edited [section 2.4.4]. The consequences of this character-replacement were obvious 1. the edited SMILES were no longer valid chemical structures and 2. the lipids that were very different had similar SMILES (leads to a false positive when compared). The MSA should have been calculated without an ad hoc limitation of 20 characters.

The first part of an MSA is a pair-wise alignment which is a time-limiting step. Methods to make fast pair wise comparisons include a seeding function of short length string frequency match (BLAST uses a length 3). It will be interesting to develop a seeding function for SMILES based on functional groups, that could make quick pair-wise comparisons in the first part, that is followed up with MSA.

An identity matrix was chosen for scoring an alignment match, which was not a bad choice, but in hindsight, the mis-match and gap (opening and extension) penalties were a poor choice. Assigning the same penalty for a mis-match, a gap opening and a gap extension (all, -10000), put these three on the same level. In future, their values must selected carefully, perhaps, optimized for each lipid class.

A pair of SMILES were compared after the multiple sequence alignment of all SMILES in the set, which was not a bad idea in the case of ceramides, but in the future, one must be careful in performing MSA on a structurally diverse set of lipids. It is better to separate lipid classes (or select SMILES of similar length) when performing MSA, otherwise, the resulting alignment could have numerous gaps.

4.2.5 Local alignment and Levenshtein - watch the match and mis-match scores

The Smith and Waterman method was tested with an identity matrix and only one set of gap penalties. This does not rule out the possibility that it would function better with appropriate parameters. Gap opening and extension had the same value (-0.5) which should be adjusted in future.

An identity matrix was used for SMILIGN, Smith-Waterman and Levenshtein alignments to score a match but it could be adjusted for each atom, for example, same valency (halogens, chalcogens etc.) should have a higher value. In the case of amino-acid sequences, large-databases are periodically screened and the log odds ratio of each substitution is used for determining the mis-match penalty value. It will be interesting to apply the same concept to lipids by measuring the re-occurrence frequency of a functional group in a population. However, such large scale lipidomics data is not yet available but should be possible in future. A more difficult challenge is to define substitution weights for unrelated SMILES characters such as '=' (double bond) and 'C' (methediyl group).

A back-tracking procedure was used when implementing sequence alignment (both Smith-Waterman and Levenshtein procedures have it). In the event of a gap and mis-match resulting in the same score, a mis-match is preferred (no gap). A mis-matched atom could have a greater impact on the lipid function, as compared to a gap which could be chemically interpreted as a neutral substitution. One must be careful in implementing the alignment procedure and consider the consequences of programming choices when adapting it for lipids.

Smith-Waterman similarity scoring function used in this study uses the number of mis-matches but not gaps. The function could be adjusted to include both gaps and mismatches.

4.2.6 2D graph comparison - a missed opportunity ?

Lipid molecules are generally visualized as line drawings which are 2D graphs. A number of 2D graph matching algorithms were available but not tested for lipids [160–162]. Line notation was preferred over graph representation in this study for three reasons 1. simplicity and 2. easier to enumerate isomer possibilities and 3. faster computational time to score similarity. The last two could be achieved with 2D graphs.

Sequences (1D) representation has the left-right orientation [section 4.1.3]. Although, fingerprints and sequence alignment based similarity scoring methods do not depend on the left-right orientation, the user must be careful when the SMILES are manually added to the list of programmatically generated SMILES, to ensure that all SMILES follow same orientation. 2D graphs do not have the left-right orientation problem.

Tandem mass spectrometry (MS^2) is increasingly being used to obtain better structure detail of lipids, especially glycolipids. For example, sn1 and sn2 composition of DAG and TAG species was determined in yeast using MS^2 [23]. Techniques to obtain substitution positions (double bond and hydroxylation) for individual molecular species are available but such methods were not used in conjunction with high throughput lipidomics. Improved structure identification work flows will lessen the isomer enumeration problem, which makes 2D graphs a better way to represent and compare lipids.

4.3 How to compare lipidomes ?

The concept of a PCA space for lipids developed in this study is useful to compare lipidomes in novel ways [section 1.3.3] [163–165]. Hausdorff distance is one measure that was tested but the other metrics are available in the theme of topological vector spaces [103, 166].

It is obvious that lipidome comparison work flow could be improved, given the various choices that are available for representing lipid structures and the parameters that could be adjusted for calculating similarity. In this section, the methods for improving the 1. lipidome visualization and 2. LUX score will be discussed.

4.3.1 PCA is good but lipidome visualization could be improved

The limitation of using only top three principal components to determine the *x, y* and *z* coordinates is apparent, especially when the total variance captured is less than 50%. The dimensional scaling methods that capture total variance in 2D or 3D, could be used instead of PCA [167]. Discriminant analysis and kernel PCA are other ways to reduce dimensionality. It will be interesting to apply these methods on lipid-distance matrices.

4.3.2 LUX Score with concentrations

This study focuses on lipid structures but that is not the only difference between a pair of lipidomes. Lipid abundances, that were the basis for correlation methods, are relevant when comparing lipidomes of the same individual (for example, before and after onset of a disease). In principle, it should be possible to combine structure and concentration differences. One way to do it is described below.

First, the Levenshtein distance d could be weighted as per the equation:

$$d_w = \frac{1}{1 + e^{50 \times (d - 0.1)}} \tag{4.1}$$

50 is the scaling factor and 0.1 is the switching factor, they could be adjusted depending on the distribution of all pair wise Levenshtein distances in a given set of lipidomes.

The concentration difference between a pair of molecules a and b could be weighted as per the equation:

$$c_w = \frac{1}{1 + e^{-50 \times |c_a - c_b|}} \tag{4.2}$$

c_a, c_b are concentration of lipids a and b, -50 is scaling factor which could be adjusted based on the distribution of all pair wise concentration differences in the two given lipidomes. The two values c_w, d_w [equations 4.1 and 4.2] could be combined to generate a concentration-weighted LUX score.

4.3.3 Feature selection, K-means clustering

In calculating the LUX score, there was no separation of lipid classes. It will be interesting to separate lipids in groups of structurally similar molecules (unbiased methods such as k-means clustering), and calculate Hausdorff distances for each cluster separately [168]. This might provide insights on the lipid-class specific differences.

It is interesting from biochemistry point of view to identify a specific lipid (or a group of lipids) that have most influence on the lipidome differences. Feature selections methods that could pick the a specific structural feature that has the most effect on the LUX score could be used in this context [169].

4.3.4 Variables in error modeling

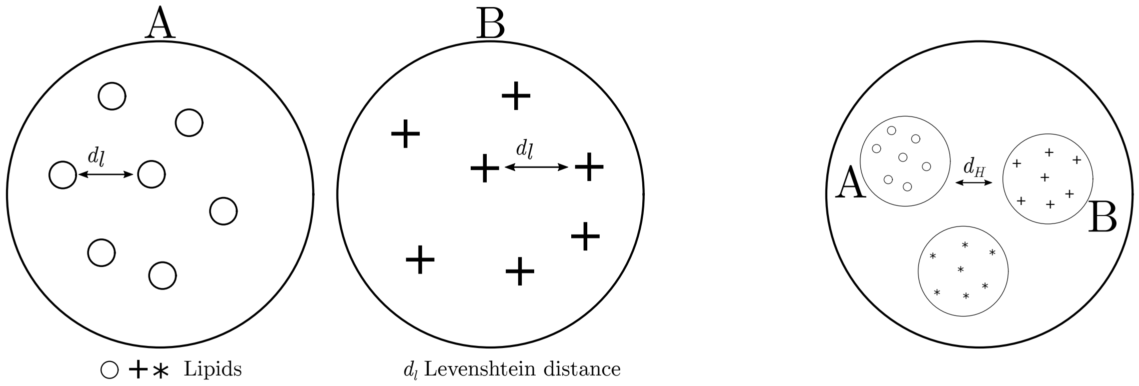
In phylogenetics, a bootstrapping procedure is used to test for robustness of the tree branching pattern. The error modeling procedure used in this study might appear similar to the boot strapping but they are not the same. Aim of error modeling was to understand the effect of lipid-detection criteria (that researchers use to identify a lipid from mass spectrum) on the LUX score calculation work flow. The procedure was important for yeast lipidome because the data was obtained from an external study and we had to test that LUX score based clustering was not an artifact of lipid detection threshold value.

When high detection threshold values were used in error modeling, the hierarchical clustering varied considerably. The t_{detect} values used in this study were carefully chosen by studying the lipid abundances and standard deviations in the data. One must identify appropriate parameters when performing similar procedure on other lipidomics data.

4.4 LUX score for comparing lipids across higher-order animal families

Computational methods developed in this study could be summarized as a generalized mathematical model for lipids and lipidomes [Fig. 4.1]. The broader aim is to compare lipid structures of model organisms and humans [170–172]. This study makes comparison between the tissues of same individual and mutants of the single species [section 3.3] but additional testing of LUX score is necessary, especially across species, genera and families to reach the broader aim.

Studies using DNA, RNA and protein sequences for comparing taxonomic families has led to an improved understanding of evolutionary relationships between organisms [173]. Methods described in this thesis could create new opportunities for comparing the metabolic relationships between model organisms and humans that use functional biomolecules such as lipids [174, 175].



(a) metric space of lipids

(b) metric space of lipidomes

Figure 4.1: Model of a metric space for lipids (a) and lipidomes (b). Let A and B are two lipidomes with 7 lipids each. Lipids are indicated by symbols (circle, plus and asterik). The separation of lipids is proportional to Levenshtein distance (d_l). The lipidomes were seperated by Hausdorff distance (d_H). In this model, lipidomes A and B do not overlap but in real data, there could be an overlap.

Bibliography

[1] F. Gunstone, J. Harwood, and A. Dijkstra. *The Lipid Handbook with CD-ROM, Third Edition*. CRC Press, 2007.

[2] F. Gunstone and B. Herslof. *Lipid Glossary 2*. Oily Press Lipid Library Series. Elsevier Science, 2000.

[3] O. G. Mouritsen. *Life - As a Matter of Fat*. The Frontiers Collection. Springer Verlag, Berlin/Heidelberg, 2005.

[4] C. Leray. *Lipids: Nutrition and Health*. CRC Press, 2014.

[5] G. P. Moss, P. a. S. Smith, and D. Tavernier. Glossary of class names of organic compounds and reactivity intermediates based on structure (IUPAC Recommendations 1995). *Pure Appl. Chem.*, 67:1307–1375, 1995.

