Tracing Molecular Properties Throughout Evolution: A Chemoinformatic Approach.

Marcelo Otero[a,b], Silvina N. Sarno[c], Sofía L. Acebedo[d,e], Javier A. Ramírez[d,e]*.

[a]Universidad de Buenos Aires. Facultad de Ciencias Exactas y Naturales. Departamento de Física. Buenos Aires, Argentina.

[b]Universidad de Buenos Aires. CONICET. Instituto de Física de Buenos Aires (IFIBA). Buenos Aires, Argentina.

[c]Escuela de Ciencia y Tecnología, Universidad Nacional de San Martín, Martín de Irigoyen 3100, 1650, San Martín, Provincia de Buenos Aires, Argentina.

[d]Universidad de Buenos Aires. Facultad de Ciencias Exactas y Naturales. Departamento de Química Orgánica. Buenos Aires, Argentina.

[e]Universidad de Buenos Aires. CONICET. Unidad de Microanálisis y Métodos Físicos Aplicados a Química Orgánica (UMYMFOR). Buenos Aires, Argentina.

* Corresponding author. E-mail: jar@qo.fcen.uba.ar. Tel: +541145763385.

Postal address: Ciudad Universitaria, Pabellón II, Tercer Piso. C1428EGA. Ciudad Autónoma de Buenos Aires, Argentina.

ABSTRACT

Evolution of metabolism is a longstanding yet unresolved question, and several hypotheses were proposed to address this complex process from a Darwinian point of view. Modern statistical bioinformatic approaches targeted to the comparative analysis of genomes are being used to detect signatures of natural selection at the gene and population level, as an attempt to understand the origin of primordial metabolism and its expansion. These studies, however, are still mainly centered on genes and the proteins they encode, somehow neglecting the small organic chemicals that support life processes. In this work, we selected steroids as an ancient family of metabolites widely distributed in all eukaryotes and applied unsupervised machine learning techniques to reveal the traits that natural selection has imprinted on molecular properties throughout the evolutionary process. Our results clearly show that sterols, the primal steroids that first appeared, have more conserved properties and that, from then on, more complex compounds with increasingly diverse properties have emerged, suggesting that chemical diversification parallels the expansion of biological complexity. In a wider context, these findings highlight the worth of chemoinformatic approaches to a better understanding the evolution of metabolism.

KEYWORDS

Metabolic evolution - Molecular properties - Chemoinformatics - Machine Learning - Steroids

1. INTRODUCTION

Every living organism is a rich source of organic small molecules with a wide range of chemical structures. This chemical diversity originates from the activity of large and numerous families of enzymes that operate in highly branched metabolic pathways (Kroymann, 2011). Although we do not know how and when these pathways originated, several and sometimes opposed models which try to explain the evolution of metabolism were developed (Fani and Fondi, 2009), models that have recently come under scrutiny thanks to the statistical analysis of the genomes of a growing number of species (Scossa and Fernie, 2020).

Nevertheless, most of these hypotheses are focused on the evolution of the metabolic enzymes rather than the metabolites they produce. In this sense, a little discussed albeit intriguing model is that of Firn and Jones, who pose an alternative evolutionary framework (Firn and Jones, 2009, 2000). The basic idea behind their model is that natural selection acts on the *physicochemical properties of the metabolites* rather than on the genes encoding the enzymes that produce them. Thus, if a mutation leads to an enzyme able to synthetize a new molecule having properties that enhance the fitness of the organism, then selection will favour the retention of individuals possessing such variant relative to those that do not, which in turn put a selective pressure on the involved enzymes themselves.

Thus, if this hypothesis holds true, it would be possible to find traces that natural selection has imprinted on the properties of metabolites. In order to perform an exploratory test of this idea, we decided to select a widely distributed family of metabolites, with an ancient origin and whose biosynthetic relationships are well known. In this sense, steroids emerged as an appealing set of compounds, as they have essential biological functions in all eukaryotes.

Moreover, in the last decades a wealth of information has been gathered about the underlying metabolic pathways that generate them, which revealed several conserved mechanisms across different taxa.

Steroid biosynthesis comprises essentially three phases. In the first one, squalene, the alicyclic triterpene which is the common precursor of steroids both in animals and plants, is oxidized by the enzyme oxidosqualene cyclase, suffering a domino-like series of ring-closing reactions with the concomitant migration of methyl groups leading to lanosterol in animals and fungi, or cycloartenol in plants (Lednicer, 2011). This is an ancient conserved pathway: nowadays, the most widely accepted hypothesis is that sterols may have evolved in eukaryotes as an adaptive response to the sharp rise in atmospheric oxygen (Brown and Galea, 2010) in early Earth about 2.3 billion years ago (Galea and Brown, 2009; Gold et al., 2017), since steroids limit oxygen diffusion across cell membranes in eukaryotes, thus controlling the intracellular levels of reactive oxygen species (Dotson et al., 2017; Khan et al., 2003; Popova et al., n.d.; Widomska et al., 2007).

Afterwards, the tetracyclic carbon skeleton is modified mainly by oxidative enzymatic transformations which are phylum-dependent: lanosterol is converted into cholesterol through a nineteen-step pathway in animals, whereas a very similar set of transformations leads to ergosterol in some fungi clades as Ascomycetes and Basidiomycetes (Weete et al., 2010), and to dinosterol in Dinoflagellata (Lu et al., 2020). Alternatively, cycloartenol is converted into the major plant steroids campesterol and sitosterol (Schaller, 2003; Tamura et al., 1992). All these sterols share the property of being able to interact with phospholipids within biological membranes modifying their fluidity, permeability and other biophysical features, which turn these steroids into essential players in living processes.

From this point on, the metabolic pathways diverge in increasingly complex ways. Apart from being fundamental components of the biomembranes, the aforementioned steroids may serve as precursors of a series of derivatives with more specialized biological functions. In this second biosynthetic phase, cholesterol is transformed into more oxidized compounds such as dafachronic acids in nematodes (Bento et al., 2010), ecdysteroids in arthropods (Honda et al., 2017; Nakagawa and Henrich, 2009; Niwa and Niwa, 2014), and oxysterols and bile acids in vertebrates (Fonseca et al., 2017). These metabolites usually serve as signaling molecules that were essential to evolution of multicellular animals, and recent evidence suggests that the necessary biosynthetic enzymatic machinery first apperared 700 million years ago and diversified for the next 400 millions years (Markov et al., 2017). In vertebrates, additional oxidative transformations may lead to the cleavage of the side chain of cholesterol to give all steroid hormones, including adrenal gland hormones and sex hormones such as estrogens and androgens (Markov et al., 2017). On the other hand, sitosterol and campesterol also serve as precursors for the biosynthesis of brassinosteroids, polyoxygenated metabolites with hormonal action in most vascular plants and some ferns and mosses (Figure 1).

