Inferring Biochemical Reactions and Metabolite Structures to Understand Metabolic Pathway Drift

Arnaud Belcour, Jean Girard, Méziane Aite, Ludovic Delage, Camille Trottier, Charlotte Marteau, Cédric J-J Leroux, Simon M. Dittami, Pierre Sauleau, Erwan Corre, et al.

To cite this version:
Arnaud Belcour, Jean Girard, Méziane Aite, Ludovic Delage, Camille Trottier, et al.. Inferring Biochemical Reactions and Metabolite Structures to Understand Metabolic Pathway Drift. iScience, 2020, 23 (2), pp.100849. 10.1016/j.isci.2020.100849. hal-01943880v2

HAL Id: hal-01943880
https://inria.hal.science/hal-01943880v2
Submitted on 24 Feb 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License
Inferring Biochemical Reactions and Metabolite Structures to Understand Metabolic Pathway Drift

Arnaud Belcour, Jean Girard, Méziane Aite, ..., Jonas Collén, Anne Siegel, Gabriel V. Markov

gabriel.markov@sb-roscoff.fr

HIGHLIGHTS
Combination of metabolite profiling and genome-scale metabolic networks analysis
New method to infer biochemical reactions and metabolites in promiscuous pathways
Red algal model for sterol and mycosporine-like amino acid biosynthesis pathways
Metabolic drift if pathway variation despite conserved initial and final metabolites
Inferring Biochemical Reactions and Metabolite Structures to Understand Metabolic Pathway Drift

Arnaud Belcour,1,6 Jean Girard,2,6 Méziane Aite,1,6 Ludovic Delage,2,6 Camille Trottier,1 Charlotte Marteau,3 Cédric Leroux,4 Simon M. Dittami,2 Pierre Sauleau,3 Erwan Corre,5 Jacques Nicolas,1 Catherine Boyen,2 Catherine Leblanc,2 Jonas Collén,2 Anne Siegel,1 and Gabriel V. Markov2,7,*

SUMMARY

Inferring genome-scale metabolic networks in emerging model organisms is challenged by incomplete biochemical knowledge and partial conservation of biochemical pathways during evolution. Therefore, specific bioinformatic tools are necessary to infer biochemical reactions and metabolic structures that can be checked experimentally. Using an integrative approach combining genomic and metabolomic data in the red algal model Chondrus crispus, we show that, even metabolic pathways considered as conserved, like sterols or mycosporine-like amino acid synthesis pathways, undergo substantial turnover. This phenomenon, here formally defined as “metabolic pathway drift,” is consistent with findings from other areas of evolutionary biology, indicating that a given phenotype can be conserved even if the underlying molecular mechanisms are changing. We present a proof of concept with a methodological approach to formalize the logical reasoning necessary to infer reactions and molecular structures, abstracting molecular transformations based on previous biochemical knowledge.

INTRODUCTION

Life is driven by a high diversity of metabolic processes, and each species or even strain may be characterized by its own metabolic particularities (e.g., Rhee et al., 2011). During evolutionary time and speciation processes, there are many ways that variations can be generated within metabolic pathways. Evolutionary models have been developed and experimentally tested to explain the arising of new pathways, but these efforts were focused on the activities of individual enzymes (Noda-Garcia et al., 2018). Changes in the metabolism may, however, also occur at a higher level of organization, notably an enzymatic replacement by non-orthologous genes encoding enzymes with identical biochemical function (Koonin et al., 1996; Figure 1, left side), or a change in enzyme order, which leads to different main biosynthetic intermediates (Figure 1, right side). This kind of variability, which we refer to as “metabolic drift” in this article, is possible due to substrate promiscuity of the enzymes and may be an important driver of evolution (Peracchi, 2018).

The concept of “drift” has previously been used in the field of animal comparative developmental biology, where it was already used to explain how morphologically similar structures can be maintained even if there are substantial variations in the molecular mechanisms underlying their formation (True and Haag, 2001). Here also, the use of “drift” is distinct from genetic drift, but, nevertheless, appropriate because chance and not selection explains how changes occur. The concept was more recently extended to plants, where such cases have been observed in leaf development (Townsley and Sinha, 2012). It was later exported to the fields of protein evolution (Hart et al., 2014) and gene expression evolution (Petit et al., 2016). Here, we hypothesize that the evolutionary concept of drift could also adequately explain the strict functional conservation of many metabolic pathways, despite variations in the underlying mechanisms (Figure 1).

To study metabolic drift by non-orthologous gene displacement (Figure 1, left part), classical comparative genomic approaches can generate hypotheses that can be experimentally checked using targeted metabolic profiling combined with inactivation of enzyme-encoding genes by reverse genetics (Markov et al., 2016; Sonawane et al., 2016). However, in case of drift by change in enzyme order (Figure 1, right part), it is necessary to combine both genomic and metabolomic data and to introduce a knowledge-based approach that implements reasoning in the manner of a biochemist. Such inter-disciplinary strategies...
have already been used to design experiments for the analysis of auxotrophic mutants in yeast (King et al., 2004) or for synthetic biology, where ab initio pathway inference is done to find a biosynthetic route that is not necessarily present in nature (Koch et al., 2017). We hypothesize that similar tools can also be used to group biochemical reaction variants based on shared ancestry and to infer undescribed pathways that may be present in emerging model species.

To test this hypothesis, we have implemented a semi-automatic analogy reasoning approach, which we use to study two distinct metabolic pathways, the sterol pathway and the mycosporine-like amino acid (MAA) pathway, in the red alga Chondrus crispus. Red algae are sufficiently distant from terrestrial plants to anticipate substantial metabolic drift compared with the known pathways in organisms such as Arabidopsis thaliana, yet C. crispus has been subject to biological studies for more than two centuries (Collen et al., 2014). Notably, its genome was sequenced and annotation was performed with a focus on metabolic features (Collen et al., 2013), and there is extensive literature available describing its metabolome (Young and Smith, 1958; Saito and Idler, 1966; Laycock and Craigie, 1977; Matsuhiro and Urzua, 1992; Karsten et al., 1998; Tasende, 2000; Krabs et al., 2004; Gaquerel et al., 2007; Banskota et al., 2014; Pina et al., 2014; Melo et al., 2015, Alcaide et al., 1968; Goldberg et al., 1982; Kremer and Kirst, 1982; Pettit et al., 1989; van Ginneken et al., 2011; Santos et al., 2015; Robertson et al., 2015; Athukorala et al., 2016; Belghiti et al., 2017; Guihéneuf et al., 2018; Lalegerie et al., 2019).

The sterol metabolism was chosen as one focus point, because there is extensive knowledge about this pathway at the comparative genomics level (Desmond and Gribaldo, 2009). Furthermore, analytical standards are available for different sterol molecules, enabling their identification by mass spectrometry (MS) (Sumner et al., 2007). MAA synthesis, on the other hand, involves combination of different building blocks, and analytical standards are lacking for this class of compounds, limiting metabolite identification to putative annotated compounds based on spectral similarity with spectral libraries (Sumner et al., 2007). Therefore the reconstruction of MAA pathway raised the problem of integrating unannotated compounds that were identified uniquely based on their m/z ratio.

By implementing a semi-automatic analogy reasoning approach that integrates both metabolite and genomic data, we here propose an exhaustive model for both metabolic pathways in C. crispus and provide strong indications for the importance of metabolic drift in shaping these pathways. We furthermore consolidated at least part of these hypotheses with targeted metabolite profiling of metabolic intermediates predicted in the models.
RESULTS
Genome-Scale Metabolic Network Enriches the Inferred Basic Integrated Metabolism for C. crispus

Genome-scale metabolic networks (GSMNs) are graph-based representations of enzymatic reactions assumed to occur in a given organism. In this framework, an enzymatic reaction denotes a chemical reaction that transforms one or several metabolic substrates into one or several metabolic products under the control of an enzyme that can be associated with a gene in the considered species.

We used the tool AuReMe (Automatic Reconstruction of Metabolic models) dedicated to “à la carte” reconstruction of GSMNs (Aite et al., 2018) to reconstruct a GSMN of C. crispus. This GSMN comprised in total 2,024 enzymatic or transport reactions (Table 1). Among them, 595 reactions were recovered from annotation-based searches from the C. crispus genome annotation (Collén et al., 2013) with the Pathway Tools suite (Karp et al., 2002) and the MetaCyc database (Caspi et al., 2016). In addition, 1,429 reactions were included in the network according to orthology evidences with protein sequences encoding enzyme reactions in the Arabidopsis thaliana GSMN (de Oliveira Dal’Molin et al., 2010), the Galdieria sulphuraria GSMN (based on genome data from Schönknecht et al., 2013), or the Ectocarpus siliculosus GSMN (Cormier et al., 2017; Aite et al., 2018). A biomass reaction was established based on the previous E. siliculosus data, defining a list of 33 compounds to be produced to consider the network functional (Pregent et al., 2014). According to this biomass reaction, the network was manually gap-filled to unblock the production of L-alpha-alanine with an alanine dehydrogenase reaction whose associated gene had been incorrectly annotated in the C. crispus genome. The predicted maximal growth rate was then 2.43 g.gDW⁻¹. h⁻¹ (gram per gram dry weight per hour). As shown in Table 1, the C. crispus GSMN is comparable in size with the E. siliculosus GSMN. A total of 254 pathways are complete, including those involved in central metabolism of carbohydrates, fatty acids, and amino acids, as well as those related with photosynthesis. The greatest contributor to the inferred reaction set was the phylogenetically close red microalga G. sulphuraria, which provided 1,361 reactions.

Building a Catalog of Evidenced Metabolites in C. crispus Using Metabolomic Data

To better understand the specificities of C. crispus, we built a catalog of metabolic compounds attested to be produced by the alga. To that goal, we reviewed the literature to collect experimental evidence of presence of all reported metabolites. To have more species-specific data on sterol and MAA synthesis we also experimentally tested for the presence of these compounds and their precursors in C. crispus using MS analysis. In this way, we assembled a set of 142 metabolites that are reported in Tables S1 and S2. Those metabolites broadly cover various classes of amino acids, carbohydrates, and lipids. We divided this dataset into two main categories: 85 database metabolites, which are already indexed into Metacyc (Table S1), and 57 orphan metabolites, not yet indexed (Table S2). In addition, we have acquired additional experimental data on two pathways for which molecules are in both categories: the sterol and the MAA pathways.

Our MS data confirmed previous findings and also pointed out possible additional precursors and intermediary metabolites in sterol and MAA biosynthesis pathways. More precisely, the results of these analyses of sterols are found in Table 2. In addition to confirming the presence of eight previously identified sterols (brassicasterol, campesterol, cholesterol, 7-dehydrocholesterol, desmosterol, lathosterol, β-sitosterol, and stigmasterol), we identified in C. crispus an immediate precursor of sterols, i.e., squalene (Figure S1). However, we did not find evidence for cycloeucalenol, ergosterol, fucosterol, and zymosterol, which are
intermediates present in other eukaryotes (Desmond and Gribaldo, 2009; Sonawane et al., 2016). We also did not find cycloartenol, contrary to a previous report in C. crispus using thin-layer chromatography (Alcaide et al., 1968). This negative finding is strengthened by the fact that we are able to identify the cycloartenol standard when added in algal extract (Figure S2).