[6] G. Rouser, G. Kritchevsky, C. Galli, and D. Heller. Determination of polar lipids: Quantitative column and thin-layer chromatography. *J. Am. Oil Chem. Soc.*, 42:215–227, 1965.

[7] W. W. Christie and X. Han. *Lipid Analysis: Isolation, Separation, Identification and Lipidomic Analysis*. Elsevier, 2010.

[8] E. Fahy, S. Subramaniam, R. C. Murphy, M. Nishijima, C. R. H. Raetz, T. Shimizu, F. Spener, G. van Meer, M. J. O. Wakelam, and E. A. Dennis. Update of the LIPID MAPS comprehensive classification system for lipids. *J. Lipid Res.*, 50:S9–S14, 2008.

[9] E. Fahy, S. Subramaniam, H. A. Brown, C. K. Glass, A. H. Merrill, R. C. Murphy, C. R. H. Raetz, D. W. Russell, Y. Seyama, W. Shaw, T. Shimizu, F. Spener, G. v. Meer, M. S. VanNieuwenhze, S. H. White, J. L. Witztum, and E. A. Dennis. A comprehensive classification system for lipids. *J. Lipid Res.*, 46:839–862, 2005.

[10] K. Kishimoto, R. Urade, T. Ogawa, and T. Moriyama. Nondestructive Quantification of Neutral Lipids by Thin-Layer Chromatography and Laser-Fluorescent Scanning: Suitable Methods for Lipidome Analysis. *Biochem. Biophys. Res. Commun.*, 281:657–662, 2001.

[11] X. Han and R. W. Gross. Global analyses of cellular lipidomes directly from crude extracts of biological samples by ESI mass spectrometry: a bridge to lipidomics. *J. Lipid Res.*, 44:1071–1079, 2003.

[12] E. A. Dennis. Lipidomics joins the omics evolution. *Proc. Natl. Acad. Sci.*, 106:2089–2090, 2009.

[13] G. Van Meer. Cellular lipidomics. *The EMBO J.*, 24:3159–3165, 2005.

[14] V. Shulaev. Metabolomics technology and bioinformatics. *Briefings Bioinforma.*, 7:128–139, 2006.

[15] X. Han and R. W. Gross. Electrospray ionization mass spectroscopic analysis of human erythrocyte plasma membrane phospholipids. *Proc. Natl. Acad. Sci.*, 91:10635–10639, 1994.

[16] B. Fuchs and J. Schiller. Application of maldi-tof mass spectrometry in lipidomics. *Eur. J. Lipid Sci. Technol.*, 111:83–98, 2009.

[17] X. Han and R. W. Gross. Shotgun lipidomics: multidimensional MS analysis of cellular lipidomes. *Expert. Rev. Proteomics*, 2:253–264, 2005.

[18] M. Hermansson, A. Uphoff, R. Käkelä, and P. Somerharju. Automated Quantitative Analysis of Complex Lipidomes by Liquid Chromatography/Mass Spectrometry. *Anal. Chem.*, 77:2166–2175, 2005.

[19] K. Yang, H. Cheng, R. W. Gross, and X. Han. Automated Lipid Identification and Quantification by Multidimensional Mass Spectrometry-Based Shotgun Lipidomics. *Anal. Chem.*, 81:4356–4368, 2009.

[20] V. Matyash, G. Liebisch, T. V. Kurzchalia, A. Shevchenko, and D. Schwudke. Lipid extraction by methyl-tert-butyl ether for high-throughput lipidomics. *J. Lipid Res.*, 49:1137–1146, 2008.

[21] G. Liebisch, M. Binder, R. Schifferer, T. Langmann, B. Schulz, and G. Schmitz. High throughput quantification of cholesterol and cholesteryl ester by electrospray ionization tandem mass spectrometry (esi-ms/ms). *Biochimica et Biophys. Acta (BBA) - Mol. Cell Biol. Lipids*, 1761:121 – 128, 2006.

[22] A. Shevchenko and K. Simons. Lipidomics: coming to grips with lipid diversity. *Nat. Rev. Mol. Cell Biol.*, 11:593–598, 2010.

[23] C. S. Ejsing, J. L. Sampaio, V. Surendranath, E. Duchoslav, K. Ekroos, R. W. Klemm, K. Simons, and A. Shevchenko. Global analysis of the yeast lipidome by quantitative shotgun mass spectrometry. *Proc. Natl. Acad. Sci. United States Am.*, 106:2136–2141, 2009.

[24] B. Brugger, B. Glass, P. Haberkant, I. Leibrecht, F. T. Wieland, and H. G. Krausslich. The HIV lipidome: A raft with an unusual composition. *Proc. Natl. Acad. Sci. United States Am.*, 103:2641–2646, 2006.

[25] A. X. da Silveira dos Santos, I. Riezman, M.-A. Aguilera-Romero, F. David, M. Piccolis, R. Loewith, O. Schaad, and H. Riezman. Systematic lipidomic analysis of yeast protein kinase and phosphatase mutants reveals novel insights into regulation of lipid homeostasis. *Mol. Biol. Cell*, 25:3234–3246, 2014.

[26] J. L. Sampaio, M. J. Gerl, C. Klose, C. S. Ejsing, H. Beug, K. Simons, and A. Shevchenko. Membrane lipidome of an epithelial cell line. *Proc. Natl. Acad. Sci. United States Am.*, 108:1903–1907, 2011.

[27] D. Schwudke, J. Oegema, L. Burton, E. Entchev, J. T. Hannich, C. S. Ejsing, T. Kurzchalia, and A. Shevchenko. Lipid profiling by multiple precursor and neutral loss scanning driven by the data-dependent acquisition. *Anal. Chem.*, 78:585–595, 2006.

[28] J. M. Foster, P. Moreno, A. Fabregat, H. Hermjakob, C. Steinbeck, R. Apweiler, M. J. O. Wakelam, and J. A. Vizcaíno. LipidHome: A Database of Theoretical Lipids Optimized for High Throughput Mass Spectrometry Lipidomics. *PLOS One*, 8:e61951, 2013.

[29] T. Kind, K.-H. Liu, D. Y. Lee, B. DeFelice, J. K. Meissen, and O. Fiehn. LipidBlast in silico tandem mass spectrometry database for lipid identification. *Nat. Methods*, 10:755–758, 2013.

[30] R. Herzog, K. Schuhmann, D. Schwudke, J. L. Sampaio, S. R. Bornstein, M. Schroeder, and A. Shevchenko. LipidXplorer: A Software for Consensual Cross-Platform Lipidomics. *PLOS One*, 7:e29851, 2012.

[31] T. Pluskal, S. Castillo, A. Villar-Briones, and M. Oresic. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinforma.*, 11:395, 2010.

[32] P. Husen, K. Tarasov, M. Katafiasz, E. Sokol, J. Vogt, J. Baumgart, R. Nitsch, K. Ekroos, and C. S. Ejsing. Analysis of Lipid Experiments (ALEX): A Software Framework for Analysis of High-Resolution Shotgun Lipidomics Data. *PLOS One*, 8:e79736, 2013.

[33] J. Hartler, M. Trötzmüller, C. Chitraju, F. Spener, H. C. Köfeler, and G. G. Thallinger. Lipid Data Analyzer: unattended identification and quantitation of lipids in LC-MS data. *Bioinforma. (Oxford, England)*, 27:572–577, 2011.

[34] G. S. V. McDowell, A. P. Blanchard, G. P. Taylor, D. Figeys, S. Fai, and S. A. L. Bennett. Predicting Glycerophosphoinositol Identities in Lipidomic Datasets Using VaLID (Visualization and Phospholipid Identification)–An Online Bioinformatic Search Engine. *BioMed Res. Int.*, 2014, 2014.

[35] Z. Ahmed, M. Mayr, S. Zeeshan, T. Dandekar, M. J. Mueller, and A. Fekete. Lipid-Pro: a computational lipid identification solution for untargeted lipidomics on data-independent acquisition tandem mass spectrometry platforms. *Bioinforma. (Oxford, England)*, 31:1150–1153, 2015.

[36] M. R. Wenk. The emerging field of lipidomics. *Nat. Rev. Drug Discov.*, 4:594–610, 2005.

[37] S. Leonelli. Model organism. In W. Dubitzky, O. Wolkenhauer, K.-H. Cho, and H. Yokota, editors, *Encyclopedia of Systems Biology*, pages 1398–1401. Springer New York, New York, NY, 2013.

[38] G. P. Rédei. *Encyclopedia of Genetics, Genomics, Proteomics and Informatics*, pages 1244–1244. Springer Netherlands, 2008.

[39] C. Klose, C. S. Ejsing, A. J. García-Sáez, H. J. Kaiser, J. L. Sampaio, M. A. Surma, A. Shevchenko, P. Schwille, and K. Simons. Yeast lipids can phase-separate into micrometer-scale membrane domains. *J. Biol. Chem.*, 285:30224–30232, 2010.

[40] M. A. Surma, C. Klose, R. W. Klemm, C. S. Ejsing, and K. Simons. Generic sorting of raft lipids into secretory vesicles in yeast. *Traffic*, 12:1139–1147, 2011.

[41] P. Tvrdik, R. Westerberg, S. Silve, A. Asadi, A. Jakobsson, B. Cannon, G. Loison, and A. Jacobsson. Role of a new mammalian gene family in the biosynthesis of very long chain fatty acids and sphingolipids. *The J. Cell Biol.*, 149:707–718, 2000.

[42] A. X. S. Santos and H. Riezman. Yeast as a model system for studying lipid homeostasis and function. *FEBS Lett.*, 586:2858–2867, 2012.

[43] A. Singh and R. Prasad. Comparative lipidomics of azole sensitive and resistant clinical isolates of candida albicans reveals unexpected diversity in molecular lipid imprints. *PLOS One*, 6:e19266, 2011.

[44] G. S. Richmond, F. Gibellini, S. A. Young, L. Major, H. Denton, A. Lilley, and T. K. Smith. Lipidomic analysis of bloodstream and procyclic form trypanosoma brucei. *Parasitol.*, 137:1357–1392, 2010.

[45] R. Welti, E. Mui, A. Sparks, S. Wernimont, G. Isaac, M. Kirisits, M. Roth, C. W. Roberts, C. Botté, E. Maréchal, and R. McLeod. Lipidomic analysis of toxoplasma gondii reveals unusual polar lipids. *Biochem.*, 46:13882–13890, 2007.