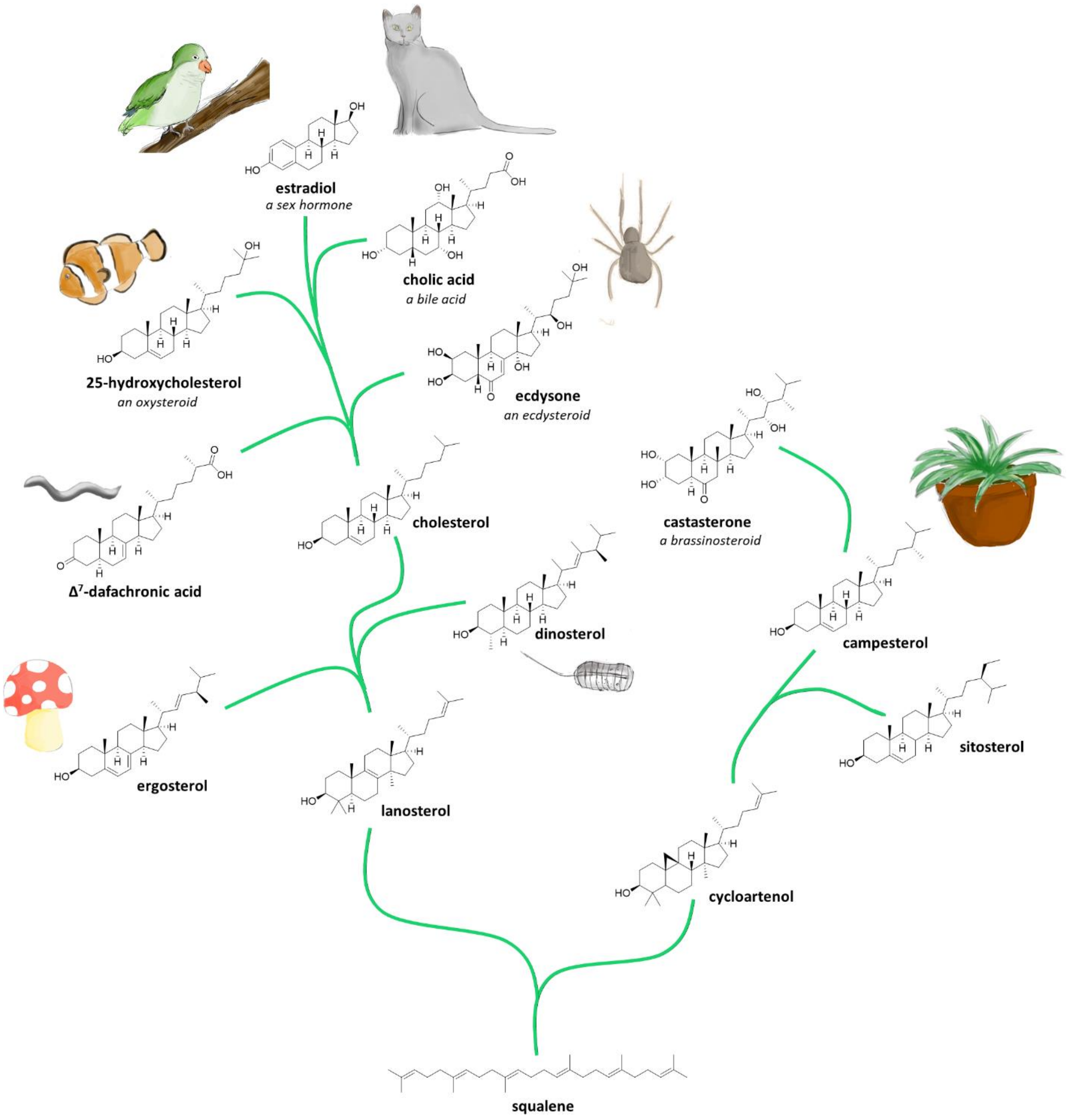

Figure 1. Diversification of Steroids from squalene.

Of note, despite the differences between these phylum-dependent steroidal metabolic pathways, it is clear that all members of this ancient family of polyhydroxylated lipids are synthesized *via* a cascade of highly conserved oxygen-dependent cytochrome P450 enzymes, which suggests that these pathways have evolved from a common unicellular ancestor (Jiang et al., 2010; Markov et al., 2009; Nelson, 2009; Thummel and Chory, 2002).

In some species, however, especially plants and marine invertebrates but also amphibians, a third phase may be present in which further transformations, which Diarey Tianero *et al*. coined as "diversity generating pathways", lead to a plethora of structurally complex steroids which are believed to have an ecological function, serving as a defensive chemical barrier against pathogens or predators (Tianero et al., 2016), and arose about 125 million years ago, as ecological networks became more intricated (Zhang et al., 2020). Several thousands of such steroids are currently known, although their biosynthetic origin remains to be established in most cases.

Given the aforementioned facts, steroids reveal a very suitable family of metabolites for performing a comparative analysis of their properties by using chemoinformatic tools in order to unravel the likely connection between structural properties of small organic biogenic molecules and metabolic evolution.

2. RESULTS AND DISCUSSION

A database of natural steroids was curated and annotated accordingly to the biosynthetic phases to which they belong. The first group comprised membrane sterols found in animals, fungi, and plants, such as cholesterol, ergosterol and sitosterol, and their biosynthetic precursors from squalene (S-steroids). The second group included steroids which have endogenous signaling (hormonal) roles, along with their biosynthetic precursors (H-steroids), whereas the third group (M-steroids) is a representative random sample of steroids from all phyla that can be classified as specialized metabolites. In total, the database included 479 compounds: 42 S-steroids (S001 to S0042), 159 H-steroids (H001 to H159) and 281 M-steroids (M001 to M281).

Sixty-four molecular descriptors were calculated for every compound. These molecular descriptors were selected in order to cover a wide range of molecular properties such as elemental analysis, hydrogen bond donor and acceptor, partitioning and distribution, global topological indices based on 2D-molecular graphs, geometric, ring and chain, and molecular complexity properties (for a list of the selected descriptors, see the Supplementary Information).

In this context, each steroid in the database could be considered as a point in a 64-dimensional chemical space defined by the selected molecular descriptors. Thus, the position of a given compound in such space reflects its physicochemical properties. Given our interest in analyzing how the members of the three families, classified according to their biosynthetic relationship, are distributed in this chemical space, we performed a Principal Component Analysis (PCA), a multivariate statistical method for variable reduction. This method consists in the creation of a new set of variables –called principal components– that are linear combinations of the original variables (orthogonal to each other), which allows the visualization of multidimensional data by using 2D-scatter plots with minimal loss of information of the original set of variables.(Härdle and Simar, 2015)

Through this analysis, we found that the first four principal components (PC1, PC2, PC3 and PC4) retain 43.1%, 13.9%, 9.9% and 9.5% of the total variance, respectively. It is remarkable to note that only four of the sixty-four principal components can explain 76.4% of the total variance of the original dataset (for further information, see the scree plot in the Supplementary Information).

Moreover, it is possible to understand how each original descriptor contributes to the new principal components. For example, PC1 is related mainly to molecular size, volume, and weight, whereas PC2 is related principally to planarity and molecular complexity. On the other hand, PC3 and PC4 share contributions from complexity, hydrophobicity, polarity, and hydrogen bond donor-acceptor properties (for more details, see the contribution plots in the Supplementary Information).

Figure 2 shows the PCA 2D-scatter plots of the four principal components, ordered from highest to lowest percentage of explained variances: (a) PC2 vs. PC1 (57.0%), (b) PC3 vs. PC1 (53.0%), (c) PC4 vs. PC1 (52.6%), (d) PC3 vs. PC2 (23.8%), (e) PC4 vs. PC2 (23.4%) and (f) PC4 vs. PC3 (19.4%). The dots correspond to the position of the compounds in the chemical space defined by the principal components, and the colored ellipses correspond to the concentration ellipses of 0.95 level.