Similar experiments were carried out for the MAA pathway, and corresponding data are summed up in Table 3. Using liquid chromatography-MS (LC-MS) profiling we confirmed the presence of six MAAs in C. crispus (see references in Table 3): asterina-330, palythene, palythin e, palythinol, porphyra-334, and shi-norine. In addition, we identified mycosporine-glycine in C. crispus. The relative abundance of MAAs varied between sampling dates (Figure 2), which is consistent with a report of MAA variation in the Galway Bay, Ireland (Guihe´neuf et al., 2018). We noticed that two unknown peaks potentially corresponding to MAAs were detected in our samples. These peaks exhibit m/z ratios consistent with peaks reported in a study on 40 red algae by Lalegerie et al. (2019). Specifically, we found a peak at m/z = 270.3 that does not match with any already identified candidate MAA, which we named MAA1, and a second one at m/z = 302.3, which we named MAA2.

Table 2. List of Sterols Profiled in This Study, and Comparisons with Previous Studies

For each compound, analytical parameters (retention time and m/z ratio) are given in Table S1.

Secondary Metabolism Evidenced by Metabolomic Data Is Only Partially Accurately Described in the C. crispus GSMN

To investigate the accuracy of GSMNs to describe the synthesis pathway of secondary metabolites of importance, we studied the capability of the C. crispus GSMN to describe synthesis pathways of metabolic
compounds whose presence have been evidenced in the alga. Among those 142 metabolites, only 85 (60%)
were indexed in the MetaCyc database version 20.5 (Table S1). The GSMN could provide synthesis path-
ways for only 59 metabolites already indexed in the Metacyc database. Those metabolites were amino
acids (19), fatty acids (12), aldehydes (6), aliphatic alcohols (3), carotenoids (3), ketones (2), carboxylic acids
(2), a halocarbon, a galactolipid, an oxylipin, a tetraps, an MAA, a nucleotide sugar, a methylketone,
and a polyol. The synthesis of the 29 remaining metabolites indexed in the MetaCyc database could not be
explained by any addition of enzymatic reactions in the database (failure of the exhaustive gap-filling pro-
cedure). The metabolites belonged to the following classes: fatty acids (8), sterols (8), alkanes (2), caroten-
oids (2), carrageens (2), aliphatic alcohols (2), an aldehyde, and a halocarbon. For the 57 metabolites that
had not yet been indexed in MetaCyc (Table S2), it was impossible to generate hypotheses on their synthe-
sis pathways. There were galactolipids (12), oxylipins (9), fatty acids (8), MAAs (6), alcohols (6), unconven-
tional amino acids (3), tetrapyrroles (3), alkanes (2), carboxylic acids (2), ketones (2), a heteroside, a sterol,
a phospholipid, and an aldehyde.

### Building a Knowledge Base of Enzymatic Reactions and Molecules in Sterol and MAA

To enable the incorporation of the orphan metabolites that were not yet in MetaCyc and derive possible
synthesis pathways for sterols and MAAs, we built two knowledge bases describing the existing knowledge
about known enzymatic reactions and molecules involving sterols or MAAs (available on https://github.
com/pathmodel/pathmodel). They are encoded in the Pathmodel datafiles sterol_pwy.lp and MAA_p-
wy.lp, both accessible in the pathmodel/data folder of the Github repository. In these knowledge bases,
molecules are described by **atoms** (identified by a number and atom types) and **bonds** (identified by
atom numbers and bond type) as highlighted in green on Figure 3. Atom numbers were assigned manually
to ensure consistency between molecules from the same family and followed IUPAC conventions when ex-
isting (Moss, 1989). Molecules can be automatically associated with a theoretical m/z ratio, calculated using
their chemical formula, as described in the program MZComputation.lp.

Enzymatic reactions, denoted by reaction, model the link between two molecules (a reactant and a product,
e.g., reaction(rxn_4243, ”sitosterol”, “stigmasterol”). They are associated with cross-references to
MetaCyc pathway IDs when possible.

The MAA knowledge base (Figure 4) contained 13 enzymatic reactions involving 12 molecules. It first con-
tained the shionine biosynthesis pathway (MetaCyc: PWY-7751), corresponding to the best understood

### Table 3. List of Mycosporine-like Amino Acids Identified in This Study and Comparisons with Previous Studies

| Analyzed Compounds | Molecular Formula | Found in This Study | Previous Evidence                       |
|--------------------|-------------------|---------------------|-----------------------------------------|
| Asterina-330       | C_{12}H_{20}N_{2}O_{6} | Yes                 | Athukorala et al. (2016) (LC-MS-MS); Guihéneuf et al. (2018) (LC-MS) |
| MAA1 compatible with m/z 270.272 | | Yes | Lalegerie et al. (2019) (HPLC) |
| MAA2 compatible with m/z = 302.3117 | | Yes | Lalegerie et al. (2019) (HPLC) |
| Mycosporine-glycine | C_{10}H_{15}NO_{6} | Yes                 | not reported |
| Palythene          | C_{12}H_{18}N_{2}O_{2} | Yes                 | Karsten et al. (1998) (UV + LC-MS); Athukorala et al. (2016) (LC-MS-MS); Guihéneuf et al. (2018) (LC-MS) |
| Usujirene/Palythene | C_{13}H_{20}N_{2}O_{5} | Yes                 | Karsten et al. (1998) (UV + LC-MS) |
| Palythinol         | C_{13}H_{22}N_{2}O_{5} (m/z = 302.3117) | No                  | Karsten et al. (1998) (UV + LC-MS), Athukorala et al. (2016) (LC-MS-MS) |
| Porphyra-334       | C_{14}H_{22}N_{2}O_{5} | Yes                 | Athukorala et al. (2016) (LC-MS-MS) |
| Shinorine          | C_{13}H_{20}N_{2}O_{6} | Yes                 | Karsten et al. (1998) (UV + LC-MS), Athukorala et al. (2016) (LC-MS-MS) |

HPLC, high-performance liquid chromatography.
part of the pathway (Shick and Dunlap, 2002). We also encoded in the knowledge base an extended version of the amino acid C3-transfer reaction (MetaCyc: RXN-17371) to incorporate the hypothesis that the enzyme performing the C3-transfer of serine can also perform the C3-transfer of threonine, leading to porphyra-334 (Brawley et al., 2017). We also incorporated additional reactions hypothesized in the literature (Carreto and Carignan, 2011), for which either the substrate or the product was an MAA described in *C. crispus*. We finally added to the database the four orphan MAAs listed in Table S2, as well as two unknown molecules (MAA1 and MAA2) for which we had an experimental support for peaks corresponding to unassigned m/z ratios, as well as mycosporine-glycine, which was here identified in *C. crispus*.

The sterol knowledge base (Figure 5) contained 15 enzymatic reactions involving 24 molecules, including the eight unproducible sterols and the orphan molecule 22-dehydrocholesterol, which was not linked to any reaction. We encoded the «early side-chain reductase» (early SSR) pathway based on the model previously published for tomato (Sonawane et al., 2016), which was added in MetaCyc upon our request (MetaCyc: PWY18C3-1). It also included portions of the canonical plant sterol biosynthesis pathway (MetaCyc: PWY-2541; Benveniste, 2004), as well as portions of the animal sterol synthesis pathway (MetaCyc: PWY66-4, Mitsche et al., 2015).

**Inferring Molecular Transformations from a Database of Enzymatic Reactions**

Based on these examples, we hypothesize that these orphan molecules challenge the ability of the GSMN to produce them because of lacks in secondary metabolism synthesis pathways that enable the description of all possible molecular transformations between compounds catalyzed by a small family of enzymes. First, we define a molecular transformation to be a chemical reaction transforming a metabolite into another metabolite by operating on a specific part of the metabolite structure, called a substructure. We define a substructure to be a set of one of several atoms associated with one or several bonds in a given molecule. For instance, substructure (“simple bond 22-23”) denotes a simple bond between atoms 22 and 23, such as in sitosterol on Figure 3A.

Importantly, an enzymatic reaction involving a single reactant and a single product can be defined as a molecular transformation by assuming that the enzyme enables the transformation of a metabolite site into another. Therefore, for each reaction involving a single reactant and a single product, we call pair of transformed substructures the substructure of the reactant (for instance, as shown in Figure 3A, simple bond between atoms 22 and 23 in sitosterol) and the substructure of the reaction product (for instance, as shown in Figure 3A, double bond 22-23 in stigmasterol).

Pairs of transformed substructures can be computed by removing all the atoms and bonds that are common to both reactant and product molecules, all atoms and bonds being previously ordered. For the sterol and MAA cases, we identified pairs of transformed substructures with a reasoning-based approach (see Methods) and then annotated them to describe a catalog of molecular transformations associated with known enzymatic reactions. For instance, the transformation replacing a simple bond between atoms 22
and 23 (as exhibited in sitosterol) into a double bond 22-23 (as exhibited in stigmasterol) is a c22 desaturation, described by the term transformation (c22_desaturation, simple bond 22-23, double-bond 22-23).

The complete list of molecular transformations for the sterol and MAA reactions is given in Table S5. The sterol enzymatic reaction database (15 reactions) yields 12 molecular transformations, whereas the MAA enzymatic reaction database (13 reactions) yields 9 molecular transformations. Reasoning on molecular transformations instead of enzymatic reactions, we reduce the complexity of the reaction set by abstracting a library of molecular transformations that can be applied to other known molecules belonging to the same chemical family.

**Inferring Molecular Compounds and Enzymatic Reactions from a Database of Molecular Transformations**

We used these databases of molecular transformations to infer putative new molecules and reactions.

To that goal, we assumed that for any molecular transformation from a substructure $S_1$ to a substructure $S_2$, and any pair of metabolites $A$ and $B$, a putative enzymatic reaction can occur from the reactant $A$ to the

---

**Figure 3. The Two Reasoning Methods Implemented in Pathmodel**

(A) Inference of a reaction between two known molecules.

(B) Inference of reaction and metabolite structure corresponding to an unassigned m/z peak. Input data encoded in the knowledge base are in black; inferred reactions and metabolite structures are in green. In both cases, reaction substructure identification followed by inference of molecular transformations are common intermediate steps.
Figure 4. A Model for MAA Biosynthesis Pathway in *C. crispus*

PWY-7751: shinorine biosynthesis pathway. The figure legend details the various data sources integrated to infer the pathways. Stars indicate reactions, and squares indicate molecules. In light blue boxes: reactions and metabolites inferred through Pathmodel.
product B as soon as (1) the molecule A contains the substructure S1, (2) the molecule B contains the substructure S2, and (3) the molecules A and B have identical structures (e.g., same numbered atoms and bonds) for all atoms and bonds to the exception of S1 and S2 substructures. An example is shown in Figure 3A: cholesterol and 22-dehydrocholesterol share a sterane skeleton. The only difference between these molecules is that cholesterol has a simple bond between atoms 22 and 23, whereas 22-dehydrocholesterol has a double bond between atoms 22 and 23. As the transformation from a simple bond to a double bond between atoms 22 and 23 has been evidenced in the reaction database (reaction from sitosterol to stigmasterol, Figure 3), we assume that a putative enzymatic reaction exists between cholesterol and 22-dehydrocholesterol and that it should be a c22_desaturation.

In addition, we assumed that, for a given molecular transformation, a given metabolite A and an experimental mass-to-charge ratio m/z corresponding to an unassigned peak, a putative metabolite B can be produced as soon as (1) the theoretical mass-to-charge ratio of B equals the experimental m/z and (2) A can be transformed into B according to a putative enzymatic reaction associated with the selected molecular transformation. An example is depicted in Figure 3B. We consider that the molecule MAA2 is a putative derived metabolite of porphyra-334 because its mass-to-charge m/z = 302.3117 corresponds to an observed peak in our measurements and that it can be obtained from porphyra-334 by removing a carboxyl group from carbon 18, a transformation named decarboxylation, which occurs in the enzymatic reaction from shinorine to asterina-330.