[46] L. Zheng, R. T'Kind, S. Decuypere, S. J. von Freyend, G. H. Coombs, and D. G. Watson. Profiling of lipids in leishmania donovani using hydrophilic interaction chromatography in combination with fourier transform mass spectrometry. *Rapid Commun. Mass Spectrom.*, 24:2074–2082, 2010.

[47] C. A. Madigan, T.-Y. Cheng, E. Layre, D. C. Young, M. J. McConnell, C. A. Debono, J. P. Murry, J.-R. Wei, C. E. Barry, G. M. Rodriguez, I. Matsunaga, E. J. Rubin, and D. B. Moody. Lipidomic discovery of deoxysiderophores reveals a revised mycobactin biosynthesis pathway in mycobacterium tuberculosis. *Proc. Natl. Acad. Sci. United States Am.*, 109:1257–1262, 2012.

[48] M. Witting and P. Schmitt-Kopplin. The Caenorhabditis elegans lipidome: A primer for lipid analysis in Caenorhabditis elegans. *Arch. Biochem. Biophys.*, 589:27–37, 2016.

[49] M. Carvalho, D. Schwudke, J. L. Sampaio, W. Palm, I. Riezman, G. Dey, G. D. Gupta, S. Mayor, H. Riezman, and A. Shevchenko. Survival strategies of a sterol auxotroph. *Dev.*, 137:3675–3685, 2010.

[50] G. Tortoriello, B. P. Rhodes, S. M. Takacs, J. M. Stuart, A. Basnet, S. Raboune, T. S. Widlanski, P. Doherty, T. Harkany, and H. B. Bradshaw. Targeted lipidomics in drosophila melanogaster identifies novel 2-monoacylglycerols and n-acyl amides. *PLOS ONE*, 8:1–10, 2013.

[51] K. A. Jeffries, D. R. Dempsey, A. L. Behari, R. L. Anderson, and D. J. Merkler. Drosophila melanogaster as a model system to study long-chain fatty acid amide metabolism. *FEBS Lett.*, 588:1596–1602, 2014.

[52] F. Beaudoin, L. V. Michaelson, S. J. Hey, M. J. Lewis, P. R. Shewry, O. Sayanova, and J. A. Napier. Heterologous reconstitution in yeast of the polyunsaturated fatty acid biosynthetic pathway. *Proc. Natl. Acad. Sci.*, 97:6421–6426, 2000.

[53] P. J. Trotter. The Genetics of Fatty Acid Metabolism in Saccharomyces Cerevisiae. *Annu. Rev. Nutr.*, 21:97–119, 2001.

[54] U. Acharya and J. K. Acharya. Enzymes of Sphingolipid metabolism in Drosophila melanogaster. *Cell. Mol. Life Sci. CMLS*, 62:128–142, 2005.

[55] C. E. Martin, C.-S. Oh, and Y. Jiang. Regulation of long chain unsaturated fatty acid synthesis in yeast. *Biochimica et Biophys. Acta*, 1771:271–285, 2007.

[56] L. R. Shen, C. Q. Lai, X. Feng, L. D. Parnell, J. B. Wan, J. D. Wang, D. Li, J. M. Ordovas, and J. X. Kang. Drosophila lacks C20 and C22 PUFAs. *J. Lipid Res.*, 51:2985–2992, 2010.

[57] R. Kraut. Roles of sphingolipids in Drosophila development and disease. *J. Neurochem.*, 116:764–778, 2011.

[58] L. A. Cowart and L. M. Obeid. Yeast Sphingolipids: Recent developments in understanding biosynthesis, regulation, and function. *Biochimica et Biophys. Acta*, 1771:421–431, 2007.

[59] M. Kniazeva, Q. T. Crawford, M. Seiber, C.-Y. Wang, and M. Han. Monomethyl branched-chain fatty acids play an essential role in Caenorhabditis elegans development. *PLOS Biol.*, 2:E257, 2004.

[60] E. V. Entchev, D. Schwudke, V. Zagoriy, V. Matyash, A. Bogdanova, B. Habermann, L. Zhu, A. Shevchenko, and T. V. Kurzchalia. Let-767 is required for the production of branched chain and long chain fatty acids in caenorhabditis elegans. *J. Biol. Chem.*, 283:17550–17560, 2008.

[61] Y. Koga and H. Morii. Recent advances in structural research on ether lipids from archaea including comparative and physiological aspects. *Biosci. Biotechnol. Biochem.*, 69:2019–2034, 2005.

[62] Y. H. Itoh, A. Sugai, I. Uda, and T. Itoh. The evolution of lipids. *Adv. Space Res.*, 28:719–724, 2001.

[63] J. Nielsen. Systems biology of lipid metabolism: From yeast to human. *FEBS Lett.*, 583:3905 – 3913, 2009.

[64] R. Hoffmann and P. Laszlo. Representation in Chemistry. *Angewandte Chemie Int. Ed. Engl.*, 30:1–16, 1991.

[65] T. Nakayama and Y. Fujiwara. Computer representation of generic chemical structures by an extended block-cutpoint tree. *J. Chem. Inf. Comput. Sci.*, 23:80–87, 1983.

[66] A. Dalby, J. G. Nourse, W. D. Hounshell, A. K. I. Gushurst, D. L. Grier, B. A. Leland, and J. Laufer. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.*, 32:244–255, 1992.

[67] M. F. Lynch. Introduction of computers in chemical structure information systems, or what is not recorded in the annals. In *The History and Heritage of Scientific and Technological Information Systems: Proceedings of the 2002 Conference*, pages 137–148. Information Today Inc. Medford NJ, 2004.

[68] W. J. Wiswesser. How the WLN began in 1949 and how it might be in 1999. *J. Chem. Inf. Comput. Sci.*, 22:88–93, 1982.

[69] S. Ash, M. A. Cline, R. W. Homer, T. Hurst, and G. B. Smith. SYBYL Line Notation (SLN): A Versatile Language for Chemical Structure Representation. *J. Chem. Inf. Comput. Sci.*, 37:71–79, 1997.

[70] H. Rohbeck. Representation of Structure Description Arranged Linearly. In P. D. J. Gmehling, editor, *Software Development in Chemistry 5*, pages 49–58. Springer Berlin Heidelberg, 1991.

[71] S. R. Heller, A. McNaught, I. Pletnev, S. Stein, and D. Tchekhovskoi. InChI, the IUPAC International Chemical Identifier. *J. Cheminform.*, 7:23, 2015.

[72] K. R. Taylor, R. J. Gledhill, J. W. Essex, J. G. Frey, S. W. Harris, and D. C. De Roure. Bringing Chemical Data onto the Semantic Web. *J. Chem. Inf. Model.*, 46:939–952, 2006.

[73] O. Casher and H. S. Rzepa. SemanticEye: A Semantic Web Application to Rationalize and Enhance Chemical Electronic Publishing. *J. Chem. Inf. Model.*, 46:2396–2411, 2006.

[74] D. Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28:31–36, 1988.

[75] P. Ertl, B. Rohde, and P. Selzer. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *J. Med. Chem.*, 43:3714–3717, 2000.

[76] J. J. Irwin and B. K. Shoichet. ZINC - A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.*, 45:177–182, 2005.

[77] D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, and M. Hassanali. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.*, 36:D901–D906, 2008.

[78] K. Degtyarenko, P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj, and M. Ashburner. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.*, 36:D344–D350, 2008.

[79] Y. Wang, J. Xiao, T. O. Suzek, J. Zhang, J. Wang, and S. H. Bryant. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.*, 37:W623–W633, 2009.

[80] M. Sud, E. Fahy, D. Cotter, A. Brown, E. A. Dennis, C. K. Glass, A. H. Merrill, R. C. Murphy, C. R. H. Raetz, D. W. Russell, and S. Subramaniam. LMSD: LIPID MAPS structure database. *Nucleic Acids Res.*, 35:D527–D532, 2007.

[81] D. Weininger, A. Weininger, and J. Weininger. Smiles .2. Algorithm For Generation Of Unique Smiles Notation. *J. Chem. Inf. Comput. Sci.*, 29:97–101, 1989.

[82] H. L. Morgan. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J. Chem. Documentation*, 5:107–113, 1965.

[83] W. D. Ihlenfeldt, Y. Takahashi, H. Abe, and S. Sasaki. Computation and management of chemical properties in CACTVS: An extensible networked approach toward modularity and compatibility. *J. Chem. Inf. Comput. Sci.*, 34:109–116, 1994.

[84] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison. Open Babel: An open chemical toolbox. *J. Cheminform.*, 3:33, 2011.

[85] P. Willett, J. M. Barnard, and G. M. Downs. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.*, 38:983–996, 1998.

[86] R. Todeschini and V. Consonni. *Handbook of Molecular Descriptors.* Wiley-VCH Verlag GmbH, 2008.

[87] D. Rogers and M. Hahn. Extended-connectivity fingerprints. *J. Chem. Inf. Model.*, 50:742–754, 2010.

[88] A. Bender, J. L. Jenkins, J. Scheiber, S. C. K. Sukuru, M. Glick, and J. W. Davies. How Similar Are Similarity Searching Methods? A Principal Component Analysis of Molecular Descriptor Space. *J. Chem. Inf. Model.*, 49:108–119, 2009.

[89] S. Vilar, E. Uriarte, L. Santana, T. Lorberbaum, G. Hripcsak, C. Friedman, and N. P. Tatonetti. Similarity-based modeling in large-scale prediction of drug-drug interactions. *Nat. Protoc.*, 9:2147–2163, 2014.

[90] D. Vidal, M. Thormann, and M. Pons. A Novel Search Engine for Virtual Screening of Very Large Databases. *J. Chem. Inf. Model.*, 46:836–843, 2006.

[91] D. Vidal, M. Thormann, and M. Pons. LINGO, an efficient holographic text based method to calculate biophysical properties and intermolecular similarities. *J. Chem. Inf. Model.*, 45:386–393, 2005.

[92] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48:443–453, 1970.

[93] D. W. Mount. *Bionformatics: Sequence and Genome Analysis.* CSHL Press, 2004.

[94] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *J. Mol. Biol.*, 147:195–197, 1981.