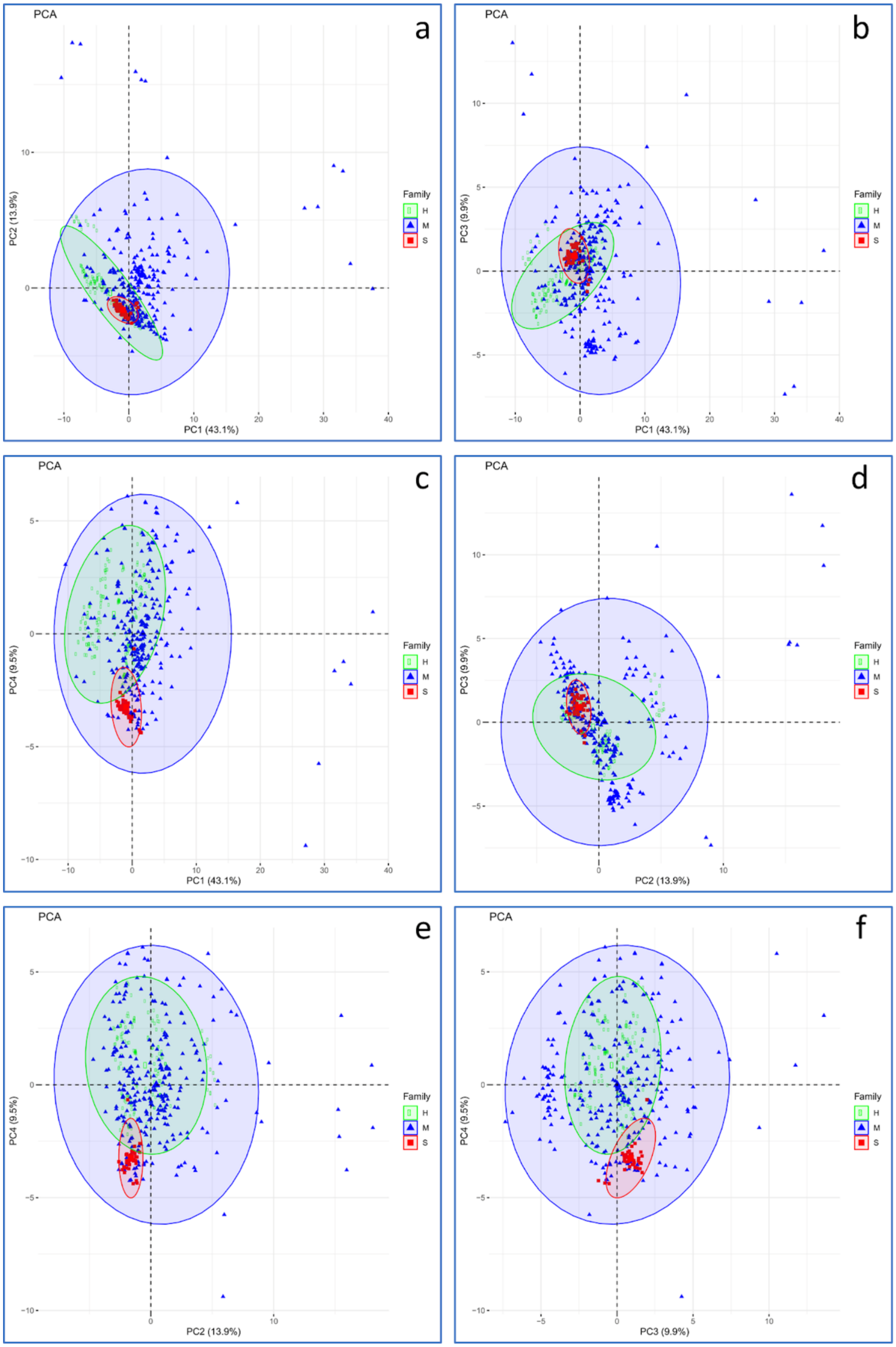

Figure 2. PCA 2D-scatter plots ordered according to the percentage of explained variances: (a) PC2-PC1 (57.0%), (b) PC3-PC1 (53.0%), (c) PC4-PC1 (52.6%), (d) PC3-PC2 (23.8%), (e) PC4-PC2 (23.4%) and (f) PC4-PC3 (19.4%). The dots correspond to the position of the compounds in the chemical space defined by the pair of principal components and the colored ellipses correspond to the concentration ellipses of 0.95 level (the 0.95 concentration ellipse is expected to enclose 95% of the data points, according to a bivariate normal distribution).

The PCA 2D-scatter plots show that S-steroids, H-steroids, and M-steroids are not equally distributed across the chemical space. A visual inspection shows that S-steroids occupy a restricted region, which means that they are more similar to one another, in contrast to H-steroids, which span through a wider area, thus showing greater diversity. Moreover, M-steroids show the largest variability in their properties. This variability can be properly quantified by calculating the standard deviations (SD) in every component, which are depicted in Figure 3a.

One of the long-standing models aimed to explain the evolution of biosynthetic pathways is the Granick hypothesis (Granick, 1957), whose central assumption is that the biosynthesis of many end-products could be explained by the forward evolution from relatively simple precursors. A prediction of the model is that the simpler compounds predated the complex ones (Scossa and Fernie, 2020), a trend that clearly emerges from our results. On the other hand, Figure 2 shows that the three families do not only differ in variability, but also in the relative position within the chemical space, which means that they have distinctive mean properties. Aiming to gain more insight into these differences, we constructed box-and-whisker plots for the 64 molecular descriptors (see Supplementary Information). Figure 3b shows the boxplots of eight selected properties and descriptors which highlight the differences and similarities among S-, H- and M-steroids.

At first sight, the narrower width of all the boxplots belonging to the S-steroids shows that their physicochemical properties are very similar. This restricted variability, which is in line with the lower standard deviations found in the PCA analysis, also translates into common biological roles. In this sense, it is known, for example, that replacing ergosterol in *S. cerevisiae* with the plant sterol campesterol or the animal sterol cholesterol leads to viable cells. Souza *et al.*

suggest that some basic functions of sterols linked to their properties, such as the ability to form membrane microdomains, have been retained along evolution, leaving little room for major changes in their structures (Souza et al., 2011). On the other hand, S-steroids have a low median value of oxygen atoms (one per molecule) when compared to H- and M-steroids, which have a median value of four per molecule, probably reflecting their appearance in an oxygen-rich atmosphere (Jiang et al., 2010).

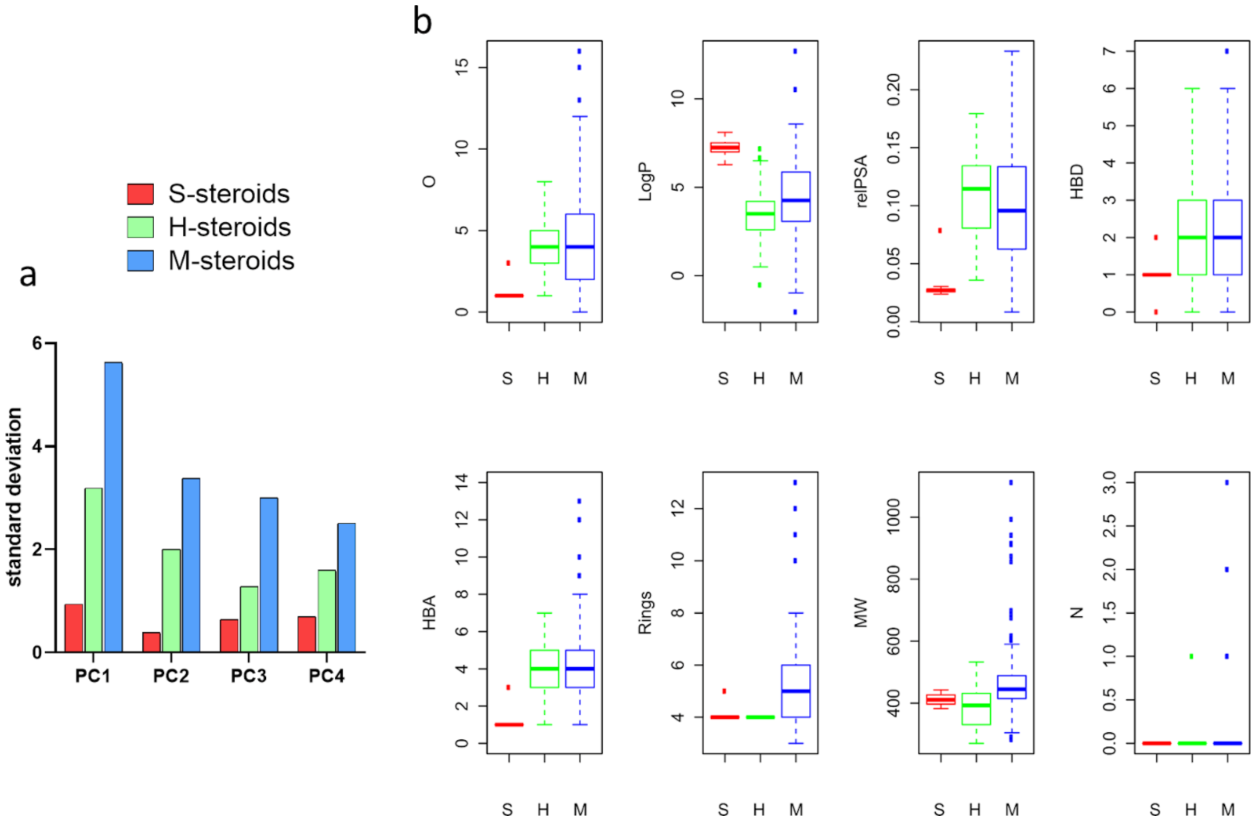

Figure 3. **a.** Standard deviations of principal components PC1, PC2, PC3 and PC4 for the three families of compounds. **b.** Comparative box-and-whisker plots for S-, H- and M-steroids illustrating the distribution of values for descriptors discussed in the text.

Other distinctive feature of S-steroids, which is directly related to oxygen content, is their much higher lipophilicity, as evidenced, for example, by their logP median values (7.2 vs. 3.5 for H-steroids and 4.3 for M-steroids). Likewise, the median relative polar surface area (relPSA, defined as the theoretical polar area normalized by the Van der Waals surface area) increases from 0.027 to 0.114 and 0.096 for S-, H- and M- steroids, respectively.

As stated before, the oxidation products of cholesterol and other membrane sterols have been selected by nature to act as signalling molecules, those which we encompassed as H-steroids. These molecules allowed multicellular organisms to respond to environmental stimuli and regulate their homeostasis, development, and reproduction. Evidence suggests that ancient oxygen-dependent cytochrome P450 enzymes, whose functions were to hydroxylate lipophilic xenobiotics (Baker, 2005), also oxidated membrane sterols to yield polyoxigenated steroids able both to traverse lipidic membranes and to act as ligands of receptors, a common feature of all steroid hormones (Baker et al., 2015; Markov et al., 2009). This increasing number of oxygenated moieties improved binding ability through hydrogen bond interactions: the median numbers of hydrogen bond donors (HBD) and acceptors (HBA) rise from one in S-steroids to two and four in H-steroids, respectively, as shown in Figure 3b.

In some species, S-steroids and H-steroids serve as intermediates for the biosynthesis of specialyzed metabolites (M-steroids) in which the steroidal skeleton may be expanded from the 6-6-6-5 tetracyclic framework to more complex polycyclic systems. The new rings can either be fussed to the parent skeleton (e.g. in tomatidine and diosgenin) or as substituents (e.g. digitoxigenin). In fact, the median value of the number of rings in M-steroids is five for the database analyzed in this work, but contains compounds with up to thirteen rings, generated by more radical transformations such as dimerizations, which also explain the presence of high molecular weight compounds (e.g. cephalostatins) among M-steroids (Chart 1).