With these two assumptions, we introduce the concept of putative synthesis pathway computed from a source molecule A, a list of target molecules, a list of molecular transformations, a list of putative metabolites, a list of corresponding theoretical mass-to-charge m/z, and a list of forbidden molecules, for which we have analytical standards that did not match experimental peaks. From these inputs, a putative synthesis pathway is a family of putative reactions connecting the source to all targets metabolites such that (1) all reactions are consistent with the database of molecular transformations, (2) all reactants and products of the reactions are allowed either metabolites or putative metabolites matching with the allowed theoretical mass-to-charge ratio m/z, (3) no transformation associated with the pathway can produce any of the forbidden molecules, and (4) a minimal number of reactions is used to connect the source to the targets.

The Pathmodel method was developed as a prototype implementation of a semi-automatic analogy reasoning approach (Figure 3). Its aim is to infer putative synthesis pathways, to connect orphan metabolites, not yet indexed in metabolic databases, according to the known enzymatic reactions. The Pathmodel method takes as input a knowledge base including a set of known metabolites, a set of known enzymatic reactions, a set of observed mass-to-charge (m/z) ratios for unknown metabolites, an initial source metabolite, a family of targeted metabolites, and a list of forbidden molecules. Metabolites are described by their numbered atoms and bounds. Pathmodel then computes the list of molecular transformations associated with the database of enzymatic reactions and putative synthesis pathways of the targeted metabolites from the source metabolite. More precisely, for each pair of metabolites not linked by a reaction in the knowledge base, the method checks whether a molecular transformation can occur between them (deductive reasoning) and further derives from known reactions candidate metabolites corresponding to observed unassigned m/z ratios (analogical reasoning). These are the bases for iterating the selection of potential reactants or products and the inference by a reasoning component of new reaction occurrences or new metabolites. It was implemented using a logic programming approach known as Answer Set Programming (Lifschitz, 2008; Gebser et al., 2012, see Methods). An example of this reasoning is given in Figure 3.

Application of Pathmodel to the MAA Synthesis Pathway

We first used Pathmodel to compute the synthesis pathway for MAA1 and palythine from sedoheptulose-7-phosphate according to the MAA enzymatic reaction database described above. We also assumed that Z-palythenic acid was defined as a forbidden molecule according to the absence of a corresponding peak in C. crispus extracts.

Figure 5. A Model for Sterol Biosynthesis in C. crispus

Early SSR: pathway involving an early sterol side-chain reduction (SSR), also present in solanacean plants (PWY18C3-1). In light blue box: late SSR pathway, involving a late sterol SSR, so far only described in C. crispus. Portions identical with the plant sterol biosynthesis pathway (PWY-2541) are also boxed. Ovals indicate molecular transformations.
The putative pathway for MAA synthesis obtained by our method is shown in Figure 4. Thanks to this approach, the knowledge base of MAA biosynthesis could be enriched with two putative molecule structures and three putative enzymatic reactions. In addition, the reactions that we compiled from the literature but are not yet indexed in MetaCyc will be submitted to the next release.

A first output of this approach is a pathway leading to a molecule structure compatible with our measured mass-to-charge ratio m/z = 270.3. It was therefore possible to infer the hypothetical structure shown in Figure 4 for this unassigned compound, named MAA1, in MS data. The predicted transformation leading from asterina-330 to MAA1 is a dehydration (in purple), a molecular transformation already observed between other MAAs such as porphyra-334 and Z-palythenic acid, a compound not identified in C. crispus (Figure 4).

Another output of this approach is to suggest a putative involvement of decarboxylation and dehydration transformations in MAA biosynthesis pathways. As no candidate enzymes were mentioned so far in the literature related to MAA biosynthesis pathways, we performed a semantic search on the GSMN from C. crispus, to identify other enzymes that may perform those molecular transformations on a serine coupled with other chemical building blocks. Serine decarboxylation indeed occurs in phospholipid metabolism and was inferred in the C. crispus GSMN based on orthology with G. sulphuraria. The candidate gene is CHC_T0000892001, annotated as phosphatidylserine decarboxylase. Interestingly, there is some evidence of catalytic promiscuity for this enzyme, enabling it to also decarboxylate a threonine residue (Heikkinen and Somerharju, 2002). Therefore, we hypothesize that this enzyme in C. crispus may also perform serine/threonine decarboxylation on a serine/threonine linked to a mycosporine-glycin.

A final interesting feature of the MAA synthesis pathways is the inferred enzymatic reaction required to decarboxylate shinorine and porphyra-334 and further dehydrate their derivatives. These reactions were added to the pathway to take into account the fact that Z-palythenic acid was absent in C. crispus extracts. Such an absence does not support dehydration occurring before decarboxylation, as proposed in other species (Carreto and Carignan, 2011), and therefore does not allow finding any chains of reaction producing palythinol, a compound previously considered to be present in C. crispus based on UV + LC-MS or LC-tandem MS data (Karsten et al., 1998; Athukorala et al., 2016). We noticed, however, that there is no synthesis-based analytical standard available for verifying this prediction. To identify alternative routes for the production of porphyra-334, we ran Pathmodel by allowing the production of putative metabolites with m/z ratio of 302.3177 (the one of palythinol). Pathmodel then predicted an additional intermediate with m/z ratio of 302.3177, named MAA2 on Figure 4. From a genomic viewpoint, switching palythinol with MAA2 does not necessitate a candidate enzyme to perform hydrogenation and demethylation on an MAA-like substrate (Figure 4) and thus reduces the number of unassigned enzymatic activities to candidate genes. We thus included this alternative route in the MAA synthesis pathway that is consistent with the possible absence of Z-palythenic acid.

**Application of Pathmodel to the Sterol Synthesis Pathway**

The Pathmodel approach was also applied to the sterol biosynthetic pathway in C. crispus (Figure 5). To that goal, we used it as an enzymatic reaction database described below. The source metabolite was cycloartenol. The targeted metabolites were 22-dehydrocholesterol, brassicasterol, and stigmasterol. In addition, forbidden metabolites were ergosterol, fucosterol, and zymosterol, which are compounds with no detection result using targeted profiling with analytical standards.

We decided to use cycloartenol even if we did not find it by gas chromatography (GC)-MS for the two following reasons. First, a cycloartenol synthase from the red alga Laurencia dendroidea was cloned and expressed in yeast cells, where it is able to transform squalene into cycloartenol, even if the authors did not report cycloartenol identification in the whole alga by GC-MS (Calegario et al., 2016). Second, unambiguous cycloartenol derivatives are known in another florideophyte red alga, Tricleocarpa fragilis (Horgen et al., 2000). Therefore, we considered more parsimonious to hypothesize that cycloartenol is present and below the limit of experimental detection rather than considering that this step is performed via an unknown intermediate.

We propose two alternative synthesis pathways from cycloartenol to cholesterol, depending on when the side-chain reductase (SSR) enzyme is acting (Figure 5).
If C. crispus uses the « early SSR » pathway (Sonawane et al., 2016), the metabolic intermediates would be identical to tomato, but there would be an important difference concerning the enzymes. Indeed, the genes encoding SSR are duplicated in Solanaceae (tomato and potato) but not in the C. crispus genome or in any red algal genome and in other plants, analyzed so far (Figure S3). The unduplicated SSR from non-solanaceous plants is known to be catalytically promiscuous, and indeed Pathmodel suggested that SSR could act on all possible intermediates (Figure 5).

Another alternative synthesis pathway inferred by Pathmodel consists in producing methylated sterols through C24-methylation on desmosterol (Figure 5). This is in agreement with the identification of a methylated sterol, 24-methylenecholesterol, in C. crispus (Tasende, 2000) and builds on with other reports about methyltransferase catalytic promiscuity across land plants and green algae (Neelakandan et al., 2009; Haubrich et al., 2015). This option highly reduces the number of non-identified methylated intermediates. Indeed, in land plants, a first methylation, involving methyltransferases, occurs directly on cycloartenol, whereas the second one occurs later on 24-methylenecholesterol, to produce methylated sterols like campesterol or brassicasterol (Benveniste, 2004) with intermediates such as cycloeucalenol or fucosterol, both compounds for which we did not find any evidence of presence. By specifying in our model not to enable the production of fucosterol, we naturally omitted this possibility. This model seems also more relevant from a quantitative viewpoint with respect to the formation of cholesterol as the main sterol, because this late methylation step would enable the production of methylated sterols using the late SSR pathway.

A Complete Set of Candidate Enzymes for the C. crispus Sterol Synthesis Pathway

To identify the enzymes associated with the sterol synthesis pathways, we carried out a comparative genomic analysis. Results are summed up in Table S6. In line with previous analyses on sterol biosynthesis gene families in eukaryotes (Desmond and Gribaldo, 2009) or more specifically in green plants (Sonawane et al., 2016), the candidate sterol synthesis enzyme set shows a mixture of conservation and divergence. Seven enzymes are encoded by genes that are conserved as 1:1 orthologs, whereas four of them either underwent lineage-specific duplications (squalene epoxidase and C-4 demethylase) or were lost or may have been replaced by distant paralogs (C24 and C24’ methylases and C22 desaturases). In one case, we found no homolog of known plant or animal enzymes performing delta-7/delta-8 isomerization in the C. crispus genome, but we found a 1:1 ortholog of ERG2, the gene that secondarily took up this function in yeast (Desmond and Gribaldo, 2009). Another similar case in the sterol synthesis pathway occurs in diatoms, where the epoxisqualene cyclase, otherwise conserved in eukaryotes, was secondarily lost and replaced by a protein belonging to the fatty acid hydroxylase superfamily (Pollier et al., 2019). We consider the ERG2 ortholog to be the best candidate for delta-7/delta-8 isomerization in C. crispus, but it is also possible that this reaction is performed by an enzyme encoded by a taxonomically restricted orphan gene (Khalturin et al., 2009).

A Complete View of Algal Primary Metabolism Associated with MAA and Sterol Synthesis

The final GSMN associated with C. crispus is constituted of the initial GSMN, enriched with models curated in detail for the sterol and the MAA synthesis pathways. It was therefore reconstructed from both the genome data and the metabolomic data, using the extraction of molecular transformations implemented in Pathmodel (Figure 6). The final GSMN contains 2,207 metabolites, including the eight initial target sterols that are now producible based on Pathmodel inferences as well as seven former orphan metabolites (six MAAs and 22-dehydrocholesterol). In this way, we increased the proportion of producible targets from 69.4% (59/85) up to 80% (74/92). The final GSMN is available as a wiki website at the following address: https://gem-aureme.genouest.org/ccrgem/index.php/Main_Page. This website is open for further community curation and will serve as the central point to later integrate additional biochemical knowledge about C. crispus. More widely, the curation effort on those pathways will also benefit the entire GSMN community. As already done for the early SSR pathway (now MetaCyc: PWY18C1-3), we will systematically suggest the inclusion of biochemical data used in Pathmodel in the next versions of MetaCyc.