[95] R. Durbin, editor. *Biological sequence analysis: probabalistic models of proteins and nucleic acids.* Cambridge University Press, Cambridge, UK, 1998.

[96] J. D. Thompson, D. G. Higgins, and T. J. Gibson. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22:4673–4680, 1994.

[97] D. G. Higgens and W. R. Taylor. Multiple sequence alignment. In D. M. Webster, editor, *Protein Structure Prediction: Methods and Protocols*, pages 1–18. Humana Press, Totowa, NJ, 2000.

[98] R. Chenna, H. Sugawara, T. Koike, R. Lopez, T. J. Gibson, D. G. Higgins, and J. D. Thompson. Multiple sequence alignment with the clustal series of programs. *Nucleic Acids Res.*, 31:3497–3500, 2003.

[99] R. C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, 32:1792–1797, 2004.

[100] V. I. Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Sov. Phys. Doklady*, 10:707, 1966.

[101] G. Navarro. A guided tour to approximate string matching. *ACM Comput. Surv.*, 33:2001, 1999.

[102] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231, 2006.

[103] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis Mach. Intell.*, 24:509–522, 2001.

[104] J. A. Grant, J. A. Haigh, , B. T. Pickup, A. Nicholls, , and R. A. Sayle. Lingos, finite state machines, and fast similarity searching. *J. Chem. Inf. Model.*, 46:1912–1918, 2006.

[105] R. A. Wagner and M. J. Fischer. The String-to-String Correction Problem. *J. ACM*, 21:168–173, January 1974.

[106] H. L. Morgan. Spelling Correction in Systems Programs. *Commun. ACM*, 13:90–94, February 1970.

[107] E. Ukkonen. On approximate string matching. In *Foundations of Computation Theory*, pages 487–495. Springer, Berlin, Heidelberg, August 1983.

[108] M. Krier and M. C. Hutter. Bioisosteric similarity of molecules based on structural alignment and observed chemical replacements in drugs. *J. Chem. Inf. Model.*, 49:1280–1297, 2009.

[109] Y. Hattori. e-1 - Metric Spaces. In K. P. Hart, J. Nagata, and J. E. Vaughan, editors, *Encyclopedia of General Topology*, pages 235–238. Elsevier, Amsterdam, 2003.

[110] J.-L. Reymond, R. van Deursen, L. C. Blum, and L. Ruddigkeit. Chemical space as a source for new drugs. *MedChemComm*, 1:30, 2010.

[111] M. Awale and J.-L. Reymond. Cluster analysis of the drugbank chemical space using molecular quantum numbers. *Bioorganic & Medicinal Chem.*, 20:5372 – 5378, 2012.

[112] M. Boehm. *Virtual Screening of Chemical Space: From Generic Compound Collections to Tailored Screening Libraries*, pages 1–33. Wiley-VCH Verlag GmbH Co. KGaA, 2011.

[113] N. Singh, R. Guha, M. A. Giulianotti, C. Pinilla, R. A. Houghten, and J. L. Medina-Franco. Chemoinformatic analysis of combinatorial libraries, drugs, natural products, and molecular libraries small molecule repository. *J. Chem. Inf. Model.*, 49:1010–1024, 2009.

[114] M. A. Koch, A. Schuffenhauer, M. Scheck, S. Wetzel, M. Casaulta, A. Odermatt, P. Ertl, and H. Waldmann. Charting biologically relevant chemical space: a structural classification of natural products (sconp). *Proc. Natl. Acad. Sci. United States Am.*, 102:17272–17277, 2005.

[115] J. W. Godden and J. Bajorath. A distance function for retrieval of active molecules from complex chemical space representations. *J. Chem. Inf. Model.*, 46:1094–1097, 2006.

[116] C. Lipinski and A. Hopkins. Navigating chemical space for biology and medicine. *Nat.*, 432:855–861, 2004.

[117] A. Bender, J. L. Jenkins, J. Scheiber, S. C. K. Sukuru, M. Glick, and J. W. Davies. How similar are similarity searching methods? a principal component analysis of molecular descriptor space. *J. Chem. Inf. Model.*, 49:108–119, 2009.

[118] G. Wong, J. Chan, B. A. Kingwell, C. Leckie, and P. J. Meikle. LICRE: unsupervised feature correlation reduction for lipidomics. *Bioinforma. (Oxford, England)*, 30:2832–2833, 2014.

[119] S. M. Lam, Y. Wang, X. Duan, M. R. Wenk, R. N. Kalaria, C. P. Chen, M. K. P. Lai, and G. Shui. The brain lipidomes of subcortical ischemic vascular dementia and mixed dementia. *Neurobiol. Aging*, 2014.

[120] T. Řezanka, I. Kolouchová, and K. Sigler. Lipidomic analysis of psychrophilic yeasts cultivated at different temperatures. *Biochimica Et Biophys. Acta*, 1861:1634–1642, 2016.

[121] M. Carvalho, J. L. Sampaio, W. Palm, M. Brankatschk, S. Eaton, and A. Shevchenko. Effects of diet and development on the Drosophila lipidome. *Mol. Syst. Biol.*, 8:600, 2012.

[122] L. F. Eggers, J. Müller, C. Marella, V. Scholz, H. Watz, C. Kugler, K. F. Rabe, T. Goldmann, and D. Schwudke. Lipidomes of lung cancer and tumour-free lung tissues reveal distinct molecular signatures for cancer differentiation, age, inflammation, and pulmonary emphysema. *Manuscr. review*, 2017.

[123] K. Tarasov, A. Stefanko, A. Casanovas, M. A. Surma, Z. Berzina, H. K. Hannibal-Bach, K. Ekroos, and C. S. Ejsing. High-content screening of yeast mutant libraries by shotgun lipidomics. *Mol. bioSystems*, 2014.

[124] K. Bozek, Y. Wei, Z. Yan, X. Liu, J. Xiong, M. Sugimoto, M. Tomita, S. Pääbo, C. Sherwood, P. Hof, J. Ely, Y. Li, D. Steinhauser, L. Willmitzer, P. Giavalisco, and P. Khaitovich. Organization and Evolution of Brain Lipidome Revealed by Large-Scale Analysis of Human, Chimpanzee, Macaque, and Mouse Tissues. *Neuron*, 85:695–702, 2015.

[125] R. t'Kindt, E. D. Telenga, L. Jorge, A. J. M. Van Oosterhout, P. Sandra, N. H. T. Ten Hacken, and K. Sandra. Profiling over 1500 Lipids in Induced Lung Sputum and the Implications in Studying Lung Diseases. *Anal. Chem.*, 87:4957–4964, 2015.

[126] G. Shui, J. W. Stebbins, B. D. Lam, W. F. Cheong, S. M. Lam, F. Gregoire, J. Kusonoki, and M. R. Wenk. Comparative Plasma Lipidome between Human and Cynomolgus Monkey: Are Plasma Polar Lipids Good Biomarkers for Diabetic Monkeys? *PLOS One*, 6:e19731, 2011.

[127] P. S. Aguilar, M. G. Heiman, T. C. Walther, A. Engel, D. Schwudke, N. Gushwa, T. Kurzchalia, and P. Walter. Structure of sterol aliphatic chains affects yeast cell shape and cell fusion during mating. *Proc. Natl. Acad. Sci.*, 107:4170–4175, 2010.

[128] L. Desfarges, P. Durrens, H. Juguelin, C. Cassagne, M. Bonneu, and M. Aigle. Yeast mutants affected in viability upon starvation have a modified phospholipid composition. *Yeast*, 9:267–277, 1993.

[129] D. P. Huttenlocher, G. A. Klanderman, and W. A. Rucklidge. Comparing Images Using the Hausdorff Distance. *IEEE Trans. Pattern Anal. Mach. Intell.*, 15:850–863, 1993.

[130] M. P. Dubuisson and A. K. Jain. A modified Hausdorff distance for object matching. In *Proceedings of the 12th IAPR International Conference on Pattern Recognition, 1994. Vol. 1 - Conference A: Computer Vision and Image Processing*, pages 566–568, 1994.

[131] W. D. Ihlenfeldt, E. E. Bolton, and S. H. Bryant. The PubChem chemical structure sketcher. *J. Cheminform.*, 1:20, 2009.

[132] M. Sud, E. Fahy, D. Cotter, A. Brown, E. A. Dennis, C. K. Glass, A. H. Merrill, R. C. Murphy, C. R. Raetz, D. W. Russell, and S. Subramaniam. LMSD: LIPID MAPS structure database. *Nucleic Acids Res.*, 35:D527–532, 2007.

[133] N. Hadadi, K. Cher Soh, M. Seijo, A. Zisaki, X. Guan, M. R. Wenk, and V. Hatzimanikatis. A computational framework for integration of lipidomics data into metabolic pathways. *Metab. Eng.*, 23:1–8, 2014.

[134] D. Oursel, C. Loutelier-Bourhis, N. Orange, S. Chevalier, V. Norris, and C. M. Lange. Lipid composition of membranes of escherichia coli by liquid chromatography/tandem mass spectrometry using negative electrospray ionization. *Rapid Commun. Mass Spectrom.*, 21:1721–1728, 2007.

[135] E. Fahy, M. Sud, D. Cotter, and S. Subramaniam. LIPID MAPS online tools for lipid research. *Nucleic Acids Res.*, 35:W606–612, 2007.

[136] C. Marella, A. E. Torda, and D. Schwudke. The LUX Score: A Metric for Lipidome Homology. *PLOS Comput. Biol*, 11:e1004511, 2015.

[137] K. Hashimoto, A. C. Yoshizawa, S. Okuda, K. Kuma, S. Goto, and M. Kanehisa. The repertoire of desaturases and elongases reveals fatty acid variations in 56 eukaryotic genomes. *J. Lipid Res.*, 49:183–191, 2008.

[138] O. Tehlivets, K. Scheuringer, and S. D. Kohlwein. Fatty acid synthesis and elongation in yeast. *Biochimica et Biophys. Acta*, 1771:255–270, 2007.

[139] S. C. Shin, S.-H. Kim, H. You, B. Kim, A. C. Kim, K.-A. Lee, J.-H. Yoon, J.-H. Ryu, and W.-J. Lee. Drosophila microbiome modulates host developmental and metabolic homeostasis via insulin signaling. *Sci.*, 334:670–674, 2011.