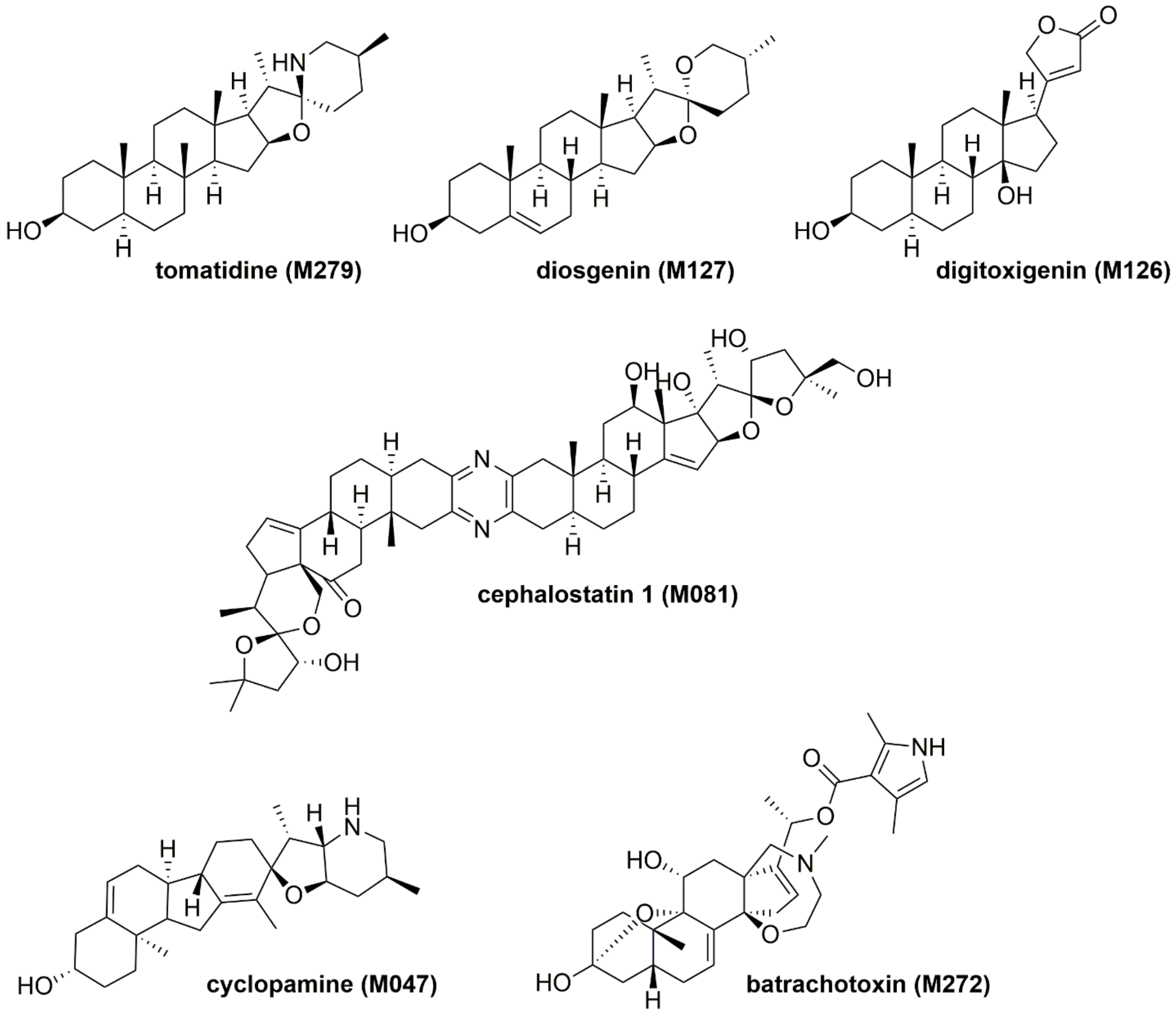

Chart 1. Some representative M-steroids.

Worth pointing out, natural steroids are not rich in nitrogen when compared to other natural products, which may reach values of sixty N atoms per molecule (Shang et al., 2018). S-steroids lack nitrogen atoms, whereas the only N-containing H-steroids are bile acids conjugated with amino acids (glycochenodeoxycholic, glycocholic, taurocholic and taurochenodeoxycholic acids). In agreement, our database of 281 M-steroids contains steroids having three N atoms at most.

M-steroids are natural products that do not seem to have an endogenous signaling function in its source organism, but usually serve as a defensive mechanism against a wide range of pathogens and predators. For example, the well-studied steroidal alkaloid cyclopamine, produced by *Veratrum californicum*, causes fatal birth defects when cattle feed upon this plant by binding to the protein Smoothened, thus disrupting the Sonic hedgehog pathway (Chen et al., 2002). Other examples of polycyclic nitrogenated steroids include batrachotoxin, a neurotoxic compound acting on sodium ion channels which is found in frogs of the genus *Phyllobates* (Li et al., 2002) (Chart 1). Despite their structural diversity, M-steroids share the common feature of being more complex and diverse than the steroids from which they derive, which is reflected by the wider boxplots and the presence of many outliers for the relevant molecular descriptors. In this sense, it is interesting to recall that according to Firn and Jones evolution has favoured mutations leading to metabolic traits that enhanced chemical diversity generation, as the more novel compounds are produced by an organism after mutation, the higher chances that some of these compounds will contain useful properties that will help increase the fitness of the producer (Firn and Jones, 2003). In some sense, the diversity-oriented synthesis strategies developed by chemists for drug discovery campaigns somehow mimic nature.

Whereas PCA is best suited to grasp the distribution of a given group of compounds in the chemical space, a cluster analysis can provide a deeper insight into the similarity of the members of such group by splitting them in clusters of molecules with related properties. Thus, we conducted a Hierarchical Cluster Analysis (HCA) of the database with the Ward's minimum variance method (Härdle and Simar, 2015) using the same 64 molecular descriptors. We visualized the HCA results through a circular dendrogram (Figure 4), in which the alphanumeric codes of S-, H- and M-steroids are colored in red, green, and blue, respectively.

At a glance, we noticed the presence of three main clusters (highlighted in yellow, orange and violet). In order to check if these clusters represent true structure or they are just an artifact of the clustering algorithm, we calculated the Jaccard coefficients (J_C) as cluster-wise measure of cluster stability (Hennig, 2008, 2007). The values for J_C resulted to be 0.98, 0.91 and 0.96 for the yellow, orange and violet clusters, respectively, showing their high stability (Zumel and Mount, 2014).

The yellow cluster contains only six members, all of them being M-steroids corresponding to high molecular weight outliers in the Mw boxplot of Figure 3b. A detailed inspection of the dendrogram reveals that the orange cluster contains only H- and M-steroids. Interestingly, every H-steroid of this cluster has a vertebrate origin; moreover, a more detailed analysis shows that these vertebrate steroids seem to be also divided in sub-clusters according to their biological role such as sexual hormones and bile acids.