DISCUSSION

In this study we present an extensive analysis of the sterol and MAA biosynthetic pathways in the red alga C. crispus, integrating automatic metabolic network reconstruction, manual curation, metabolite profiling, and semi-automatic analogy reasoning approach based on molecular similarity and dissimilarity to generate hypotheses on metabolic pathways associated with secondary metabolism and metabolites,
which are not described in GSMN reference databases such as Metacyc. Three main conclusions can be derived from the presented research. (1) Our findings underline the usefulness of semi-automatic analogy reasoning approaches to link orphan metabolites to existing pathways through the prediction of molecular transformations to infer reactions based on molecular similarity and dissimilarity. (2) These models support our hypothesis of drift as an evolutionary mechanism shaping the metabolism of living organisms. (3) We propose models of sterol and MAA biosynthesis in *C. crispus* and partially validate these models by metabolite profiling.

**Semi-automatic Analogy Reasoning Approaches to Integrate Orphan Metabolites**

Our study demonstrates that data on metabolite occurrence can be explicitly incorporated into the quality criteria for evaluating a GSMN. To deal with metabolic drift during metabolic network reconstruction, the incorporation of metabolite data is essential, because it puts further constraints on partially known pathways. Putting more emphasis on metabolites, especially the missing ones, creates methodological challenges regarding *ab initio* inferences of pathways when enzymes are not yet known, and we have shown that it is now possible to build tools to specifically address those challenges. The next issue is about the scalability of our approach. The Pathmodel version we present here is a working prototype that can already be applied to other metabolic pathways in *C. crispus* or in other organisms where genomic and metabolomic data are available.

**Application of the Pathmodel Approach to Other Studies**

The results presented in this study are a first step toward the further development of the underlying bioinformatic tools and their application to additional model biosynthesis pathways. Indeed, the Pathmodel tool was developed to support reasoning based on the metabolic pathway drift hypothesis to automatically infer reactions and metabolites. A key feature of the successful application of this strategy was the

---

**Figure 6. Reconstruction Scheme for the Genome-Scale Metabolic Network of *C. crispus***

Green boxes indicate starting data and resulting knowledge. Blue boxes indicate the tools that were used to analyze and integrate genomic and metabolomic data. The part overshadowed in gray indicates tools that are already integrated in the AuReMe workflow (Aité et al., 2018).
precision and the quality of the biochemical and biological knowledge encoded in the reaction and metabolite databases used as entries to Pathmodel. In particular, all metabolites have to be described with a homogeneous ordering of atoms to predict molecular transformations. Generalizing this approach to any other application will similarly require interactions between chemists, biologists, and computer scientists. Further improvements should be made to minimize the burden in manually entering molecular structures. It is not yet possible to fully automate the atom numbering during metabolic reaction due to the intrinsic complexity of metabolic pathways. For example, the split of molecules during biochemical processes can generate inconsistent numbering (Figure S5A) that can only be handled by resetting atom numbering from one step to another. This is already done in MetaCyc, but in a way that does not make possible to automatically number in the same manner atoms from different reactions that share a molecular transformation (Figure S5B). Other methods, like the CLCA approach (Kumar and Maranas, 2014) already implemented ways to compare reactions sharing molecular transformations, and thus would provide lists of candidate reactions for molecular transformations, but the atom numbering during the comparisons does not allow simultaneous atom mapping between reactant and product (Figure S5C) that we need in Pathmodel to abstract the molecular transformation. Therefore, following IUPAC atom encoding as much as possible seems to be the best way to combine atom mapping and abstraction of molecular transformations (Figure S5D).

The second key feature of Pathmodel is to be focused on a selected pathway rather than on a complete genome-scale metabolic network. The selection of the relevant pathway to be considered, for instance, from preliminary evidences extracted from metabolomic analysis, is therefore a key pre-processing step to combine and filter the predictions of Pathmodel with genomic and metabolomic data. Practically, Pathmodel can already be used on any type of incomplete biochemical pathway on a molecule class for which there is knowledge about some biochemical reactions and orphan molecules not yet connected to each other by any reaction. The main limitation will be the user time necessary to properly encode the starting knowledge base.

**Metabolic Drift as a Driver for Pathway Evolution**

Whatever the actual topology of the sterol and MAA pathways in *C. crispus*, each discussed hypothesis has implications regarding metabolic pathway drift. All possible sterol pathways provide further strong candidate case studies for drift by non-homologous enzyme replacement, and the pathways inferred by Pathmodel provide candidate case studies for drift by enzyme inversion. The unresolved point with the sterol pathways is that, among eukaryotes, there is no consensus yet about the ancestral order of enzymatic reactions. Experimental data are too disparate across the tree of life to enable firm conclusions on this. In that respect, the MAA pathway is interesting, because if our hypothesis about decarboxylation of porphyra-334 before dehydration is true, this would mean that an enzymatic inversion took place in other lineages where porphyra-334 is first dehydrated to Z-palythenic acid and then decarboxylated to palythene. Here the limit is that, to date, enzymes are unknown for both reactions, so the system is not yet genomically tractable. Identifying close enzymatic inversions would be important to provide a mechanism for gradual divergence of pathways. Indeed, experimental analyses on *E. coli* have shown that drastic pathway rewiring by enzyme knockout or gene overexpression can led to toxic intermediates (Kim et al., 2010).

**Limitations of the Study**

Our arguments for the metabolic pathway drift hypothesis rely on comparisons between pathways from *C. crispus* and from distantly related species belonging to a few other phyla, such as land plants or animals. Additional support should become available from detailed comparisons based on additional eukaryotic lineages and, when possible, from multiple species in the same phylum. Finally, it remains to be determined to what extent chance and selection contributed in generating metabolic pathway diversity.

**METHODS**

All methods can be found in the accompanying Transparent Methods supplemental file.

**SUPPLEMENTAL INFORMATION**

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2020.100849.
ACKNOWLEDGMENTS

We thank Cécile Hervé for help in collecting field samples from C. crispus, Gaëlle Correc for help in sample preparation, Karine Cahier for help during GC-MS analyses on the MetaboMer-Corsaire platform, Jeanne Got and Marie Chevallier for help in using preliminary versions of AuReMe, and Clémence Frioux for help in analyzing FBA artifacts. G.V.M. is also grateful to Ralf J. Sommer for giving him the possibility to start developing Pathmodel during a previous postdoctoral stay at the Max-Planck Institute for Developmental Biology in Tübingen. We thank Alessio Peracchi for commenting on the putative role of PLP-enzymes in the MAA biosynthesis pathway and the two anonymous reviewers for their improvement suggestions. This work benefited from the support of the French Government via the National Research Agency investment expenditure program IDEALG (ANR-10-BTBR-04) and from Région Bretagne via the grant « SAD 2016 - METALG (9673) ».

AUTHOR CONTRIBUTIONS

G.V.M., L.D., S.M.D., P.S., E.C., J.N., C.B., C. Leblanc, A.S., and J.C. conceived the project. J.G. conducted the sterol profiling with help from L.D., C. Leroux, C. Leblanc, J.C., and G.V.M. P.S. conducted the MAA profiling with help from C.M. G.V.M. performed the genome-scale metabolic network reconstruction with help from M.A., C.T., S.M.D., J.C., and A.S. A.B., G.V.M., and J.N. wrote the Pathmodel software. A.B., J.G., L.D., and G.V.M. curated the biosynthetic pathway models. G.V.M. wrote the manuscript with input and edits from all authors.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: December 21, 2018
Revised: November 11, 2019
Accepted: January 13, 2020
Published: February 21, 2020

SUPPORTING CITATIONS

The following reference appears in the Supplemental Information: Anisimova and Gascuel, 2006; Gouy et al., 2010; Guindon and Gascuel, 2003; Kamehisa et al., 2017; King et al., 2016; Le and Gascuel, 2010; Loira et al., 2015; Moretti et al., 2016; Prigent et al., 2017; Sievers and Higgins, 2014.

REFERENCES

Aite, M., Chevallier, M., Frioux, C., Trottier, C., Got, J., Cortés, M.P., Mendoza, S.N., Carrier, G., Dameron, O., Guillaume, N., et al. (2018). Traceability, reproducibility and wiki-exploration for “à-la-carte” reconstructions of genome-scale metabolic models. PLoS Comput. Biol. 14, e1006146.

Aliche, A., Devys, M., and Barbier, M. (1968). Remarques sur les stérols des algues rouges. Phytochemistry 7, 329–330.

Anisimova, M., and Gascuel, O. (2006). Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. Syst. Biol. 55, 539–552.

Athukorala, Y., Trang, S., Kwok, C., and Yuan, Y.W. (2016). Antiproliferative and antioxidant activities and mycosporine-like amino acid profiles of wild-harvested and cultivated edible Canadian marine red macroalgae. Molecules 21, E119.

Banskota, A.H., Stefanova, R., Sperker, S., Lali, S., Craigie, J.S., and Hafting, J.T. (2014). Lipids isolated from the cultivated red alga Chondrus crispus inhibit nitric oxide production. J. Appl. Physcol. 26, 1565–1571.

Belghit, I., Rasinger, J.D., Heesch, S., Biancarosa, I., Liland, N., Torstensen, B., Waagba, R., Lock, E.-J., and Bruckner, C.G. (2017). In-depth metabolic profiling of marine macroalgae confirms strong biochemical differences between brown, red and green algae. Algal Res. 26, 240–249.

Boveniste, P. (2004). Biosynthesis and accumulation of sterols. Annu. Rev. Plant Biol. 55, 429–457.

Brawley, S.H., Blouin, N.A., Ficko-Blean, E., Wheeler, G.L., Lohr, M., Goodson, H.V., Jenkins, J.W., Blaby-Haas, G.E., Hellwell, K.E., Chan, C.X., et al. (2017). Insights into the red alga and eukaryotic evolution from the genome of Porphyra umbilicalis (Bangioiphyceae, Rhodophyta). Proc. Natl. Acad. Sci. U S A 114, E6361–E6370.

Calegario, G., Poiller, J., Arendt, P., de Oliveira, L.S., Thompson, C., Soares, A.R., Pereira, R.C., Goossens, A., and Thompson, F.L. (2016). Cloning and functional characterization of cycloartenol synthase from the red seaweed Laurencia dendroidea. PLoS One 11, e0165954.

Carreto, J.L., and Carignan, M.O. (2011). Mycosporine-like amino acids: relevant secondary metabolites. Chemical and ecological aspects. Mar. Drugs 9, 387–446.

Caspi, R., Millington, R., Ferrer, L., Foerster, H., Fulcher, C.A., Keseler, I.M., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L.A., et al. (2016). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. 44, D471–D480.

Collin, J., Porcel, B., Carré, W., Ball, S.G., Chapparo, C., Tonon, T., Barbezyron, T., Michel, G., Noel, B., Valentín, K., et al. (2013). Genome structure and metabolic features in the red seaweed Chondrus crispus shed light on evolution of the Archaeplastida. Proc. Natl. Acad. Sci. U S A 110, 5247–5252.

Collin, J., Cornish, M.L., Craie, J., Ficko-Blean, E., Herve, C., Krueger-Handfield, S.A., Leblanc, C., Michel, G., Potin, P., Tonon, T., and Boyen, C. (2014). Chondrus crispus – a present and historical model organism for red seaweeds. Adv. Bot. Res. 71, 53–90.
Cormier, A., Avia, K., Sterle, L., Derrien, T., Wucher, V., Andres, G., Monsoor, M., Godfrey, O., Lipinska, A., Perrienne, M.M., et al. (2017). Re-annotation, improved large-scale assembly and establishment of a catalogue of noncoding loci for the genome of the model brown alga Ectocarpus. New Phytol. 214, 219–232.

de Oliveira Dal’Molin, C.G., Quek, L.E., Palfreyman, R.W., Brunblum, S.M., and Nielsen, L.K. (2010). AraGEM, a genome-scale reconstruction of the primary metabolic network in Arabidopsis. Plant Physiol. 152, 579–589.