[140] H. Fyrst, X. Zhang, D. R. Herr, H. S. Byun, R. Bittman, V. H. Phan, G. L. Harris, and J. D. Saba. Identification and characterization by electrospray mass spectrometry of endogenous Drosophila sphingadienes. *J. Lipid Res.*, 49:597–606, 2008.

[141] M. Sud, E. Fahy, and S. Subramaniam. Template-based combinatorial enumeration of virtual compound libraries for lipids. *J. Cheminform.*, 4:23, 2012.

[142] F. J. Damerau. A technique for computer detection and correction of spelling errors. *Commun. ACM*, 7:171–176, 1964.

[143] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2014.

[144] U. Ligges and M. Mächler. Scatterplot3d - an r package for visualizing multivariate data. *J. Stat. Softw.*, 8:1–20, 2003.

[145] F. Hausdorff. *Grundzüge der Mengenlehre.* Veit and Company, Leipzig, 1914.

[146] A. K. Jain, M. N. Murty, and P. J. Flynn. Data Clustering: A Review. *ACM Comput. Surv.*, 31:264–323, 1999.

[147] E. Paradis, J. Claude, and K. Strimmer. APE: analyses of phylogenetics and evolution in R language. *Bioinforma.*, 20:289–290, 2004.

[148] R. Schneiter, V. Tatzer, G. Gogg, E. Leitner, and S. D. Kohlwein. Elo1p-Dependent Carboxy-Terminal Elongation of C14:1Δ9 to C16:1Δ11 Fatty Acids inSaccharomyces cerevisiae. *J. Bacteriol.*, 182:3655–3660, 2000.

[149] C. S. Oh, D. A. Toke, S. Mandala, and C. E. Martin. ELO2 and ELO3, homologues of the Saccharomyces cerevisiae ELO1 gene, function in fatty acid elongation and are required for sphingolipid formation. *The J. Biol. Chem.*, 272:17376–17384, 1997.

[150] J. M. Cherry, E. L. Hong, C. Amundsen, R. Balakrishnan, G. Binkley, E. T. Chan, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. R. Engel, D. G. Fisk, J. E. Hirschman, B. C. Hitz, K. Karra, C. J. Krieger, S. R. Miyasato, R. S. Nash, J. Park, M. S. Skrzypek, M. Simison, S. Weng, and E. D. Wong. Saccharomyces Genome Database: the genomics resource of budding. *Nucleic Acids Res.*, 40:D700–705, 2012.

[151] C. B. Brachmann, A. Davies, G. J. Cost, E. Caputo, J. Li, P. Hieter, and J. D. Boeke. Designer deletion strains derived from Saccharomyces cerevisiae S288c: a useful set of strains and plasmids for PCR-mediated gene disruption and other applications. *Yeast (Chichester, England)*, 14:115–132, 1998.

[152] M. Watanabe, D. Watanabe, S. Nogami, S. Morishita, and Y. Ohya. Comprehensive and quantitative analysis of yeast deletion mutants defective in apical and isotropic bud growth. *Curr. Genet.*, 55:365–380, 2009.

[153] L. Ni and M. Snyder. A genomic study of the bipolar bud site selection pattern in Saccharomyces cerevisiae. *Mol. Biol. Cell*, 12:2147–2170, 2001.

[154] G. Liebisch, J. A. Vizcaíno, H. Köfeler, M. Trötzmüller, W. J. Griffiths, G. Schmitz, F. Spener, and M. J. O. Wakelam. Shorthand notation for lipid structures derived from mass spectrometry. *J. Lipid Res.*, 54:1523–1530, 2013.

[155] C. Southan. InChI in the wild: an assessment of InChIKey searching in Google. *J. Cheminform.*, 5:10, 2013.

[156] V. Hähnke, M. Rupp, M. Krier, F. Rippmann, and G. Schneider. Pharmacophore alignment search tool: Influence of canonical atom labeling on similarity searching. *J. Comput. Chem.*, 31:2810–2826, 2010.

[157] S. S. S. J. Ahmed and V. Ramakrishnan. Systems biological approach of molecular descriptors connectivity: optimal descriptors for oral bioavailability prediction. *PLOS One*, 7:e40654, 2012.

[158] J. D. Holliday, C. Hu, and P. Willett. Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2d fragment bit-strings. *Comb. Chem. & High Throughput Screen.*, 5:155–166, 2002.

[159] N. Brown, editor. *Bioisosteres in Medicinal Chemistry*. Wiley-VCH Verlag GmbH & Co. KGaA, 2012.

[160] S. C. Basak, S. Bertelsen, and G. D. Grunwald. Application of graph theoretical parameters in quantifying molecular similarity and structure-activity relationships. *J. Chem. Inf. Comput. Sci.*, 34:270–276, 1994.

[161] S. C. Basak, V. Magnuson, G. Niemi, and R. Regal. Determining structural similarity of chemicals using graph-theoretic indices. *Discret. Appl. Math.*, 19:17–44, 1988.

[162] M. Randić and C. L. Wilkins. Graph theoretical approach to recognition of structural similarity in molecules. *J. Chem. Inf. Comput. Sci.*, 19:31–37, 1979.

[163] A. Strehl, E. Strehl, J. Ghosh, and R. Mooney. Impact of Similarity Measures on Web-page Clustering. In *In Workshop on Artificial Intelligence for Web Search (AAAI 2000*, pages 58–64. AAAI, 2000.

[164] H. Becker, M. Naaman, and L. Gravano. Learning Similarity Metrics for Event Identification in Social Media. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 291–300, New York, NY, USA, 2010. ACM.

[165] C. M. Dobson. Chemical space and biology. *Nat.*, 432:824–828, 2004.

[166] Z. Zhu, C. Zhao, and Y. Hou. Research on Similarity Measurement for Texture Image Retrieval. *PLOS One*, 7:e45302, 2012.

[167] R. T. Rockafellar and R. J.-B. Wets. *Variational analysis.* Springer, Berlin; New York, 1998.

[168] J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Sci.*, 290:2319–2323, 2000.

[169] S. Wold, M. Sjöström, and L. Eriksson. Pls-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.*, 58:109–130, 2001.

[170] R. P. Kühnlein. Thematic review series: Lipid droplet synthesis and metabolism: from yeast to man. Lipid droplet-based storage fat metabolism in Drosophila. *J. Lipid Res.*, 53:1430–1436, 2012.

[171] S. Hindle, S. Hebbar, and S. T. Sweeney. Invertebrate models of lysosomal storage disease: what have we learned so far? *Invertebr. neuroscience: IN*, 11:59–71, 2011.

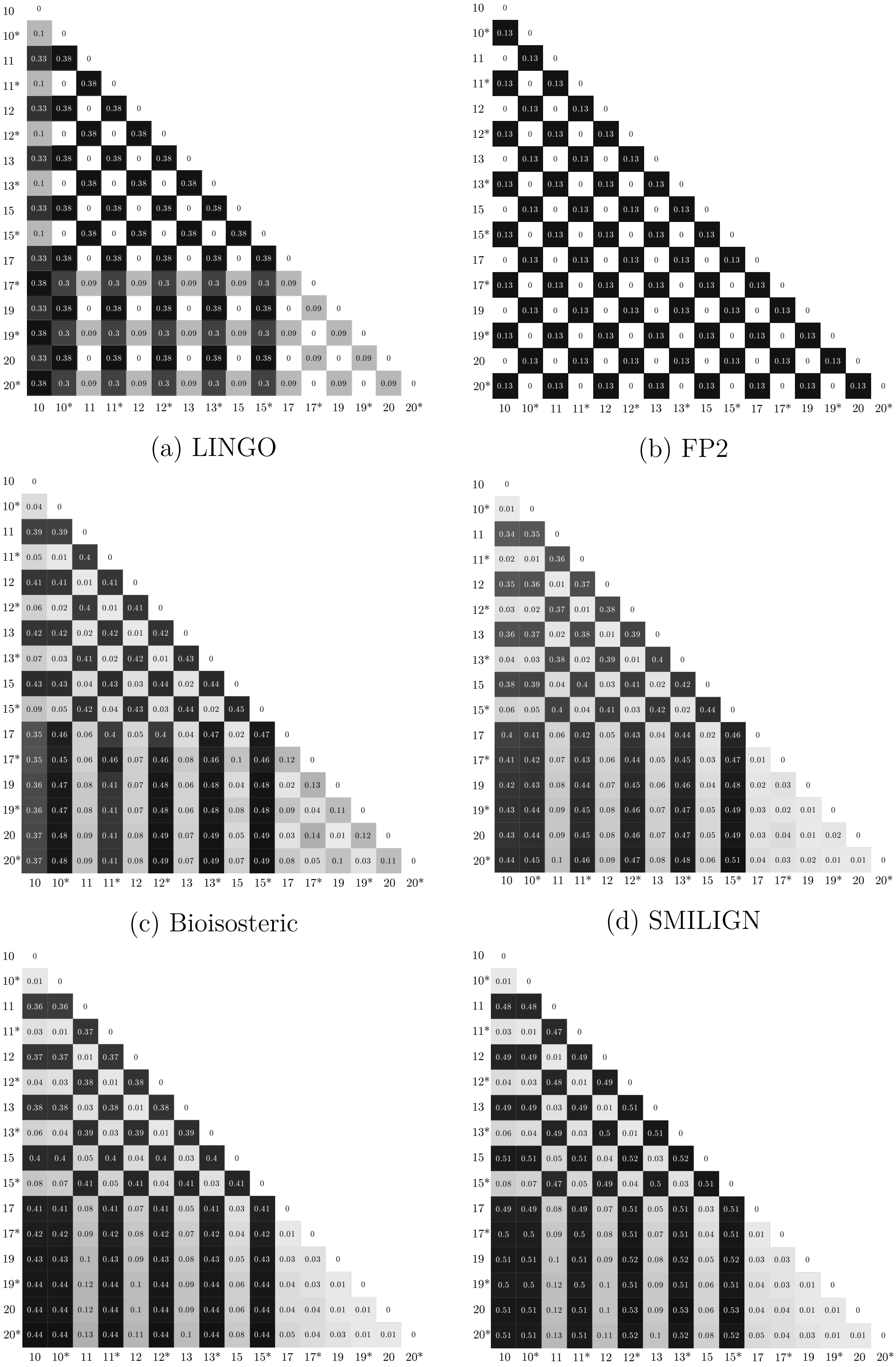
[172] M. E. Lopez and M. P. Scott. Genetic dissection of a cell-autonomous neurodegenerative disorder: lessons learned from mouse models of Niemann-Pick disease type C. *Dis. Model. & Mech.*, 6:1089–1100, 2013.