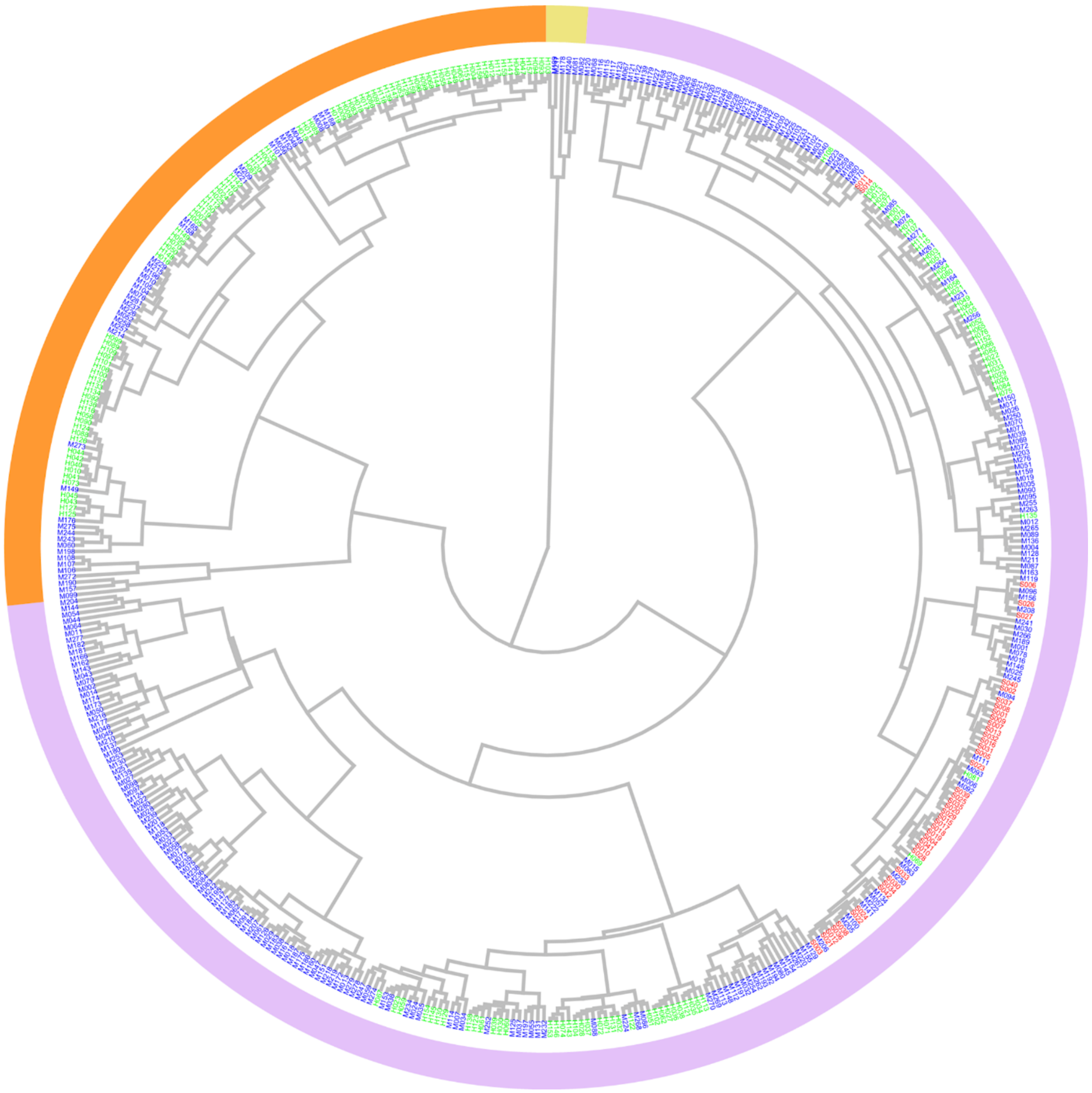

Figure 4. Circular dendrogram of S-steroids (red), H-steroids (green) and M-steroids (blue). The three principal clusters are colored in yellow, orange and violet.

Finally, the largest violet cluster shows the greater diversity, containing most of the M-steroids, all the sterols and their precursors (S-steroids) and all the H-steroids from invertebrate animals and plants, along with a small group of vertebrate H-steroids such as oxysterols. These H-steroids are not randomly mixed but also seem to be distributed in sub-clusters according to phylum and biological function, like the H-steroids belonging to the orange cluster.

At this point, it is clear that an HCA based on comparing simple structural and physicochemical descriptors calculated from the 2D-structures of the database was able to cluster the set of H-steroids into groups containing biologically related compounds. This intriguing fact led us to perform a further HCA focusing exclusively on the H-steroids, which was also depicted as a circular dendrogram (Figure 5a). This analysis confirmed a significant correlation between the biological role of the compounds and the internal structure of the dendrogram. In this sense, a clockwise inspection around the graphic allows to identify nine distinctive clusters, most of them related to different biological functions. For example, the dark green and light blue clusters only contain estrogens and ecdysteroids, respectively. Once again, the resulting clusters showed a high stability according to their calculated J_C values (caption of Figure 5a).

To dig deeper into these noteworthy results, we clustered the same set of compounds using the k-means method, an alternative non-hierarchical unsupervised machine learning technique in which the number of clusters (k) is defined *a priori* (Kubat, 2017). Figure 5b consistently shows that when the analysis is performed for k = 9, the resulting clusters are identical to the clusters found in the HCA.

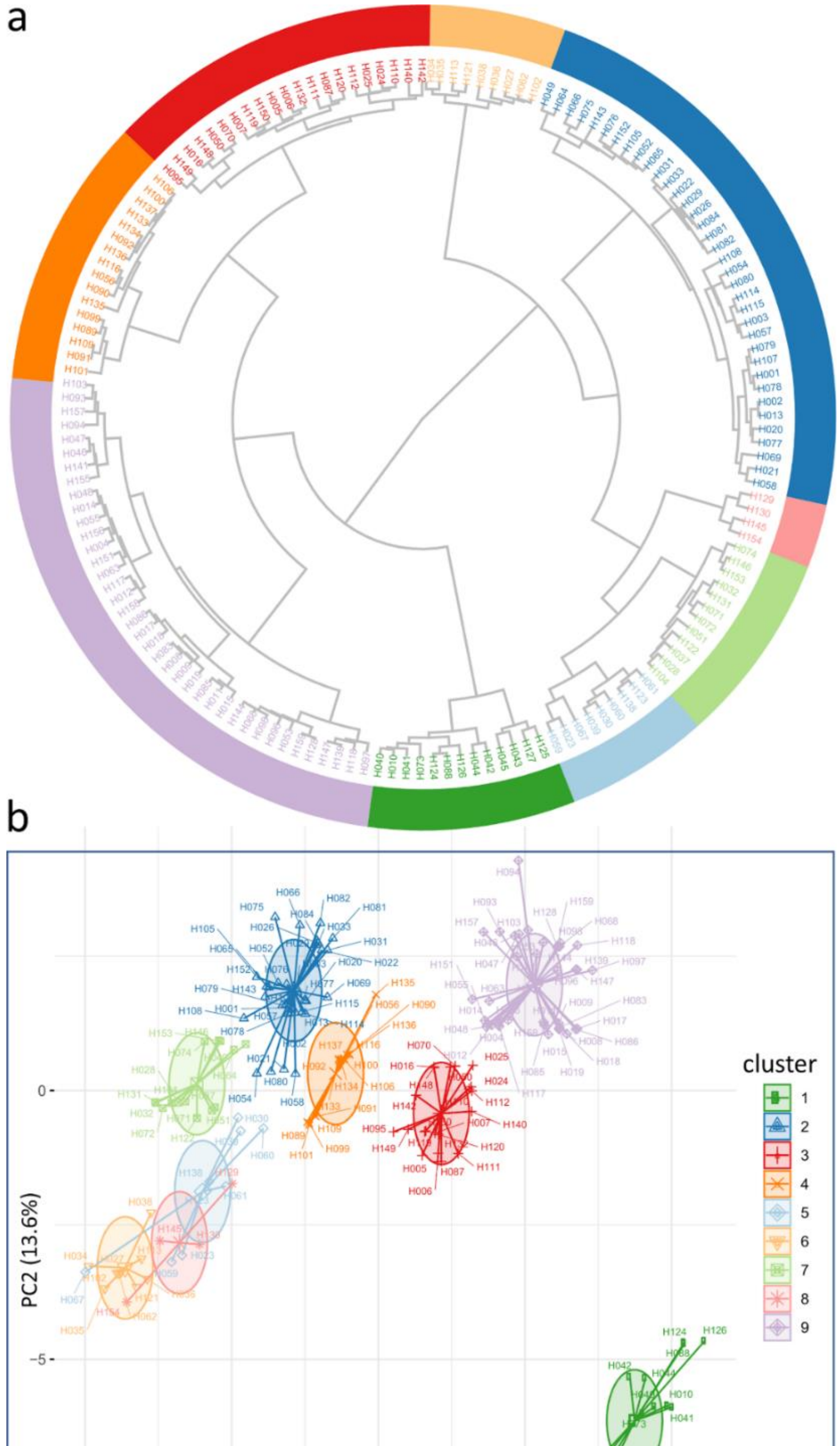

Figure 5: (a) Circular dendrogram of the H-steroids with the nine principal clusters. J_C values are 0.94 (#1); 0.98 (#2); 0.88 (#3); 0.99 (#4); 0.78 (#5); 0.99 (#6); 0.80 (#7); 0.91 (#8); 0.84 (#9). (b) PCA cluster plot with nine clusters obtained with k-means clustering method. The same coloring for each cluster was used in both graphics.

As a most noticeable result, estrogens do not only conform, as seen in the dendrogram, a defined cluster (cluster #1), but also lie far from the rest of the H-steroids in the chemical space, probably because estrogens are the only H-steroids with an aromatic ring. (e.g. estrone in Chart 2).