Desmond, E., and Gribaldo, S. (2009). Phylogenomics of sterol synthesis: insights into the origin, evolution, and diversity of a key eukaryotic feature. Genome Biol. Evol. 1, 364–381.

Dittami, S.M., Corre, E., Brillet-Guéquelin, L., Lipinska, A.P., Pontonzeau, N., Aite, M., Avia, K., Caron, C., Cho, C.H., Colén, J., et al. (2020). The genome of Ectocarpus subulatus highlights unique mechanisms for stress tolerance in brown algae. Mar. Genomics. https://doi.org/10.1016/j.margen.2020.100740.

Gaquerel, E., Hervé, C., Labrière, C., Boyen, C., Potin, P., and Sulkunen, J.P. (2007). Evidence for oxygen synthesis and induction of a new polyunsaturated fatty acid hydroxylase activity in Chondrus crispus in response to methyljasmonate. Biochim. Biophys. Acta 1771, 565–575.

Gebser, M., Kaminis, R., Kaufmann, B., and Schaub, T. (2012). Answer set solving in practice. Synth. Lect. Artif. Intell. Mach. Learn. 6, 1–238.

Goldberg, A., Hubby, C., Cobb, D., Millard, P., Ferrara, N., Galdi, G., Premuzic, E.T., and Gaffney, J.S. (1982). Sterol distribution in red algae Trichecercarps fragilis. J. Nat. Prod. 63, 210–216.

Imam, S., Schaabule, S., Valenzuela, J., López García de Lomana, A., Carter, W., Price, N.D., and Baliga, N.S. (2015). Refined genome-scale reconstruction of Chlamydomonas metabolism provides a platform for systems-level analyses. Plant J. 84, 1239–1256.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 45, D353–D361.

Karp, P. D., Paley, and Romero. (2002). The Pathway Tools software. Bioinformatics 18, S225–S232.

Karsten, U., Franklin, L.A., Lönig, K., and Wiencke, C. (1998). Natural ultraviolet radiation and photosynthetically active radiation induce formation of mycosporine-like amino acids in the macroalga Chondrus crispus (Rhodophyta). Planta 205, 257–262.

Khalturn, K., Hemmrich, G., Fraune, S., Augustin, R., and Bosch, T.C.G. (2009). More than just orphans: are taxonomically-restricted genes important in evolution? Trends Genet. 25, 404–413.

Kim, J., Kershner, J.P., Novokov, Y., Shoemaker, R.K., and Copley, S.D. (2010). Three serendipitous pathways in E. coli can bypass a block in pyridoxal-5-phosphate synthesis. Mol. Syst. Biol. 4, 436.

King, R.D., Whelan, K.E., Jones, F.M., Reiser, P.G.K., Bryant, C.H., Muggleton, S.H., Kell, D.B., and Oliver, S.G. (2004). Functional genomic hypothesis generation and experimentation by a robot scientist. Nature 427, 247–252.

King, Z.A., Lu, J., Dräger, A., Miller, P., Fedorowicz, S., Lerman, J.A., Ebrahim, A., Palsson, B.O., and Lewis, N.E. (2016). BiGG Models: a platform for integrating, standardizing and sharing genome-scale models. Nucleic Acids Res. 45, DS51–DS52.

Koch, M., Duiguou, T., Carbonnell, P., and Faulon, J.L. (2017). Molecular structures enumeration and virtual screening in the chemical space with RetroPath2.0. J. Cheminform. 9, 64.

Kronin, E.V., Mushogean, A.R., and Bork, P. (1996). Non-orthologous gene displacement. Trends Genet. 12, 334–336.

Kräbs, G., Watanabe, M., and Wiencke, C. (2004). A monochromatic action spectrum for the photoinduction of the UV-absorbing mycosporine-like amino acid shinorine in the red alga Chondrus crispus. Photochem. Photobiol. 79, 515–519.

Kremer, B.P., and Kist, G.O. (1982). Biosynthesis of photosynthates and taxonomy of algae. Z. Naturforsch. 37c, 761–771.

Kumar, A., and Maranas, C.D. (2014). CLCA: maximum common molecular substructure queries within the MetRxn Database. J. Chem. Inf. Model. 54, 3417–3438.

Laglegerie, F., Lajii, S., Bedoux, G., Taupin, L., Siger-Pouvreau, V., and Connan, S. (2019). Photo-protective compounds in red macroalgae from Brittany: considerable diversity in mycosporine-like amino acids (MAAs). Mar. Environ. Res. 147, 37–48.

Laycock, M.V., and Craigie, J.S. (1977). The occurrence and seasonal variation of gigantin and L-citrullinyl-L-arginine in Chondrus crispus Stackh. Can. J. Biochem. 55; 27–30.

Le, S.Q., and Gascuel, O. (2010). Accounting for solvent accessibility and secondary structure in protein phylogenetics is clearly beneficial. Syst. Biol. 59, 277–287.

Lifschitz, V. (2008). What is answer set programming? In AAAI/08: Proceedings of the 23rd national conference on Artificial intelligence, A. Cohn, ed. (Chicago: AAAI Press), pp. 1594–1597.

Loira, N., Zhukova, A., and Sherman, D.J. (2015). Pantograph: a template-based method for genome-scale metabolic model reconstruction. J. Bioinform. Comput. Biol. 13, 1550006.

Markov, G.V., Meyer, J.M., Panda, O., Artyukhin, A.B., Claesen, M., Witte, H., Schroeder, F.C., and Sommer, R.J. (2016). Functional conservation and divergence of daf-22 paralogs in Pristionchus pacificus during development. Mol. Biol. Evol. 33, 2506–2514.

Matsuhiro, B., and Urzua, C. (1992). Heterogeneity of carrageenans from Chondrus crispus. Phytochemistry 31, 531–534.

Melo, T., Alves, E., Azevedo, V., Martins, A.S., Neves, B., Domingues, P., Calado, R., Abreu, M.H., and Domínguez, M.R. (2015). Lipidomics as a new approach for the bioprospecting of marine macroalgae – Unraveling the polar lipid and fatty acid composition of Chondrus crispus. Algal Res. 8, 181–191.

Mitsche, M.A., McDonald, J.G., Hobbs, H.H., and Cohen, J.C. (2015). Flux analysis of cholesterol biosynthesis in vivo reveals multiple tissue and cell-type specific pathways. Elife 4, e07999.

Moretti, S., Martin, O., Van Du Tran, T., Bridge, A., Morpat, A., and Pagni, M. (2016). MetaNetX/MNXref – reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. Nucleic Acids Res. 44, D523–D526.

Moss, G.P. (1989). IPAC-IUB Joint Commission on Biochemical Nomenclature. Nomenclature of steroids (Recommendations 1989). Eur. J. Biochem. 167, 1783–1822.

Neelakandan, A.K., Song, Z., Wang, J., Richards, M.H., Wu, X., Valliyodan, B., Nguyen, H.T., and Nes, W.D. (2009). Cloning, functional expression and phylogenetic analysis of plant sterol 24C-methyltransferases involved in sitosterol biosynthesis. Phytochemistry 70, 1982–1998.

Noda-Garcia, L., Liebermeister, W., and Tawfik, D.S. (2018). Metabolite-enzyme coevolution from single enzymes to metabolic pathways and networks. Annu. Rev. Biochem. 87, 187–216.
Pettit, T., Jones, A., and Harwood, J. (1989). Lipid metabolism in the red marine algae Chondrus crispus and Polysiphonia lanosa as modified by temperature. Phytochemistry 28, 2053–2089.

Pina, A., Costa, A., Lage-Yustu, M., and López-Hernández, A. (2014). An evaluation of edible red seaweed (Chondrus crispus) components and their modification during the cooking process. LWT - Food Sci. Technol. 56, 175–180.

Pollier, J., Vancaester, E., Kuzhiumparambil, U., Vickers, C.E., Vande Poele, K., Goossens, A., and Fabris, M. (2019). A widespread alternative squalene epoxidase participates in eukaryote steroid biosynthesis. Nat. Microbiol. 4, 226–233.

Pirgent, S., Collet, G., Dittami, S.M., Delage, L., Ethis de Corny, F., Dameron, O., Eveillard, D., Thiele, S., Cambefort, J., Boyen, C., et al. (2014). The genome-scale metabolic network of Ectocarpus siliculosus (EctoGEM): a resource to study brown algal physiology and beyond. Plant J. 80, 367–381.

Pirgent, S., Frioux, C., Dittami, S.M., Thiele, S., Larhlimi, A., Collet, G., Gutknecht, F., Got, J., Eveillard, D., Bourdon, J., et al. (2017). Meneco, a topology-based gap-filling tool applicable to degraded genome-wide metabolic networks. PloS Comput. Biol. 13, e1005276.

Rhee, K.Y., de Carvalho, L.P.S., Bryk, R., Ehr, S., Marrero, J., Park, S.W., Schnappinger, D., Venugopal, A., and Nathan, C. (2011). Central carbon metabolism in Mycobacterium tuberculosis: an unexpected frontier. Trends Microbiol. 19, 307–314.

Robertson, R.C., Guihéneuf, F., Bahar, B., Schmid, M., Stengel, D.B., Fitzgerald, G.F., Ross, R.P., and Stanton, C. (2015). The anti-inflammatory effect of algae-derived lipid extracts on lipopolysaccharide (LPS)-stimulated human THP-1 macrophages. Mar. Drugs 13, 5402–5424.

Saito, A., and Idler, D.R. (1966). Sterols in Irish moss (Chondrus crispus). Can. J. Biochem. 44, 1195–1199.

Santos, S.A., Vilela, C., Freire, C.S., Abreu, M.H., Rocha, S.M., and Silvestre, A.J. (2015). Chlorophyll a and Rhodophylla macroalgae: a source of health promoting phytochemicals. Food Chem. 183, 122–128.

Schönknecht, G., Chen, W.H., Ternes, C.M., Bariay, G.G., Shrestha, R.P., Stanke, M., Brautigam, A., Baker, B.J., Banfield, J.F., Garavito, R.M., et al. (2013). Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote. Science 339, 1207–1210.

Shick, J.M., and Dunlap, W.C. (2002). Mycospore-like amino acids and related gadusols: biosynthesis, accumulation, and UV-protective functions in aquatic organisms. Annu. Rev. Physiol. 64, 223–262.

Sievers, F., and Higgins, D.G. (2014). Clustal Omega, accurate alignment of very large numbers of sequences. Methods Mol. Biol. 1079, 105–116.

Sonawane, P.D., Pollier, J., Panda, S., Szymanski, J., Massalha, H., Yona, M., Unger, T., Malitsky, S., Arendt, P., Pauwels, L., et al. (2016). Plant cholesterol biosynthetic pathway overlaps with phytosterol metabolism. Nat. Plants 3, 16205.

Sumner, L.W., Amberg, A., Barrett, D., Besle, M.H., Beger, R., Dykinson, C.A., Fan, T.W., Fieth, O., Goodacre, R., Griffin, J.L., et al. (2007). Proposed minimum reporting standards for chemical analysis chemical analysis working group (CAWG) metabolomics standards initiative (MSI). Metabolomics 3, 211–221.

Tasende, M. (2000). Fatty acid and sterol composition of gametophytes and sporophytes of Chondrus crispus (Gigartinaeaceae, Rhodophyta). Sci. Mar. 64, 421–426.