[173] S. A. Rahman, S. M. Cuesta, N. Furnham, G. L. Holliday, and J. M. Thornton. EC-BLAST: a tool to automatically search and compare enzyme reactions. *Nat. Methods*, 11:171–174, 2014.

[174] A. R. Joyce and B. O. Palsson. The model organism as a system: integrating'omics' data sets. *Nat. Rev. Mol. Cell Biol.*, 7:198–210, 2006.

[175] O. Fiehn. Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. *Comp. Funct. Genomics*, 2:155–168, 2001.

Supplementary Results



(a) LINGO

(b) FP2

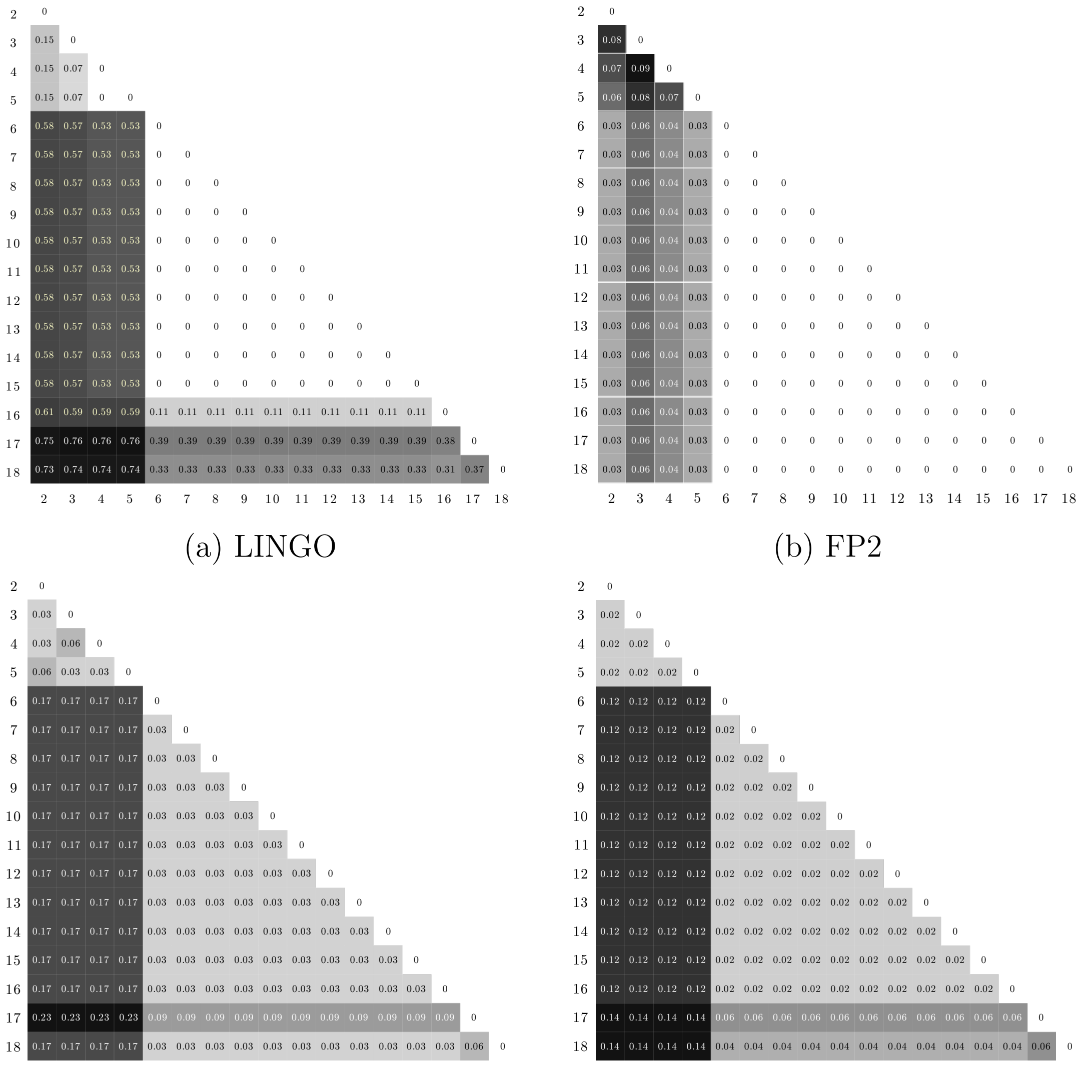
(c) Bioisosteric

(d) SMILIGN

(e) Smith-Waterman

(f) Levenshtein

Figure 4.2: Pairwise distances from CACTVS canonical SMILES of 16 PI. a. 34 non-identical pairs have zero LINGO distance. b. The matrix is similar for the Bioisosteric, SMILIGN, Smith-Wateran and Levenshtein distances . The checker board pattern on the top half of the matrix and alternating blue-yellow vertical bars on the lower half are distinct.

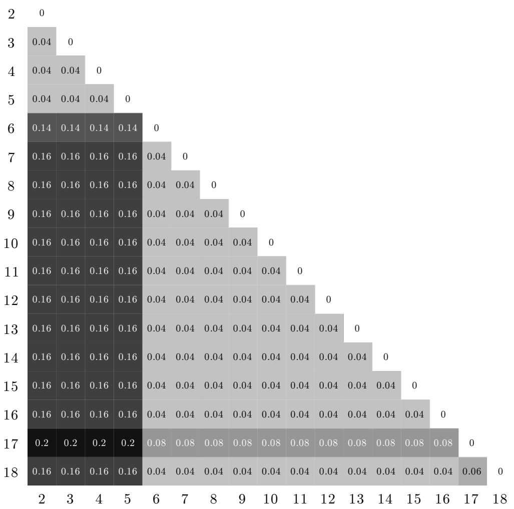


(a) LINGO

(b) FP2

(c) Bioisosteric

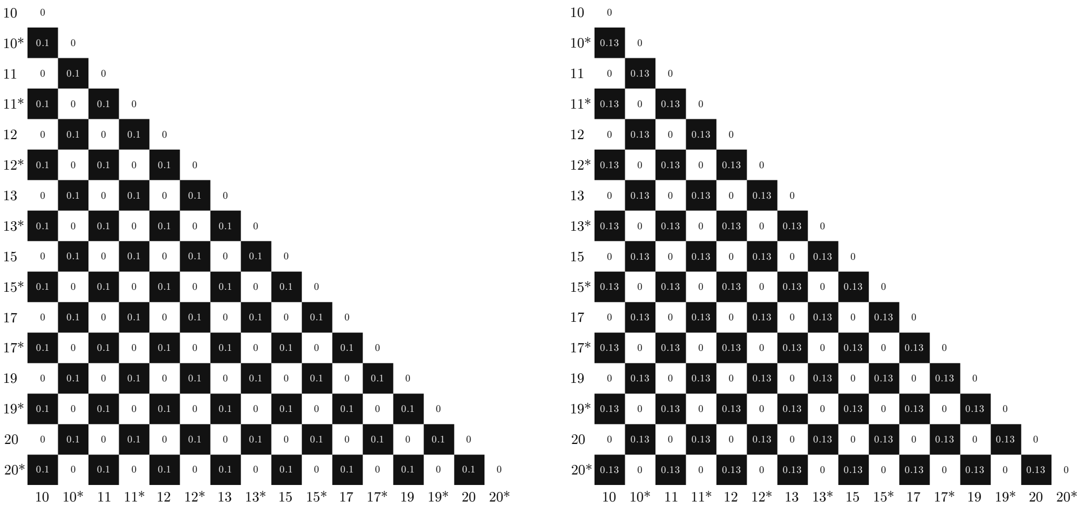
(d) SMILIGN



	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
2	0																
3	0.04	0															
4	0.04	0.04	0														
5	0.04	0.04	0.04	0													
6	0.14	0.14	0.14	0.14	0												
7	0.16	0.16	0.16	0.16	0.04	0											
8	0.16	0.16	0.16	0.16	0.04	0.04	0										
9	0.16	0.16	0.16	0.16	0.04	0.04	0.04	0									
10	0.16	0.16	0.16	0.16	0.04	0.04	0.04	0.04	0								
11	0.16	0.16	0.16	0.16	0.04	0.04	0.04	0.04	0.04	0							
12	0.16	0.16	0.16	0.16	0.04	0.04	0.04	0.04	0.04	0.04	0						
13	0.16	0.16	0.16	0.16	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0					
14	0.16	0.16	0.16	0.16	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0				
15	0.16	0.16	0.16	0.16	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0			
16	0.16	0.16	0.16	0.16	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0		
17	0.2	0.2	0.2	0.2	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0	
18	0.16	0.16	0.16	0.16	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.06	0

(e) Smith-Waterman

Figure 4.3: Pairwise distances from Open Babel canonical SMILES of 17 ceramides. The lowest non-zero distance in the matrix is between the pairs {3-4, 3-5} [Fig. **??**]. The lowest distance for the pair {3-4} is expected, because the structure difference between these two molecules is one methanediyl. The second set of lower distances is observed with the pairs associated with molecule 16 [row 16]. A similar trend was observed with the LINGO distance matrix with CACTVS SMILES, where, the pairs associated with molecule 16 had lower distances