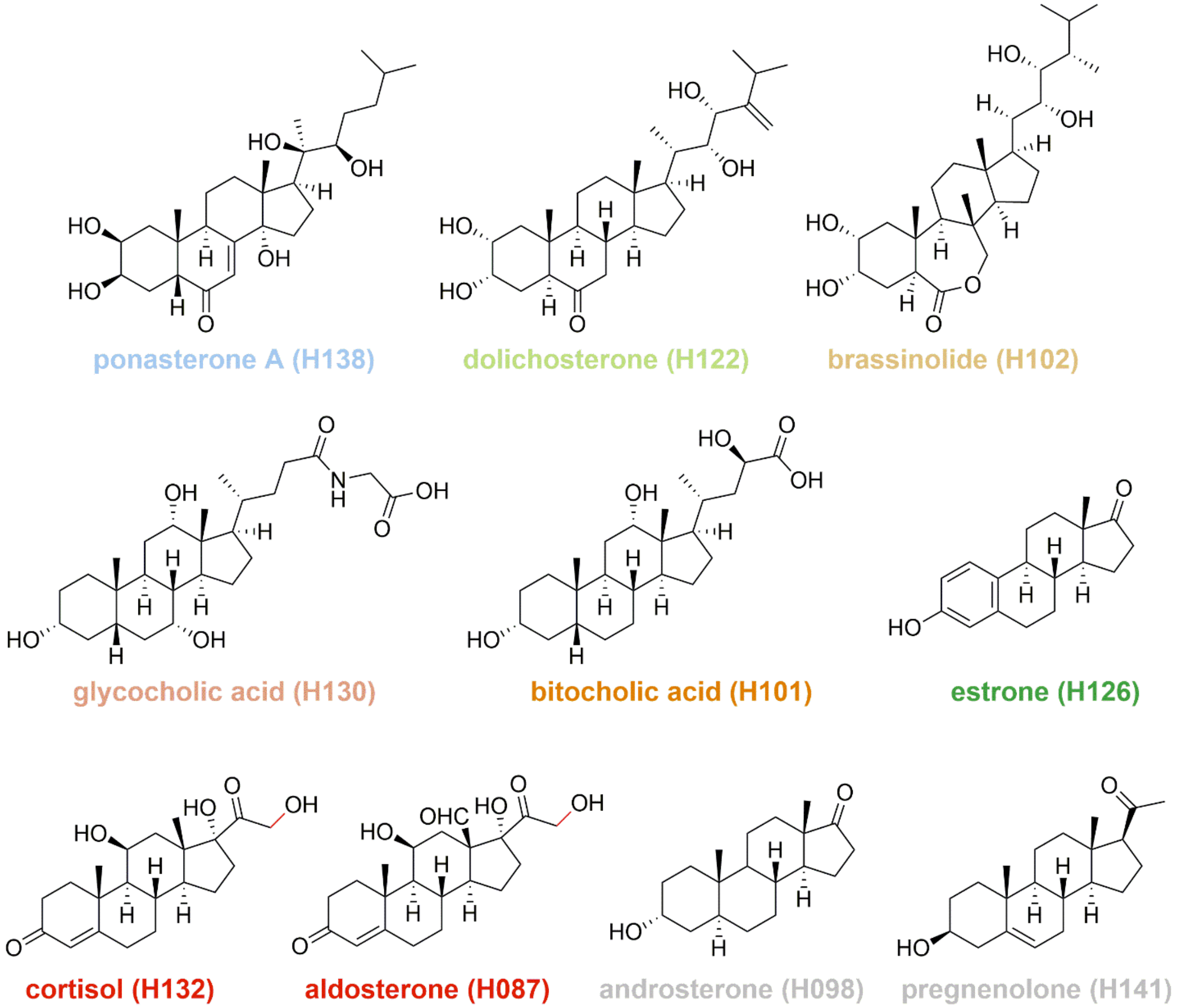

Chart 2. Representative members of clusters #1 to #9.

Ecdysteroids, the moulting steroidal hormones in arthropods, also constitute a separate cluster (cluster #5). In contrast, plant hormones brassinosteroids are split in two different clusters (clusters #6 and #7). A detailed analysis shows that cluster #7 consists of brassinosteroids having a six-membered B ring (such as dolichosterone in Chart 2), while brassinosteroids in cluster #6 bear a 7-oxalactone functionality in ring B (e.g. brassinolide). Although ecdysteroids and brassinosteroids are hormones of very distant phylogenetic organisms (Thummel and Chory, 2002), they are usually considered structurally very similar families: for example, ponasterone A (Chart 2) has an A ring hydroxylated at carbons 2 and 3, a 6-keto group and also a dihydroxylated side chain, common features also present in many brassinosteroids. Even when they are close in the chemical space, the analysis performed in this work found enough differences between these families of compounds to clearly separate ecdysteroids from brassinosteroids into different clusters.

In addition, bile acids also split into two different clusters (clusters #4 and #8), with the smallest one (cluster #8) containing four bile acids which are conjugated to amino acids. Finally, cluster #9 comprises most of the androgens and progestagens (such as androsterone and pregnenolone) and cluster #3 includes most of the corticosteroids (e.g. cortisol and aldosterone).

Even if the remaining cluster #2 cannot be easily associated to a defined biological role, a biosynthetic pathway map (Figure 5) in which H-steroids were colored according to the cluster they belong revealed that compounds of cluster #2 share a common relative position within the biosynthetic map, as they are early biosynthetic intermediates closer to the S-steroids than the rest of the H-steroids.

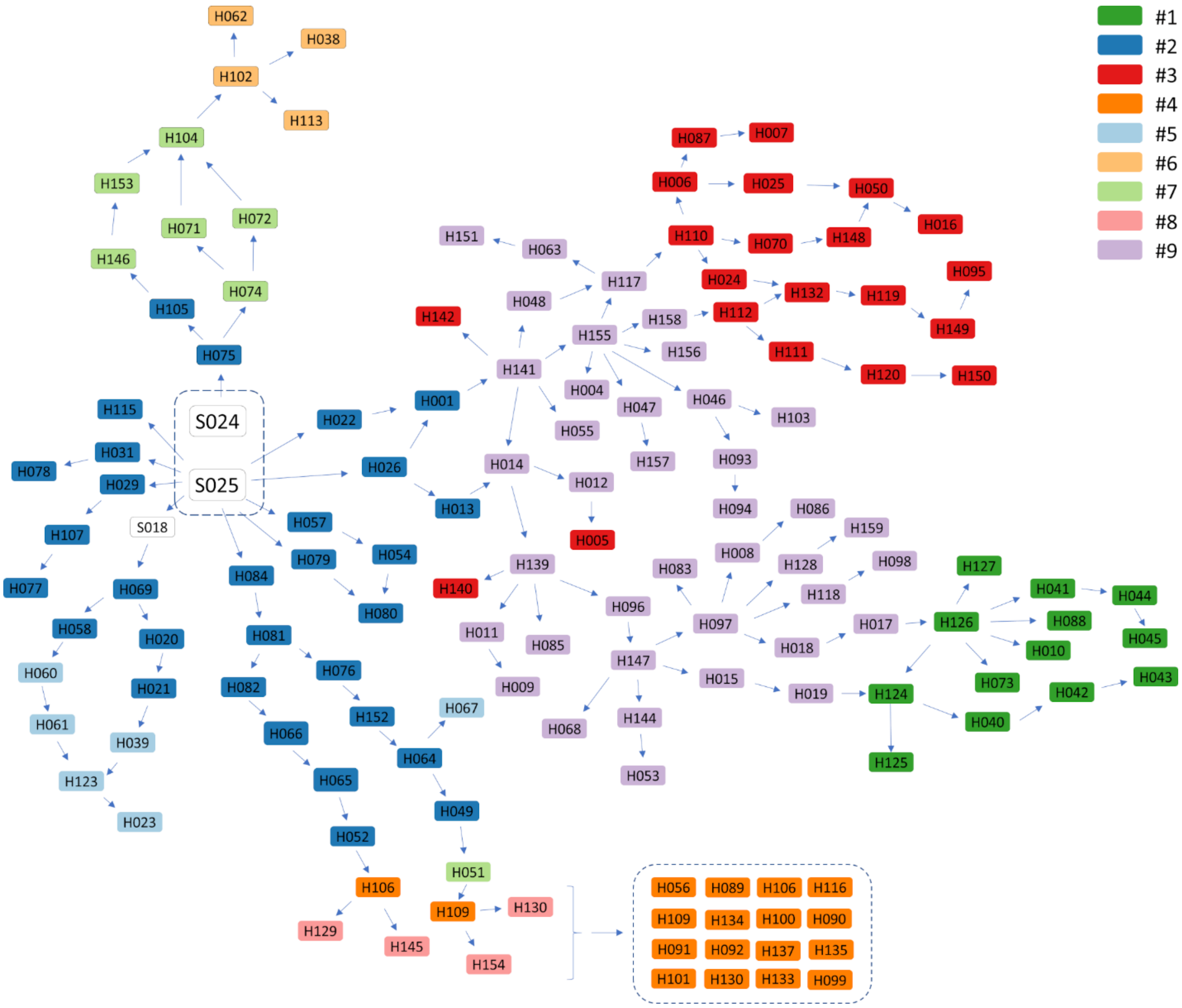

Figure 6. Biosynthetic pathways of the H-steroids colored according to the cluster to which each individual compound belongs.