Townsley, B.T., and Sinha, N.R. (2012). A new development: evolving concepts in leaf ontogeny. Annu. Rev. Plant Biol. 63, 535–562.

True, J.R., and Haag, E.S. (2001). Developmental system drift and flexibility in evolutionary trajectories. Evol. Dev. 3, 109–119.

van Ginneken, V.J., Helser, J.P., de Visser, W., van Keulen, H., and Brandenburg, W.A. (2011). Polysaturated fatty acids in various macroalgal species from north Atlantic and tropical seas. Lipids Health Dis. 10, 104.

Young, E.G., and Smith, D.G. (1958). Amino acids, peptides, and proteins of Irish moss, Chondrus crispus. J. Biol. Chem. 233, 406-410.

Pettit, C., Rey, C., Lambert, A., Pelletier, M., Pantalacci, S., and Sémon, M. (2016). Comparing transcriptomes to probe into the evolution of developmental program reveals an extensive developmental system drift. In JOBIM2016 Conference Proceedings of the 17èmes Journées Ouvertes en Biologie, Informatique et Mathématiques, G. Perrière and F. Picard, eds. (Published online by the Société Française de Bioinformatique (SFBI), pp. 118–120.
Supplemental Information

Inferring Biochemical Reactions and Metabolite Structures to Understand Metabolic Pathway Drift

Arnaud Belcour, Jean Girard, Méziane Aite, Ludovic Delage, Camille Trottier, Charlotte Marteau, Cédric Leroux, Simon M. Dittami, Pierre Sauleau, Erwan Corre, Jacques Nicolas, Catherine Boyen, Catherine Leblanc, Jonas Collén, Anne Siegel, and Gabriel V. Markov
TRANSPARENT METHODS

Genome-Scale Metabolic Network reconstruction

Genome-Scale Metabolic Network (GSMN) reconstruction was performed using the AuReMe pipeline (Aite et al., 2018). A set of 85 targets coming from the literature was used as an input and is provided in Table S1. Orphan metabolites that are experimentally supported but do not have a MetaCyc ID are listed in Table S2. The process encompassed the following steps:

1) an annotation-based draft network was generated using the PathoLogic program from the Pathway Tools suite, using the gbk file from the *Chondrus crispus* genome annotation (Collén et al., 2013) and the metabolic reaction database MetaCyc20.5 (Caspi et al., 2016).

2) an orthology-based network was generated using the protein sequences and metabolic network of *Arabidopsis thaliana* (AraGEM, De Oliveira d'al Molin et al., 2010), using the Pantograph software (Loira et al., 2015) to combine the output of ortholog searches with the Inparanoid and OrthoMCL softwares.

3) an orthology-based network was generated using the protein sequences from the well-annotated red microalga *Galdieria sulphuraria* (Schönknecht et al., 2013) and its metabolic network reconstructed using Pathway Tools. This *G. sulphuraria* annotation-based network was then used as a template to generate a *C. crispus* network using Pantograph. We decided to build this template GSMN after realizing that the genbank file was especially annotation-rich.

4) an orthology-based network was also generated using the protein sequences from the version 2 of the annotated genome of *Ectocarpus siliculosus* (Cormier et al., 2017), as well as version 2 of its metabolic network (Aite et al., 2018).

5) the four preliminary networks were merged together in the AuReMe environment, and an additional gap-filling step was performed using Meneco (Prigent et al., 2017), constraining the network to produce the 85 metabolites from the literature that were indexed in the Metacyc database (Table S1).

Sampling of algae

For sterol analyses, samples from *C. crispus* were collected from a population on the shore at Roscoff, France, in front of the Station Biologique (48°43′38″ N; 3°59′04″ W). Algal cultures were maintained in 10 L flasks in a culture room at 14°C using filtered seawater and aerated with 0.22 μm-filtered compressed air to avoid CO₂ depletion. Photosynthetically active radiation (PAR) was provided by Philips daylight fluorescence tubes at a photon flux density of 40 μmol.m⁻².s⁻¹ for 10 h.d⁻¹. The algal samples were freeze dried, ground to powder using a cryogrinder and stored at
For MAAs analysis, more than 50 g (wet weight) of *C. crispus* were collected along the Brittany coasts (France) at Ploemeur (47°42′07″ N; 3°24′31″ W) in July 2013, Roscoff (48°43′38″ N; 3°59′04″ W) in April and August 2013, and Tregunc (47°50′25″N; 3°54′08″ W) in September 2013.

**Standards and reagents**

Cholesterol, stigmasterol, β-sitosterol, 7-dehydrocholesterol, lathosterol (5α-cholest-7-en-3β-ol), squalene, campesterol, brassicasterol, desmosterol, lanosterol, fucosterol, cycloartenol, 5α-cholestane (internal standard) were acquired from Sigma-Aldrich (Saint-Quentin-Fallavier, France), cycloartanol and cycloeucalenol from Chemfaces (Wuhan, China) and zymosterol from Avanti Polar Lipids (Alabaster, USA). The C7-C40 Saturated Alkanes Standards were acquired from Supelco (Bellefonte, USA). Reagents used for extraction, saponification, and derivation steps were *n*-hexane, ethyl acetate, acetonitrile, methanol (Carlo ERBA Reagents, Val de Reuil, France), (trimethylsilyl)diazomethane, toluene (Sigma-Aldrich, Saint-Quentin-Fallavier, France) and N,O-bis(trimethylsilyl)trifluoroacetamide with trimethylcholorosilane (BSTFA:TMCS (99:1)) (Supelco, Bellefonte, USA).

**Standard preparation**

Stock solutions of cholesterol, stigmasterol, β-sitosterol, 7-dehydrocholesterol, lathosterol (5α-cholest-7-en-3β-ol), squalene, campesterol, brassicasterol, desmosterol, lanosterol, fucosterol, cycloartenol and 5α-cholestane were prepared in hexane with a concentration of 5 mg.mL⁻¹. Working solutions were made at a concentration of 1 mg.mL⁻¹, in hexane, by diluting stock solutions. The C7-C40 Saturated Alkanes Standard stock had a concentration of 1 mg.mL⁻¹ and a working solution was made at a concentration of 0.1 mg.mL⁻¹. All solutions were stored at -20°C.

**Sample preparation**

For sterol analyses, dried algal samples (60 mg) were extracted with 2mL ethyl acetate by continuous agitation for 1 hour at 4°C. After 10 min of centrifugation at 4000 rpm, the solvent was removed, the extracts were saponified in 3 mL of methanolic potassium hydroxide solution (1M) by 1 hour incubation at 90°C. The saponification reaction was stopped by plunging samples into an ice bath for 30 min minimum. The unsaponifiable fraction was extracted with 2 mL of hexane and 1.2 mL of water and centrifuged at 2000 rpm for 5 min. The upper phase was collected, dried under N₂, and resuspended with 120 μL of (trimethylsilyl)diazomethane, 50 μL of methanol:toluene (2:1 (v/v)) and 5 μL of 5α-cholestane (1 mg.mL⁻¹) as internal standard. The mixture was vortexed for 30
seconds, and heated at 37°C for 30 min. After a second evaporation under N₂, 50 µL of acetonitrile and 50 µL of BSTFA:TMCS (99:1) were added to the dry residue, vortexed for 30 seconds and heated at 60°C for 30 min. After final evaporation under N₂, the extract was resuspended in 100 µL of hexane, transferred into a sample vial and stored at -80°C until the GC-MS analysis.

For MAAs, one gram of dried algae was extracted twice for two hours under continuous shaking with 10 mL of acetone. After 5 min of centrifugation at 3000 rpm, acetone was discarded and samples were re-extracted twice with 10 mL water/acetone (30/70, v/v) for 24 hours under continuous shaking at 120 rpm. Water/acetone supernatants were pooled, added to one gram of silica and evaporated to dryness by rotary evaporation. Extracts were then purified by silica gel chromatography column with dichloromethane/methanol mixtures and MAAs were eluted with 200 mL of dichloromethane/methanol (15/85, v/v). After rotary evaporation, samples were re-suspended in water/methanol (50/50, v/v) and filtrated using 0.45 µm syringes filter. Solution were adjusted to a final concentration of 1 mg.mL⁻¹ and stored at 3°C until LC-MS analysis.

**Sterol analysis by gas chromatography-mass spectrometry**

The sterols were analyzed on a 7890 Agilent Technologies gas chromatography coupled with a 5975C Agilent Technologies mass spectrometer (GC-MS). A HP-5MS capillary GC column (30 m x 0.25 mm x 0.25 µm) from J&W Scientific (CA, USA) was used for separation and UHP helium was used as carrier gas at flow rate to 1 mL.min⁻¹. The temperature of the injector was 280°C and the detector temperature was 315°C. After injection, the oven temperature was kept at 60°C for 1 min. The temperature was increased from 60°C to 100°C at a rate of 25°C.min⁻¹, then to 250°C at a rate of 15°C.min⁻¹, then to 315°C at a rate of 3°C.min⁻¹ and then held at 315°C for 2 min, resulting in a total run time of 37 min. Electronic impact mass spectra were measured at 70eV and an ionization temperature of 250°C. The mass spectra scanned from m/z 50 to m/z 500. Peaks were identified based on the comparisons with the retention times and the mass spectra (Table S3).

**MAA analysis by liquid chromatography-mass spectrometry**

High Resolution Mass Spectrometry was carried out on a microTOF-Q II (Bruker Daltonics, Germany) coupled to an Ultimate 3000 LC System (Dionex, Germany). Experiments were performed on a Gemini C6-Phenyl column (250 mm x 4.6 mm x 5 µm) (Phenomenex, Germany). The gradient was as follows: methanol/water (20:80, v/v) with 0.2% acid acetic for two minutes to 100% methanol with 0.2% acid acetic in 23 minutes. The UV detector was set to 330 nm, flow rate was kept constant at 0.4 mL.min⁻¹ and column temperature set at 30°C. MS spectra were recorded in positive ESI mode with a drying gas temperature of 220°C, a nitrogen flow of 12 L.min⁻¹, a nebulizer pressure set to 60 psi, and a collision energy of 20 eV. MAAs were identified by HR-MS
on the basis of the detection of the pseudo-molecular ion \([M+H]^+\) with a \(m/z\) value varying less than \(\pm 0.02\) Da compared to the theoretical \(m/z\) value. In the absence of commercially available standards, relative quantification of MAAs in each sample was estimated by calculating the ratio between the area under the curve of the Extracted Ion Chromatogram (EIC) corresponding to the selected MAAs and the sum of the areas under the curve of the EIC of all MAAs detected in the algal extract. The same procedure was applied to UV detection (Table S4).

**Flux-balance analysis**

A biomass reaction was established based on the previous *E. siliculosus* data, defining a list of 33 compounds to be produced in order to consider the network functional (Prigent et al., 2014). One compound, L-alpha-alanine, was not producible, thus blocking biomass production. This was due to the absence of the alanine dehydrogenase reaction. The corresponding enzyme (CHC_T00008930001) was present in the *C. crispus* network but annotated as an NAD(P) transhydrogenase. We completed the annotation through the manual curation form to enable it to dehydrogenate alanine and to restore producibility of the biomass ([https://gem-aureme.genouest.org/ccrgem/index.php/Manual-ala_dehy](https://gem-aureme.genouest.org/ccrgem/index.php/Manual-ala_dehy)).