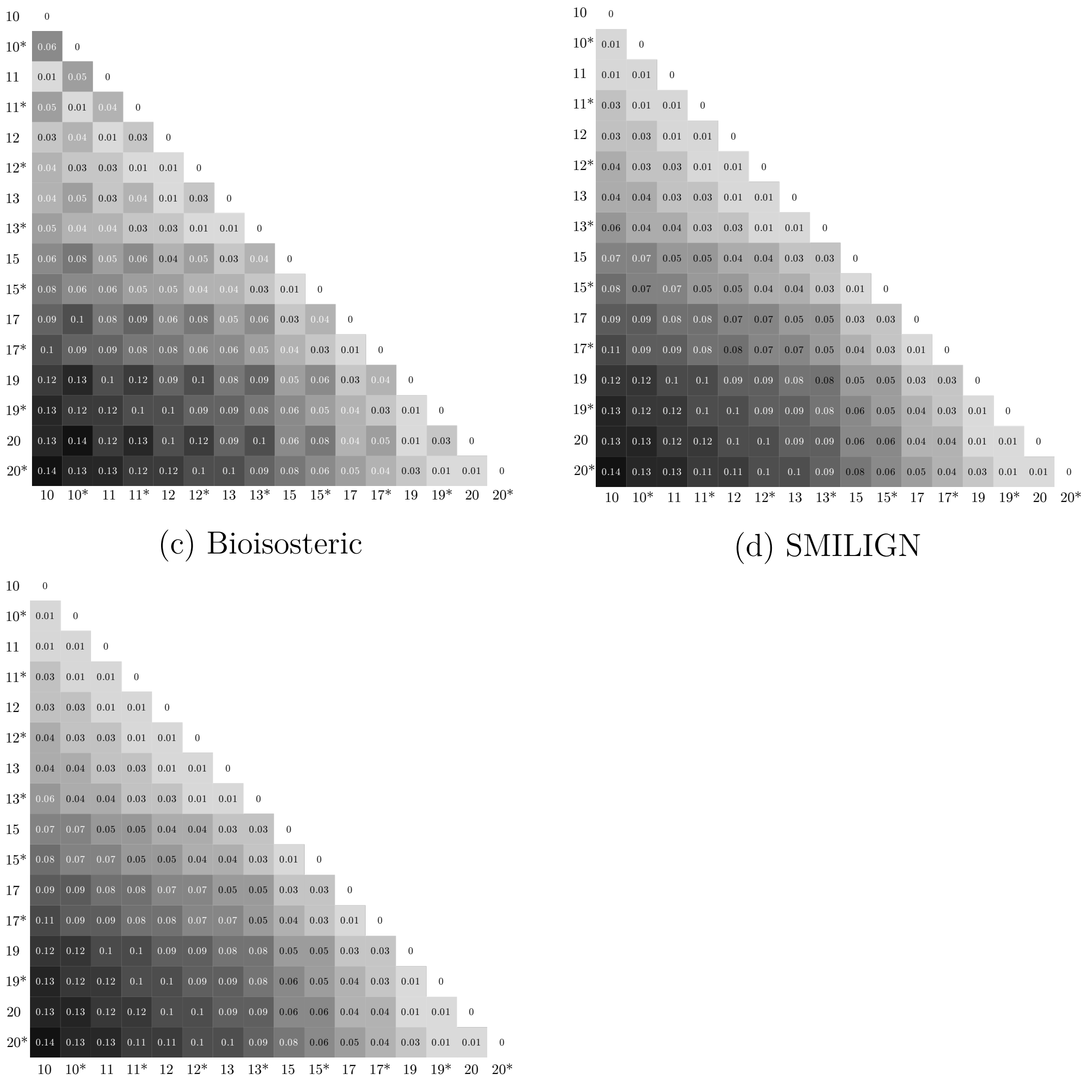


	10	10*	11	11*	12	12*	13	13*	15	15*	17	17*	19	19*	20	20*
10	0															
10*	0.1	0														
11	0	0.1	0													
11*	0.1	0	0.1	0												
12	0	0.1	0	0.1	0											
12*	0.1	0	0.1	0	0.1	0										
13	0	0.1	0	0.1	0	0.1	0									
13*	0.1	0	0.1	0	0.1	0	0.1	0								
15	0	0.1	0	0.1	0	0.1	0	0.1	0							
15*	0.1	0	0.1	0	0.1	0	0.1	0	0.1	0						
17	0	0.1	0	0.1	0	0.1	0	0.1	0	0.1	0					
17*	0.1	0	0.1	0	0.1	0	0.1	0	0.1	0	0.1	0				
19	0	0.1	0	0.1	0	0.1	0	0.1	0	0.1	0	0.1	0			
19*	0.1	0	0.1	0	0.1	0	0.1	0	0.1	0	0.1	0	0.1	0		
20	0	0.1	0	0.1	0	0.1	0	0.1	0	0.1	0	0.1	0	0.1	0	
20*	0.1	0	0.1	0	0.1	0	0.1	0	0.1	0	0.1	0	0.1	0	0.1	0

(a) LINGO

	10	10*	11	11*	12	12*	13	13*	15	15*	17	17*	19	19*	20	20*
10	0															
10*	0.13	0														
11	0	0.13	0													
11*	0.13	0	0.13	0												
12	0	0.13	0	0.13	0											
12*	0.13	0	0.13	0	0.13	0										
13	0	0.13	0	0.13	0	0.13	0									
13*	0.13	0	0.13	0	0.13	0	0.13	0								
15	0	0.13	0	0.13	0	0.13	0	0.13	0							
15*	0.13	0	0.13	0	0.13	0	0.13	0	0.13	0						
17	0	0.13	0	0.13	0	0.13	0	0.13	0	0.13	0					
17*	0.13	0	0.13	0	0.13	0	0.13	0	0.13	0	0.13	0				
19	0	0.13	0	0.13	0	0.13	0	0.13	0	0.13	0	0.13	0			
19*	0.13	0	0.13	0	0.13	0	0.13	0	0.13	0	0.13	0	0.13	0		
20	0	0.13	0	0.13	0	0.13	0	0.13	0	0.13	0	0.13	0	0.13	0	
20*	0.13	0	0.13	0	0.13	0	0.13	0	0.13	0	0.13	0	0.13	0	0.13	0

(b) FP2



	10	10*	11	11*	12	12*	13	13*	15	15*	17	17*	19	19*	20	20*
10	0															
10*	0.06	0														
11	0.01	0.05	0													
11*	0.05	0.01	0.04	0												
12	0.03	0.04	0.01	0.03	0											
12*	0.04	0.03	0.03	0.01	0.01	0										
13	0.04	0.05	0.03	0.04	0.01	0.03	0									
13*	0.05	0.04	0.04	0.03	0.03	0.01	0.01	0								
15	0.06	0.08	0.05	0.06	0.04	0.05	0.03	0.04	0							
15*	0.08	0.06	0.06	0.05	0.05	0.04	0.04	0.03	0.01	0						
17	0.09	0.1	0.08	0.09	0.06	0.08	0.05	0.06	0.03	0.04	0					
17*	0.1	0.09	0.09	0.08	0.08	0.06	0.06	0.05	0.04	0.03	0.01	0				
19	0.12	0.13	0.1	0.12	0.09	0.1	0.08	0.09	0.05	0.06	0.03	0.04	0			
19*	0.13	0.12	0.12	0.1	0.1	0.09	0.09	0.08	0.06	0.05	0.04	0.03	0.01	0		
20	0.13	0.14	0.12	0.13	0.1	0.12	0.09	0.1	0.06	0.08	0.04	0.05	0.01	0.03	0	
20*	0.14	0.13	0.13	0.12	0.12	0.1	0.1	0.09	0.08	0.06	0.05	0.04	0.03	0.01	0.01	0

(c) Bioisosteric

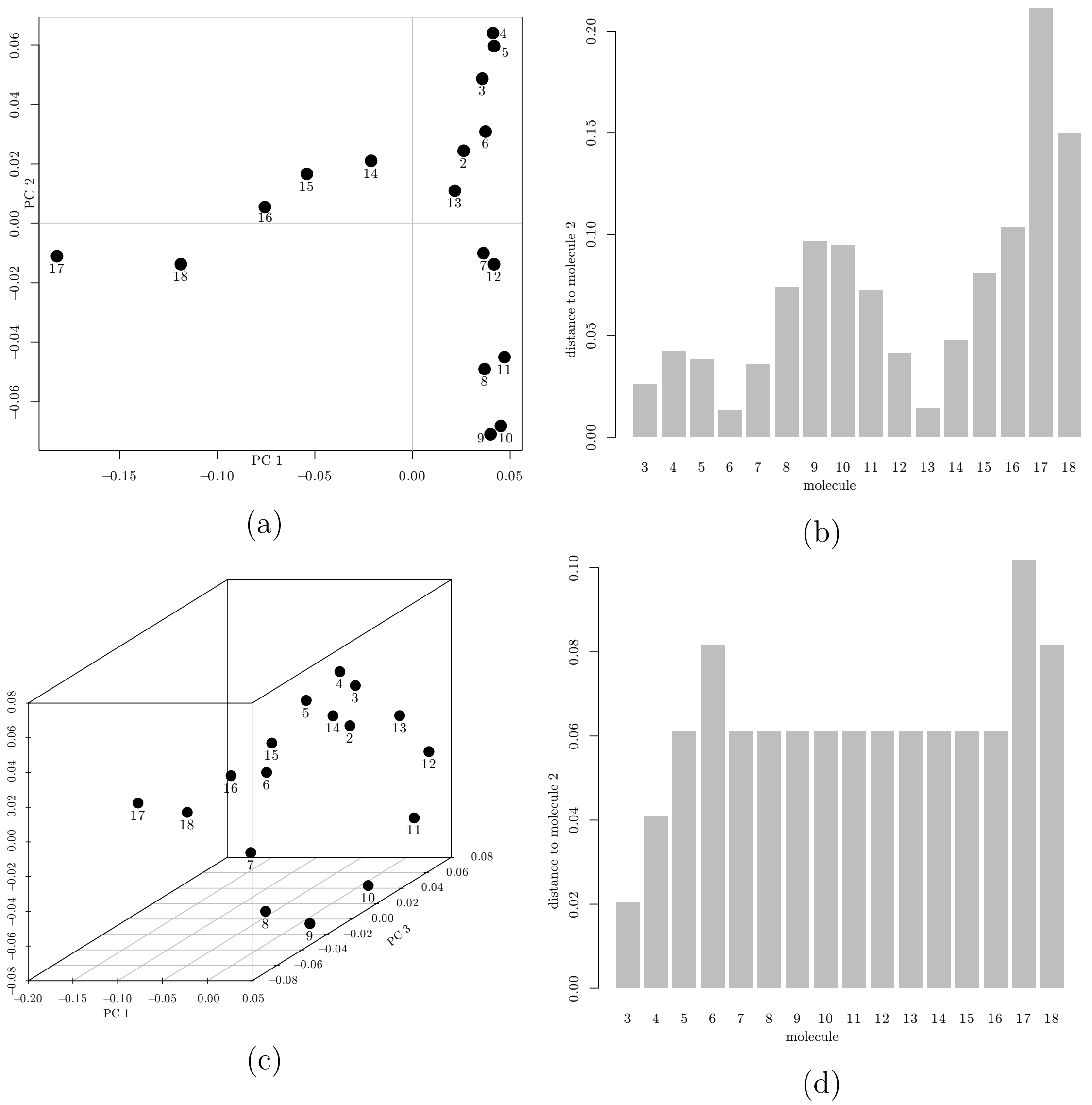
	10	10*	11	11*	12	12*	13	13*	15	15*	17	17*	19	19*	20	20*
10	0															
10*	0.01	0														
11	0.01	0.01	0													
11*	0.03	0.01	0.01	0												
12	0.03	0.03	0.01	0.01	0											
12*	0.04	0.03	0.03	0.01	0.01	0										
13	0.04	0.04	0.03	0.03	0.01	0.01	0									
13*	0.06	0.04	0.04	0.03	0.03	0.01	0.01	0								
15	0.07	0.07	0.05	0.05	0.04	0.04	0.03	0.03	0							
15*	0.08	0.07	0.07	0.05	0.05	0.04	0.04	0.03	0.01	0						
17	0.09	0.09	0.08	0.08	0.07	0.07	0.05	0.05	0.03	0.03	0					
17*	0.11	0.09	0.09	0.08	0.08	0.07	0.07	0.05	0.04	0.03	0.01	0				
19	0.12	0.12	0.1	0.1	0.09	0.09	0.08	0.08	0.05	0.05	0.03	0.03	0			
19*	0.13	0.12	0.12	0.1	0.1	0.09	0.09	0.08	0.06	0.05	0.04	0.03	0.01	0		
20	0.13	0.13	0.12	0.12	0.1	0.1	0.09	0.09	0.06	0.06	0.04	0.04	0.01	0.01	0	
20*	0.14	0.13	0.13	0.11	0.11	0.1	0.1	0.09	0.08	0.06	0.05	0.04	0.03	0.01	0.01	0