In addition, Figure 6 clearly shows that the arrangement of the rest of the compounds in defined regions of the biosynthetic pathway map also closely correlates with the cluster they belong to. As a result, some interesting trends can be unveiled: it is known that cleavage of the side chain present in cholesterol, catalyzed by the enzyme CYP11A, is only found in vertebrates, and that these steroidogenic pathways producing glucocorticoids and reproductive steroids (included in clusters #1, #3 and #9) are several hundred million years old (Goldstone et al., 2016). Our analysis shows that this unique biosynthetic step introduces such important structural changes on the properties so as to establish a well-defined frontier between cluster #2 and the aforementioned clusters. In a similar way, aromatization of the six-membered A ring of androgens (included in cluster #9) by the enzyme CYP19A1 leads to a set of compounds (estrogens in cluster #1) that lie far from the rest of H-steroids in the chemical space.

In recent work on the evolution of steroidogenic enzymes and the timing of their appearance, Markov *et al*. suggest that the first pathway in appearing led to the synthesis of oxysterols, followed by dafachronic acids, ecdysones, and progestagens, and that most of the vertebrate bile acids, sex, and adrenal steroids surfaced during a later phase of diversification (Markov et al., 2017). Our findings also indicate that such diversification of enzymatic pathways paralleled the broadening of the properties of steroids they produce, allowing the emergence of more complex life forms able to regulate their life cycles more efficiently, but also to better fit to the environment. An important example is the evolution of the salt and water conservation mechanisms mediated by aldosterone which granted terrestrial vertebrates to conquer land (Rossier et al., 2015).

As stated before, modern statistical bioinformatic approaches targeted to the comparative analysis of genomes are being used to detect signatures of natural selection at the gene and population level, as an attempt to understand the origin of primordial metabolism and its expansion to form the metabolic networks existing nowadays, with impressive results. Nonetheless, these studies are mainly centered on genes and the proteins they encode, somehow neglecting the small organic chemicals that are, after all, central players in life processes. On the other hand, chemoinformatic techniques are nowadays at the roots of drug design and medicinal chemistry, areas in which assessing the complexity and diversity of compounds is paramount but are seldomly used to address basic biological questions. Only in recent times these techniques have been applied to the study of metabolites and natural products (Ertl and Schuhmann, 2019; González-Medina and Medina-Franco, 2019; Saldívar-González et al., 2019; Shang et al., 2018).

3. CONCLUSIONS

In this work, we have applied chemoinformatic tools to analyze, as a proof of concept, a set of natural steroids and found that even from simple physicochemical and topological properties derived from their molecular graphs, several traits that evolution has imprinted at the metabolite level can be unveiled, suggesting that natural selection also acts on molecular properties as advanced by Firn and Jones.

In this sense, our results clearly show that sterols, the primal steroids that first appeared, have more conserved properties and that, from then on, more complex compounds with increasingly diverse properties have emerged, which is in line with some models that were developed for explaining the evolution of metabolism. It is foreseeable that as more biogenic small organic compounds are discovered and their biosynthetic origins are disclosed, these chemoinformatic approaches could complement the usual genome-based studies, thus contributing to a better understanding of the complex chemical framework of life.

4. METHODS

4.1. Database curation

Structures of S- and H- steroids were extracted from the KEGG PATHWAY database (Kanehisa, 2000) and PubChem (Kim et al., 2019). Structures of M-steroids were collected from the *Dictionary of Steroids* (Hill et al., 1991) and from reports published in the *Journal of Natural Products* between years 2000 and 2019. Structures were compiled in isomeric SMILES format, and were downloaded from databases when available, or generated from ChemDraw drawings (ChemDraw 15.0.0.106, PerkinElmer Informatics).

4.2. Calculation of the molecular descriptors and statistical analysis

Calculation of the 64 physicochemical descriptors for S-, H- and M- steroids was performed with ChemAxon´s JChem for Excel (release 20.11.0.644, 2020, ChemAxon; http://www.chemaxon.com). Statistical analysis, unsupervised machine learning techniques (dimensionality reduction and clustering) and data visualization were carried out with R version 3.6.1 using the packages *factoextra*, *factoMineR*, *dendextend, fpc* and *ggplot2*. The database with the calculated molecular descriptors and the R scripts are available in the Suplemmentary Information.

CONFLICTS OF INTEREST

The authors have no conflicts of interest to declare.

AUTHOR CONTRIBUTION

M.O and J.A.R. conceived the presented ideas. S.N.S and S.L.O. curated the database M.O performed the calculations. All authors discussed the results and contributed to the final manuscript.

ACKNOWLEDGMENTS

This work was supported by grants from Universidad de Buenos Aires (UBACyT 20020170100679BA and 20020170100247BA). We are grateful to Ms. Maria Marta Rancez for her assistance with English language editing and to Miss Ana Ramírez for helping with the preparation of the Figure 1.

Supplementary information

List of the sixty-four properties and molecular descriptors used in this work. Table containing the names and codes of the natural steroids collected in the database. Additional statistical results. R scipts. Database with calculated molecular descriptors.

REFERENCES

Baker, M.E., 2005. Xenobiotics and the evolution of multicellular animals: Emergence and diversification of ligand-activated transcription factors. Integr. Comp. Biol. 45, 172–178. https://doi.org/10.1093/icb/45.1.172

Baker, M.E., Nelson, D.R., Studer, R.A., 2015. Origin of the response to adrenal and sex steroids:

Roles of promiscuity and co-evolution of enzymes and steroid receptors. J. Steroid Biochem. Mol. Biol. 151, 12–24. https://doi.org/10.1016/j.jsbmb.2014.10.020

Bento, G., Ogawa, A., Sommer, R.J., 2010. Co-option of the hormone-signalling module dafachronic acid-DAF-12 in nematode evolution. Nature 466, 494–497. https://doi.org/10.1038/nature09164

Brown, A.J., Galea, A.M., 2010. Cholesterol as an evolutionary response to living with oxygen. Evolution (N. Y). 64, 2179–2183. https://doi.org/10.1111/j.1558-5646.2010.01011.x

Chen, J.K., Taipale, J., Cooper, M.K., Beachy, P.A., 2002. Inhibition of Hedgehog signaling by direct binding of cyclopamine to Smoothened. Genes Dev. 16, 2743–8. https://doi.org/10.1101/gad.1025302

Dotson, R.J., Smith, C.R., Bueche, K., Angles, G., Pias, S.C., 2017. Influence of Cholesterol on the Oxygen Permeability of Membranes: Insight from Atomistic Simulations. Biophys. J. 112, 2336–2347. https://doi.org/10.1016/j.bpj.2017.04.046

Ertl, P., Schuhmann, T., 2019. A Systematic Cheminformatics Analysis of Functional Groups Occurring in Natural Products. J. Nat. Prod. 82, 1258–1263. https://doi.org/10.1021/acs.jnatprod.8b01022

Fani, R., Fondi, M., 2009. Origin and evolution of metabolic pathways. Phys. Life Rev. 6, 23–52. https://doi.org/10.1016/j.plrev.2008.12.003

Firn, R.D., Jones, C.G., 2009. A Darwinian view of metabolism: Molecular properties determine fitness. J. Exp. Bot. 60, 719–726. https://doi.org/10.1093/jxb/erp002

Firn, R.D., Jones, C.G., 2003. Natural products - A simple model to explain chemical diversity.

Nat. Prod. Rep. 20, 382. https://doi.org/10.1039/b208815k

Firn, R.D., Jones, C.G., 2000. The evolution of secondary metabolism - A unifying model. Mol. Microbiol. 37, 989–994. https://doi.org/10.1046/j.1365-2958.2000.02098.x

Fonseca, E., Ruivo, R., Lopes-Marques, M., Zhang, H., Santos, M.M., Venkatesh, B., Castro, L.F.C., 2017. LXRα and LXRβ nuclear receptors evolved in the common ancestor of gnathostomes. Genome Biol. Evol. 9, 222–230. https://doi.org/10.1093/gbe/evw305

Galea, A.M., Brown, A.J., 2009. Special relationship between sterols and oxygen: Were sterols an adaptation to aerobic life? Free Radic. Biol. Med. 47, 880–889. https://doi.org/10.1016/j.freeradbiomed.2009.06.027

Gold, D.A., Caron, A., Fournier, G.P., Summons, R.E., 2017. Paleoproterozoic sterol biosynthesis and the rise of oxygen. Nature 543, 420–423. https://doi.org/10.1038/nature21412

Goldstone, J. V., Sundaramoorthy, M., Zhao, B., Waterman, M.R., Stegeman, J.J., Lamb, D.C., 2016. Genetic and structural analyses of cytochrome P450 hydroxylases in sex hormone biosynthesis: Sequential origin and subsequent coevolution. Mol. Phylogenet. Evol. 94, 676–687. https://doi.org/10.1016/j.ympev.2015.09.012

González-Medina, M., Medina-Franco, J.L., 2019. Chemical Diversity of Cyanobacterial Compounds: A Chemoinformatics Analysis. ACS Omega 4, 6229–6237. https://doi.org/10.1021/acsomega.9b00532

Granick, S., 1957. Speculations On The Origins And Evolution Of Photosynthesis. Ann. N. Y. Acad. Sci. 69, 292–308. https://doi.org/10.1111/j.1749-6632.1957.tb49665.x

Härdle, W.K., Simar, L., 2015. Applied Multivariate Statistical Analysis. Springer Berlin

Heidelberg, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-45171-7

Hennig, C., 2008. Dissolution point and isolation robustness: Robustness criteria for general cluster analysis methods. J. Multivar. Anal. 99, 1154–1176. https://doi.org/10.1016/j.jmva.2007.07.002

Hennig, C., 2007. Cluster-wise assessment of cluster stability. Comput. Stat. Data Anal. 52, 258–271. https://doi.org/10.1016/j.csda.2006.11.025

Hill, R.A., Makin, H.L.J., Kirk, D.N., Murphy, G.M., 1991. Dictionary of Steroids. Taylor & Francis.