**Global metabolic networks comparisons**

In order to compare the global features of the GSM from *C. crispus* with other ones, it is necessary to use the same reference database. This is the case for *E. siliculosus* and *E. subulatus* for which the reconstructions are based on MetaCyc (Caspi et al., 2016) while *A. thaliana* and *Chlamydomonas reinhardtii* are respectively from KEGG (Kanehisa et al., 2017) and BiGG (King et al., 2016). To get access to MetaCyc pathway information for *A. thaliana* and *C. reinhardtii*, their networks were mapped using the sbml_mapping function implemented in the AuReMe workflow (Aite et al., 2018). This function provides a dictionary of corresponding reactions from a database to another one using the MetaNetX cross-reference database (Moretti et al., 2016). This dictionary was then used in AuReMe to create a new genome-scale metabolic network based on the new reference database for *A. thaliana* and *C. reinhardtii*. Those new networks, who are comparable in size with the published ones (+/- 10 reactions and enzymes in our counts) enabled to estimate the number of pathways as defined in MetaCyc for both species.

**Ab-initio inference of metabolic reactions: implementation of a Semi-Automatic Analogy Reasoning Approach**

The Pathmodel method was developed to infer new reactions based on molecular similarity and dissimilarity. This knowledge-based approach is founded on two modes of reasoning (deductive and
analogical) and was implemented using a logic programming approach known as Answer Set Programming (ASP) (Lifschitz et al., 2008; Gebser et al., 2012). It is a declarative approach oriented toward combinatorial (optimization) problem-solving and knowledge processing. ASP combines both a high-level modeling language with high performance solving engines so that the focus is on the problem specification rather than the algorithmic part. ASP expresses a problem as a set of logical rules (clauses). Problem solutions appear as particular logical models (so-called stable models or answer sets) of this set. An ASP program consists of rules $h :- b_1, \ldots, b_m \text{ not } b_{m+1}, \ldots, \text{ not } b_n$, where each $b_i$ and $h$ are literals and not stands for default negation. In fact, each proposition is a predicate, encoded by a function whose arguments can be constant atoms or variables over a finite domain. The rule states that the head $h$ is proven to be true ($h$ is in an answer set) if the body of the rule is satisfied, i.e. $b_1, \ldots, b_m$ are true and it cannot be proved that $b_{m+1}, \ldots, b_n$ are true.

The main predicates used in Pathmodel to represent molecules and reactions forming a knowledge base are *bond*, *atom* and *reaction*. The theoretical m/z ratio of a molecule is determined by logical rules, which were encoded in the program MZComputation.lp.

As depicted in Figure 3, several logical rules are then applied to all possible reactions and potential reactants. These are the bases for the selection of potential reactants or products and the inference by a reasoning component of reaction occurrences or metabolites, using either deductive or analogical reasoning in the PathModel.lp program. Resulting products that do not belong to the knowledge base but that correspond to an observed m/z ratio are considered as inferred metabolites and reactions. The finally encoded reactions result from iterative interactions between analogical model construction, automated inference, and manual validation of inferred reactions with respect to experimental results.

By comparing reactants and products, the program ReactionSiteExtraction.lp characterizes two structures of the reaction site containing atoms and bonds involved in the reaction. The predicates $\text{diffAtomBeforeReaction}$, $\text{diffBondBeforeReaction}$, $\text{diffAtomAfterReaction}$ and $\text{diffBondAfterReaction}$ compare atoms and bonds between the reactant and the product and extract the two structures. Then these two structures are compared to the structure of all other molecules in the knowledge base (predicates $\text{siteBeforeReaction}$ and $\text{siteAfterReaction}$). These predicates characterize sub-structures of the molecules that can be part of a reaction.

By deductive reasoning, the reference molecule pair of each reaction is compared to the structures of a potential reactant-product pair sharing a common chemical structure. The presence of the reaction site in the two putative molecules is checked using the predicates $\text{siteBeforeReaction}$ and $\text{siteAfterReaction}$. Furthermore, if the product and the reactant have the same overall structure, except for the reaction site, the program will infer that the reaction actually occurs between the
reactant and the product.

By analogical reasoning, all possible reactions are applied to potential reactants, and resulting products are filtered using their structures and m/z ratios. The predicate `newMetaboliteName` creates all the possible products from a known molecule using all the reactions in the knowledge base. These possible metabolites are filtered using their m/z ratios, which must correspond to an unassigned m/z ratio (predicate `possibleMetabolite`) and checked if they share the same structure as a known molecule (predicate `alreadyKnownMolecule`). If they do not fit with an already known molecule, they will be added as new molecules and a new reaction variant.

Given a source molecule and a target molecule, the program will take several inference steps iteratively applying either analogical or deductive reasoning modes. To connect the source and the target molecules along a pathway, Pathmodel infers missing reactions and metabolites using a minimal number of reactions. To further constraint the number of possible pathways, a predicate `absentmolecules` was added to avoid pathways with compounds for which targeted profiling with analytical standards gives strong evidence for real absence (here ergosterol, fucosterol and zymosterol). The source code is available in the following Github repository: https://github.com/pathmodel/pathmodel

It includes a specific tutorial to replicate the analysis reported in the article: https://github.com/pathmodel/pathmodel#tutorial-on-article-data-chondrus-crispus-sterol-and-mycosporine-like-amino-acids-pathways

**De novo gene prediction and manual curation of gene sequence models**

Missing genes from the sterol synthesis pathway (squalene monooxygenase and sterol C-4 methyl oxidase) were found by targeted tblastn using orthologs from other organisms as a query. The new gene predictions are provided in supplementary dataset 1 and will be included in the next version of *C. crispus* genome browser (http://mmo.sb-roscoff.fr/jbrowse/?data=data%2Fpublic%2Fchondrus). The split protein sequence of sterol delta-7 reductase was also restored as a single protein prediction, merging the two adjacent partial predictions.

**Phylogenetic analyses**

Collected sequences were aligned using Clustal Omega (Sievers and Higgins, 2014) and alignments were checked manually and edited with Seaview (Gouy et al., 2010). Phylogenetic trees were built using PHYML (Guindon and Gascuel, 2003) using the LG model (Le and Gascuel, 2010) with a gamma law. The reliability of nodes was assessed by likelihood-ratio test (Anisimova and Gascuel, 2006).
SUPPLEMENTAL REFERENCES

Anisimova, M., and Gascuel, O. (2006). Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. Syst. Biol. 55, 539 – 552.

Gouy, M., Guindon, S., and Gascuel, O. (2010). SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol. Biol. Evol. 27, 221 – 224.

Guindon, S., and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52, 696 – 704.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 45, D353 – D361.

King, Z.A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J.A., Ebrahim, A., Palsson, B.O., and Lewis, N.E. (2016). BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. Nucleic Acids Res. 44, D515 – D522.

Le, S.Q., and Gascuel, O. (2010). Accounting for solvent accessibility and secondary structure in protein phylogenetics is clearly beneficial. Syst. Biol. 59, 277 – 287.

Loira, N., Zhukova, A., and Sherman, D.J. (2015) Pantograph: A template-based method for genome-scale metabolic model reconstruction. J. Bioinform. Comput. Biol. 13, 1550006.

Moretti, S., Martin, O., Van Du Tran, T., Bridge, A., Morgat, A., and Pagni, M. (2016). MetaNetX/MNXref – reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. Nucleic Acids Res. 44, D523 – D526.

Prigent, S., Frioux, C., Dittami, S.M., Thiele, S., Larhlimi, A., Collet, G., Gutknecht, F., Got, J., Eveillard, D., Bourdon, J., et al. (2017). Meneco, a topology-based gap-filling tool applicable to degraded genome-wide metabolic networks. PLoS Comput. Biol. 13, e1005276.

Sievers, F., and Higgins, D.G. (2014). Clustal Omega, accurate alignment of very large numbers of sequences. Methods Mol. Biol. 1079, 105 – 116.