(d) SMILIGN

	10	10*	11	11*	12	12*	13	13*	15	15*	17	17*	19	19*	20	20*
10	0															
10*	0.01	0														
11	0.01	0.01	0													
11*	0.03	0.01	0.01	0												
12	0.03	0.03	0.01	0.01	0											
12*	0.04	0.03	0.03	0.01	0.01	0										
13	0.04	0.04	0.03	0.03	0.01	0.01	0									
13*	0.06	0.04	0.04	0.03	0.03	0.01	0.01	0								
15	0.07	0.07	0.05	0.05	0.04	0.04	0.03	0.03	0							
15*	0.08	0.07	0.07	0.05	0.05	0.04	0.04	0.03	0.01	0						
17	0.09	0.09	0.08	0.08	0.07	0.07	0.05	0.05	0.03	0.03	0					
17*	0.11	0.09	0.09	0.08	0.08	0.07	0.07	0.05	0.04	0.03	0.01	0				
19	0.12	0.12	0.1	0.1	0.09	0.09	0.08	0.08	0.05	0.05	0.03	0.03	0			
19*	0.13	0.12	0.12	0.1	0.1	0.09	0.09	0.08	0.06	0.05	0.04	0.03	0.01	0		
20	0.13	0.13	0.12	0.12	0.1	0.1	0.09	0.09	0.06	0.06	0.04	0.04	0.01	0.01	0	
20*	0.14	0.13	0.13	0.11	0.11	0.1	0.1	0.09	0.08	0.06	0.05	0.04	0.03	0.01	0.01	0

(e) Smith-Waterman

Figure 4.4: Pairwise distances from CACTVS canonical SMILES of 16 PI. a. 34 non-identical pairs have zero LINGO distance. b. The matrix is similar for the Bioisosteric, SMILIGN, Smith-Wateran and Levenshtein distances . The checker board pattern on the top half of the matrix and the alternating blue-yellow vertical bars on the lower half are distinct.



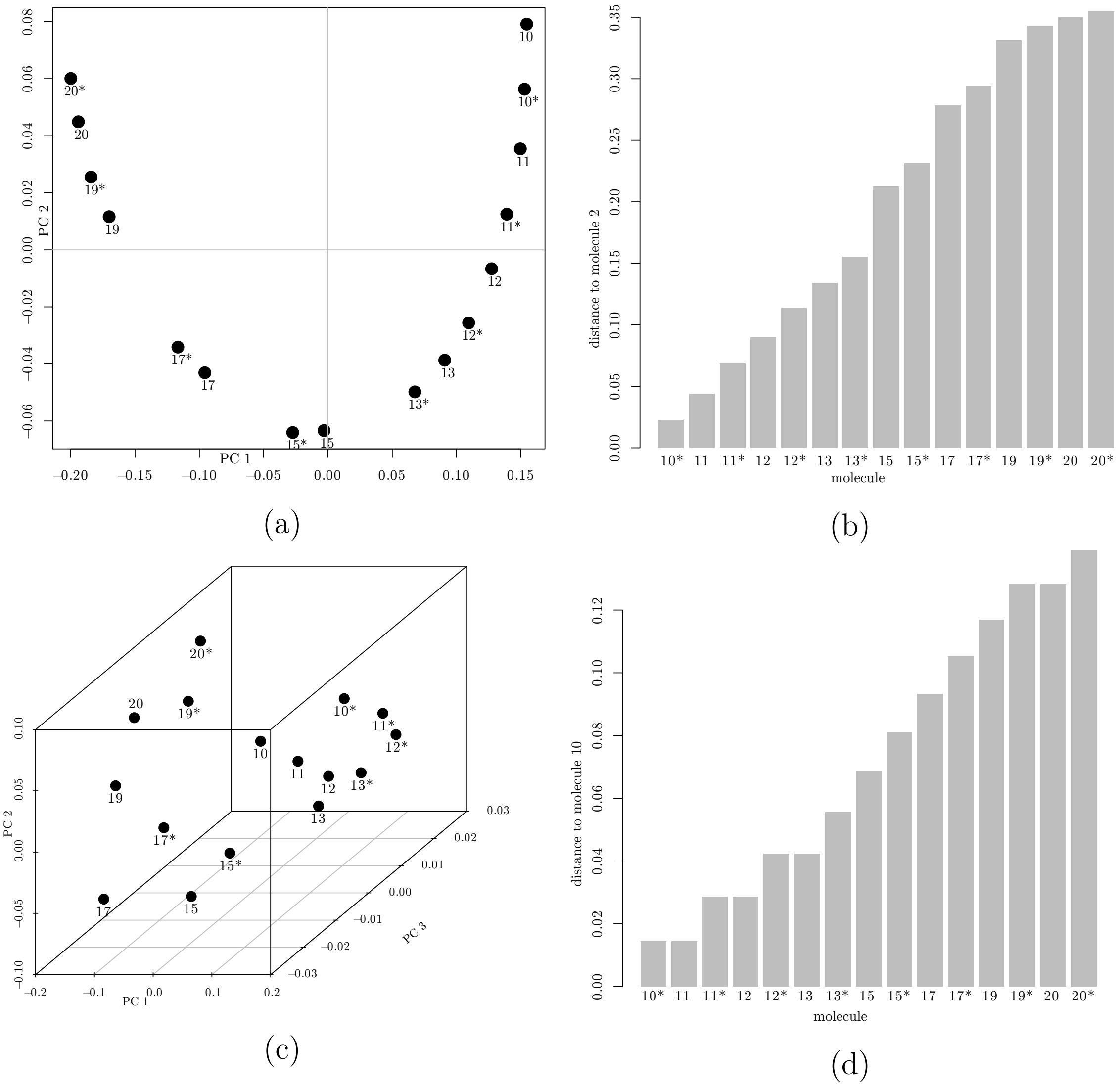
(a)

(b)

(c)

(d)

Figure 4.5: Distribution of 17 Ceramides in 2D and 3D Coordinate Space (S-W). Molecule cordinates in x, y and z axis derived from top three pricipal components of pairwise Smith Waterman distance matrix. Euclidean distance is calculated from the x, y and z coordinates.



(a)

(b)

(c)

(d)

Figure 4.6: Distribution of 16 PI in 2D and 3D Coordinate Space (S-W). Molecule cordinates in x, y and z axis derived from top three pricipal components of pairwise Smith Waterman distance matrix. Euclidean distance is calculated from the x, y and z coordinates.

List of Figures

List of Tables

No harmful chemicals were used in this study.

No animals were used in this study.

Thanks to thesis advisors, their critique at various stages of the study helped me re-evaluate methods and generate new ideas. Thanks to Prof. Andrew Torda and Dr. Geoffry Hyde for the help with English language and useful feedback. Thanks to Dr. Nicole Zehethofer, Dr. Jochen Behrens for proof reading the thesis and Dr. Lena Heinbrockel for German language help.

Thanks to Dr. Ashutosh Srivastava, Dr. Anil Dasanna and Dr. Aridman Pandit, discussions with him set the foundational idea for starting this project. Thanks to Dr. Sandeep krishna and Dr. Eli Lebow, discussions with them helped me convert ideas to mathematical models and algorithms. Thanks to members of Stack Exchange and Perl Monks for help with debugging. Thanks to colleagues Dr. Pritha Ghosh, Dr. Karthik Kamath for the support and of course for lively cafeteria conversations.

Thanks to Dr. Diego Goyeneche-Patiño, Dr. Monika Mackeiwitz and Dr. Lekha Damodaran for the warm welcome in Germany. Thanks to Gudrun Lehwark-Yvetot, Doreen Beyer and Dr. Bianca Schneider for the help with bureaucracy.

Thanks to friends Dr. Prasanth Ghanta, Dr. Avinash Reddy, Bhava Vemuru, Dr. Anand Kantari, Dr. Arun Shastry and Sriram Pattaswamy for the support. Thanks to mentors Dr. Suguna Challa, Dr. Balasubramanyam Muttuswamy, Prof. Basavaraju R, Prof. Janardhana Chelli, Prof. Panduranga Rao, Mr. Rao K. S, Ms. Chamundeswari and Dr. R. Ramu. Special thanks to Satyasai Baba for the guidance. Finally, I would like to thank my family for the love and support. I will gladly dedicate this work to them.

I hereby declare on oath, that I have written the present dissertation on my own and have not used resources and aids, other than those acknowledged. The submitted written version corresponds to the version on the electronic storage medium. I hereby declare that I have not previously applied or pursued for a doctorate (Ph.D. studies).

April 20, 2017

Chakravarthy Marella