Honda, Y., Ishiguro, W., Ogihara, M.H., Kataoka, H., Taylor, D.M., 2017. Identification and expression of nuclear receptor genes and ecdysteroid titers during nymphal development in the spider Agelena silvatica. Gen. Comp. Endocrinol. 247, 183–198. https://doi.org/10.1016/j.ygcen.2017.01.032

Jiang, Y.Y., Kong, D.X., Qin, T., Zhang, H.Y., 2010. How does oxygen rise drive evolution? Clues from oxygen-dependent biosynthesis of nuclear receptor ligands. Biochem. Biophys. Res. Commun. 391, 1158–1160. https://doi.org/10.1016/j.bbrc.2009.11.041

Kanehisa, M., 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27–30. https://doi.org/10.1093/nar/28.1.27

Khan, N., Shen, J., Chang, T.Y., Chang, C.C., Fung, P.C.W., Grinberg, O., Demidenko, E., Swartz, H., 2003. Plasma Membrane Cholesterol: A Possible Barrier to Intracellular Oxygen in Normal and Mutant CHO Cells Defective in Cholesterol Metabolism. Biochemistry 42, 23–29. https://doi.org/10.1021/bi026039t

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B.A., Thiessen, P.A., Yu, B., Zaslavsky, L., Zhang, J., Bolton, E.E., 2019. PubChem 2019 update: improved access to chemical data. Nucleic Acids Res. 47, D1102–D1109. https://doi.org/10.1093/nar/gky1033

Kroymann, J., 2011. Natural diversity and adaptation in plant secondary metabolism. Curr. Opin. Plant Biol. 14, 246–251. https://doi.org/10.1016/j.pbi.2011.03.021

Kubat, M., 2017. Unsupervised Learning, in: An Introduction to Machine Learning. Springer International Publishing, Cham, pp. 273–295. https://doi.org/10.1007/978-3-319-63913-0_14

Lednicer, D., 2011. Steroid Chemistry at a Glance, Chemistry At a Glance. Wiley.

Li, H.L., Hadid, D., Ragsdale, D.S., 2002. The batrachotoxin receptor on the voltage-gated sodium channel is guarded by the channel activation gate. Mol. Pharmacol. 61, 905–912. https://doi.org/10.1124/mol.61.4.905

Lu, Y., Jiang, J., Zhao, H., Han, X., Xiang, Y., Zhou, W., 2020. Clade-Specific Sterol Metabolites in Dinoflagellate Endosymbionts Are Associated with Coral Bleaching in Response to Environmental Cues. mSystems 5, 1–16. https://doi.org/10.1128/mSystems.00765-20

Markov, G. V., Gutierrez-Mazariegos, J., Pitrat, D., Billas, I.M.L., Bonneton, F., Moras, D., Hasserodt, J., Lecointre, G., Laudet, V., 2017. Origin of an ancient hormone/receptor couple revealed by resurrection of an ancestral estrogen. Sci. Adv. 3, 1–14. https://doi.org/10.1126/sciadv.1601778

Markov, G. V, Tavares, R., Dauphin-Villemant, C., Demeneix, B.A., Baker, M.E., Laudet, V., 2009. Independent elaboration of steroid hormone signaling pathways in metazoans. Proc Natl Acad Sci U S A 106, 11913–11918. https://doi.org/10.1073/pnas.0812138106

Nakagawa, Y., Henrich, V.C., 2009. Arthropod nuclear receptors and their role in molting. FEBS J. 276, 6128–6157. https://doi.org/10.1111/j.1742-4658.2009.07347.x

Nelson, D.R., 2009. The cytochrome P450 homepage. Hum. Genomics 4, 59–65. https://doi.org/10.1186/1479-7364-4-1-59

Niwa, R., Niwa, Y.S., 2014. Enzymes for ecdysteroid biosynthesis: Their biological functions in insects and beyond. Biosci. Biotechnol. Biochem. 78, 1283–1292. https://doi.org/10.1080/09168451.2014.942250

Popova, A. V, Velitchkova, M., Zanev, Y., n.d. Effect of membrane fluidity on photosynthetic oxygen production reactions. Z. Naturforsch. C. 62, 253–60.

Rossier, B.C., Baker, M.E., Studer, R.A., 2015. Epithelial sodium transport and its control by aldosterone: The story of our internal environment revisited. Physiol. Rev. 95, 297–340. https://doi.org/10.1152/physrev.00011.2014

Saldívar-González, F.I., Valli, M., Andricopulo, A.D., Da Silva Bolzani, V., Medina-Franco, J.L., 2019. Chemical Space and Diversity of the NuBBE Database: A Chemoinformatic Characterization. J. Chem. Inf. Model. 59, 74–85. https://doi.org/10.1021/acs.jcim.8b00619

Schaller, H., 2003. The role of sterols in plant growth and development. Prog. Lipid Res. 42, 163–175. https://doi.org/10.1016/S0163-7827(02)00047-4

Scossa, F., Fernie, A.R., 2020. The evolution of metabolism: How to test evolutionary hypotheses at the genomic level. Comput. Struct. Biotechnol. J. 18, 482–500. https://doi.org/10.1016/j.csbj.2020.02.009

Shang, J., Hu, B., Wang, J., Zhu, F., Kang, Y., Li, D., Sun, H., Kong, D.X., Hou, T., 2018.

Cheminformatic Insight into the Differences between Terrestrial and Marine Originated Natural Products. J. Chem. Inf. Model. 58, 1182–1193. https://doi.org/10.1021/acs.jcim.8b00125

Souza, C.M., Schwabe, T.M.E., Pichler, H., Ploier, B., Leitner, E., Guan, X.L., Wenk, M.R., Riezman, I., Riezman, H., 2011. A stable yeast strain efficiently producing cholesterol instead of ergosterol is functional for tryptophan uptake, but not weak organic acid resistance. Metab. Eng. 13, 555–569. https://doi.org/10.1016/j.ymben.2011.06.006

Tamura, T., Akihisa, T., Kokke, W., 1992. Naturally Occurring Sterols and Related Compounds from Plants, in: Physiology and Biochemistry of Sterols. AOCS Publishing, pp. 172–228. https://doi.org/10.1201/9781439821831.ch7

Thummel, C.S., Chory, J., 2002. Steroid signaling in plants and insects - Common themes, different pathways. Genes Dev. 16, 3113–3129. https://doi.org/10.1101/gad.1042102

Tianero, M.D., Pierce, E., Raghuraman, S., Sardar, D., McIntosh, J.A., Heemstra, J.R., Schonrock, Z., Covington, B.C., Maschek, J.A., Cox, J.E., Bachmann, B.O., Olivera, B.M., Ruffner, D.E., Schmidt, E.W., 2016. Metabolic model for diversity-generating biosynthesis. Proc. Natl. Acad. Sci. U. S. A. 113, 1772–1777. https://doi.org/10.1073/pnas.1525438113

Weete, J.D., Abril, M., Blackwell, M., 2010. Phylogenetic distribution of fungal sterols. PLoS One 5, 3–8. https://doi.org/10.1371/journal.pone.0010899

Widomska, J., Raguz, M., Subczynski, W.K., 2007. Oxygen permeability of the lipid bilayer membrane made of calf lens lipids. Biochim. Biophys. Acta - Biomembr. 1768, 2635–2645. https://doi.org/10.1016/j.bbamem.2007.06.018

Zhang, Y., Deng, T., Sun, L., Landis, J.B., Moore, M.J., Wang, H., Wang, Y., Hao, X., Chen, J., Li, S., Xu, M., Puno, P.-T., Raven, P.H., Sun, H., 2020. Phylogenetic patterns suggest frequent multiple origins of secondary metabolites across the seed plant “tree of life.” Natl. Sci. Rev. https://doi.org/10.1093/nsr/nwaa105

Zumel, N., Mount, J., 2014. Unsupervised methods, in: Practical Data Science with R. Manning Publications.