Schönknecht, G., Chen, W.H., Ternes, C.M., Barbier, G.G., Shrestha, R.P., Stanke, M., Bräutigam, A., Baker, B.J., Banfield, J.F., Garavito, R.M., et al. (2013). Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote. Science 339, 1207 – 1210.
| Table S1 related to Fig 6. Database metabolites. |
|---------------------------------------------|
| **Usual name** | **Category** | **MetaCyC ID** | **References** |
| dodecanoic acid | 12:0 fatty acid | DODECANOATE* | Santos et al., 2015 ; Robertson et al., 2015 |
| Myristic acid | 14:0 fatty acid | CPD-7836 | Pettitt et al., 1989; Tasende, 2000; Van Ginneken et al., 2011; Robertson et al., 2015 ; Belghit et al., 2017 |
| Pentadecanoic acid | 15:0 fatty acid | CPD-8462 | Santos et al., 2015 ; Belghit et al., 2017 |
| Palmitic acid | 16:0 fatty acid | PALMITATE | Pettitt et al., 1989; Tasende, 2000; Van Ginneken et al., 2011; Robertson et al., 2015 ; Belghit et al., 2017 |
| Heptadecanoic acid | 17:0 fatty acid | CPD-7830 | Santos et al., 2015 |
| Stearic acid | 18:0 fatty acid | STEARIC ACID | Tasende et al., 2000 ; Robertson et al., 2015 |
| Eicosanoic acid | 20:0 fatty acid | ARACHIDIC ACID* | Santos et al., 2015 |
| Docosanoic acid | 22:0 fatty acid | DOCOSANOATE* | Santos et al., 2015 |
| Tricosanoic acid | 23:0 fatty acid | CPD-7834* | Santos et al., 2015 |
| Tetracosanoic acid | 24:0 fatty acid | TETRACOSANOATE | Santos et al., 2015 |
| Palmitoleic acid | 16:1(n-7) fatty acid | CPD-9245 | Pettitt et al., 1989; Tasende, 2000; Robertson et al., 2015 ; Belghit et al., 2017 |
| Oleic acid | 18:1(n-9) fatty acid | OLEATE-CPD | Tasende et al., 2000; Van Ginneken et al., 2011; Robertson et al., 2015 ; Belghit et al., 2017 |
| Linoleic acid | 18:2(n-6) fatty acid | LINOLEIC ACID | Tasende et al., 2000 ; Robertson et al., 2015 ; Belghit et al., 2017 |
| Alpha Linolenic acid | 18:3(n-3) fatty acid | LINOLENIC ACID* | Tasende et al., 2000 |
| γ-linolenic acid | 18:3(n-6) fatty acid | CPD-8117* | Robertson et al., 2015 ; Belghit et al., 2017 |
| Octadecatetraenoic acid | 18:4(n-3) fatty acid | CPD-12653* | Tasende et al., 2000 ; Robertson et al., 2015 ; Belghit et al., 2017 |
| Arachidonic acid | 20:4(n-6) fatty acid | ARACHIDONIC ACID | Tasende et al., 2000 ; Banskota et al., 2014 ; Robertson et al., 2015 ; Belghit et al., 2017 |
| Eicosapentaenoic acid | 20:5(n-3) fatty acid | 5Z3Z11Z14Z17Z-EICOSAPENTAENOATE* | Tasende et al., 2000 ; Banskota et al., 2014 ; Robertson et al., 2015 ; Belghit et al., 2017 |
| Octanedic acid | fatty acid | CPD0-1264* | Santos et al., 2015 |
| Nonanedioic acid | fatty acid | CPD0-1265* | Santos et al., 2015 |
| Cycloartenol | sterol | CYCLOARTENOL* | Salto and Idler, 1966; Alcaide et al., 1968 |
| Cholesterol | sterol | CHOLESTEROL* | Salto and Idler, 1966; Tasende et al., 2000 ; Santos et al., 2015 |
| 7-Dehydrocholesterol | sterol | 7-DEHYDROCHOLESTEROL* | Tasende et al., 2000 |
| Brassinosterol | sterol | BRASSICASTEROL* | Salto and Idler, 1966 ; Tasende et al., 2000 |
| Campesterol | sterol | CAMPESTEROL* | Tasende et al., 2000 ; Santos et al., 2015 |
| 24-Methylenecholesterol | sterol | 24-METHYLENECHOLESTEROL* | Tasende et al., 2000 |
| Sitosterol | sterol | SITOSTEROL* | Salto and Idler, 1966; Tasende et al., 2000 ; Santos et al., 2015 |
| Stigmasterol | sterol | STIGMASTEROL* | Tasende et al., 2000 |
| 15-keto-prostaglandin E2 | oxylipin | HYDROXY-9,11-DIOXOPROSTA-13-ENOATE* | Gaquerel et al., 2007 |
| lutein | carotenoid | LUTEIN* | Banskota et al., 2014 |
| Chlorophyll a | tetrapyrrrole | CHLOROPHYLL-A | Melo et al., 2015 ; Robertson et al., 2015 |
| all-trans-beta-carotene | carotenoid | CPD1F-129 | Robertson et al., 2015 |
| 9-cis-beta-carotene | carotenoid | CPD-14646 | Robertson et al., 2015 ; Belghit et al., 2017 |
| zeaxanthin | carotenoid | CPD1F-130 | Robertson et al., 2015 |
| 2,6,8-trimethyl-1,3-cyclohexadiene-1-carboxaldehyde (safranal) | carotenoid | CPD-8669* | Pina et al., 2014 |
| Alanine | aminoacid | L-ALPHA-ALANINE | Young et al., 1958, Belghit et al., 2017 |
| Arginine | aminoacid | ARG | Young et al., 1958, Belghit et al., 2017 |
| Aspartic acid | aminoacid | L-ASPARTATE | Young et al., 1958, Belghit et al., 2017 |
| Citrulline | aminoacid | L-CITRULLINE | Young et al., 1958 ; Belghit et al., 2017 |
| Metabolite               | Type                | Reference                                      |
|-------------------------|---------------------|------------------------------------------------|
| Cystine                 | Aminoacid           | Young et al., 1958                            |
| Glutamic acid           | Aminoacid           | Young et al., 1958 ; Belghit et al., 2017     |
| Glycine                 | Aminoacid           | Young et al., 1958 ; Belghit et al., 2017     |
| Histidine               | Aminoacid           | Young et al., 1958 ; Belghit et al., 2017     |
| Isoleucine              | Aminoacid           | Young et al., 1958 ; Belghit et al., 2017     |
| Leucine                 | Aminoacid           | Young et al., 1958 ; Belghit et al., 2017     |
| Lysine                  | Aminoacid           | Young et al., 1958 ; Belghit et al., 2017     |
| Methionine              | Aminoacid           | Young et al., 1958 ; Belghit et al., 2017     |
| Ornithine               | Aminoacid           | Young et al., 1958 ; Belghit et al., 2017     |
| Phenylalanine           | Aminoacid           | Young et al., 1958                            |
| Proline                 | Aminoacid           | Young et al., 1958 ; Belghit et al., 2017     |
| Serine                  | Aminoacid           | Young et al., 1958 ; Belghit et al., 2017     |
| Threonine               | Aminoacid           | Young et al., 1958 ; Belghit et al., 2017     |
| Tyrosine                | Aminoacid           | Young et al., 1958 ; Belghit et al., 2017     |
| Valine                  | Aminoacid           | Young et al., 1958 ; Belghit et al., 2017     |
| Shinorine               | Mycosporine-like aminoacid | Kråks et al., 2004 |
| UDP-α-D-galactose       | Nucleotide sugar    | Collén et al., 2014                           |
| D-galactosyl-1,2-diacylglycerol | Galactolipid | Banskota et al., 2014 |
| ⅰ-carrageose           | Carrageenan         | Matsuhiro et al., 1992                        |
| ⅰ-carrageose           | Carrageenan         | Matsuhiro et al., 1992                        |
| Glycerol                | Polyl               | Santos et al., 2015                           |
| Heptadecane             | Alcan              | Santos et al., 2015                           |
| 6,10,14-Trimethyl-2-pentadecanone | Methylketone | Santos et al., 2015 |
| Hexadecan-1-ol          | Long chain aliphatic alcohol | Santos et al., 2015 |
| 9-Octadecan-1-ol        | Long chain aliphatic alcohol | Santos et al., 2015 |
| Docosan-1-ol            | Long chain aliphatic alcohol | Santos et al., 2015 |
| Octacosan-1-ol          | Long chain aliphatic alcohol | Santos et al., 2015 |
| Acetaldehyde            | Aldehyde           | Pina et al., 2014                             |
| 2-methylpropanal        | Aldehyde           | Pina et al., 2014                             |
| Butanal                 | Aldehyde           | Pina et al., 2014                             |
| 3-methylbutanal         | Aldehyde           | Pina et al., 2014                             |
| Pentanal                | Aldehyde           | Pina et al., 2014                             |
| Hexanal                 | Aldehyde           | Pina et al., 2014                             |
| Benzaldehyde            | Aldehyde           | Pina et al., 2014                             |
| Ethanol                 | Short chain aliphatic alcohol | Pina et al., 2014 |
| 1-butanol               | Short chain aliphatic alcohol | Pina et al., 2014 |
| 1-pentanol              | Short chain aliphatic alcohol | Pina et al., 2014 |
| 2-butanone              | Short chain ketone  | Pina et al., 2014                             |
| 3,5-octadien-2-one      | Short chain ketone  | Pina et al., 2014                             |
Table S1 related to Fig 6. Database metabolites.

| Compound              | Class       | CPD   | Reference         |
|-----------------------|-------------|-------|-------------------|
| dichloromethane       | halocarbon  | CPD-681 | Pina et al., 2014 |
| chloroform            | halocarbon  | CPD-843* | Pina et al., 2014 |
| glycerate             | carboxylic acid | GLYCERATE | Belghi et al., 2017 |
| 2-methylpropanoic acid| carboxylic acid | ACET | Pina et al., 2014 |
| 2-methylbutanoic acid | carboxylic acid | ISOBUTYRATE | Pina et al., 2014 |
| hexane                | acyclic     | CPD-9288* | Pina et al., 2014 |
| 2,2,4-trimethylpentane| acyclic     | CPD-19039* | Pina et al., 2014 |

* non predicted in initial GSMN reconstruction
| Usual name                        | Category                  | References                     |
|----------------------------------|---------------------------|--------------------------------|
| Heneicosanoic acid               | 21:0 fatty acid           | Santos et al., 2015            |
| N/A                              | 16:1 fatty acid           | Robertson et al., 2015         |
| N/A                              | 18:1(n-7) fatty acid      | Robertson et al., 2015         |
| 10-nonadecenoate                 | 19:1(n-9) fatty acid      | Belght et al., 2017            |
| Eicosadienoic acid               | 20:2(n-6) fatty acid      | Robertson et al., 2015         |
| Eicosatrienoic acid              | 20:3(n-6) fatty acid      | Robertson et al., 2015         |
| Docosadienoate                   | 22:2(n-6) fatty acid      | Belght et al., 2017            |
| Octadeca-9-enolic acid           | fatty acid                | Santos et al., 2015            |
| 22-Dehydrocholesterol*           | sterol                    | Tasende et al., 2000           |
| 11-hydroxy-octadecadienoic acid  | oxylipin                  | Gaquerel et al., 2007          |
| 13-hydroxy-9Z,11E-octadecadienoic acid | oxylipin                  | Gaquerel et al., 2007; Belght et al., 2017 |
| 13S-hydroxy-9Z,11E,15Z-octadecatrienoic acid | oxylipin                  | Belght et al., 2017            |
| 13-oxo-9Z,11E-octadecadienoic acid | oxylipin                  | Belght et al., 2017            |
| 13-hydroxyeicosatetraenoic acid  | oxylipin                  | Gaquerel et al., 2007          |
| 13-hydroxyeicosapentaenoic acid  | oxylipin                  | Gaquerel et al., 2007          |
| 15-hydroxydocosahexaenoic acid   | oxylipin                  | Gaquerel et al., 2007          |
| 11-hydroxyoctadecadienoic acid   | oxylipin                  | Gaquerel et al., 2007          |
| Hydroxypheophytin a               | tetrapyrrole              | Melo et al., 2015              |
| Pheophytin d                      | tetrapyrrole              | Melo et al., 2015              |
| Hydroxypheophytin d               | tetrapyrrole              | Melo et al., 2015              |
| Monogalactosyldiacylglycerol 2 (MGDG2) | galactolipid              | Pettitt et al., 1989           |
| Digalactosyldiacylglycerols (DGDG) | galactolipid              | Pettitt et al., 1989           |
| Sulfoquinovosyldiacylglycerol 1 (SQDG1) | galactolipid              | Pettitt et al., 1989           |
| Sulfoquinovosyldiacylglycerol 2 (DQDG) | galactolipid              | Pettitt et al., 1989           |
| (2S)-1,2-bis-O-eicosapentaenoyl-3-O-β-D-galactopyranosylglycerol | galactolipid              | Banskota et al., 2014          |
| (2S)-1-O-eicosapentaenoyl-2-O-arachidonoyl-3-O-β-D-galactopyranosylglycerol | galactolipid              | Banskota et al., 2014          |
| (2S)-1-O-(6Z,9Z,12Z,15Z-octadecatrienoyl)-2-O-palmitoyl-3-O-β-D-galactopyranosylglycerol | galactolipid              | Banskota et al., 2014          |
| (2S)-1-O-eicosapentaenoyl-2-O-palmitoyl-3-O-β-D-galactopyranosylglycerol | galactolipid              | Banskota et al., 2014          |
| (2S)-1,2-bis-O-arachidonoyl-3-O-β-D-galactopyranosylglycerol | galactolipid              | Banskota et al., 2014          |
| (2S)-1-O-arachidonoyl-2-O-palmitoyl-3-O-β-D-galactopyranosylglycerol | galactolipid              | Banskota et al., 2014          |
| (2S)-1-O-arachidonoyl-2-O-palmitoyl-3-O-β-D-galactopyranosylglycerol | galactolipid              | Banskota et al., 2014          |
| diphasphatidylglycerol, phosphatidic acid | phospholipid              | Pettitt et al., 1989           |
| l-citrullinyl-l-arginine           | aminoacid                 | Laycock et al., 1977           |
| Gigartinine                      | aminoacid                 | Laycock et al., 1977           |
| Amide N                          | aminoacid                 | Young et al., 1958             |
| Asterina-330*                    | Mycosporine-like aminoacid | Athukorala et al., 2016; Guheneuf et al., 2018 |
Table S2 related to Fig 6. Orphan metabolites.

| Compound                  | Class                  | Reference                                      |
|---------------------------|------------------------|------------------------------------------------|
| MAA1*                     | Mycosporine-like aminoacid | This study                                    |
| MAA2*                     | Mycosporine-like aminoacid | This study                                    |
| Palythine*                | Mycosporine-like aminoacid | Karsten et al., 1998; Athukorala et al., 2016; Guihéneuf et al., 2018 |
| Palythene                 | Mycosporine-like aminoacid | Karsten et al., 1998                         |
| Porphyra-334*             | Mycosporine-like aminoacid | Athukorala et al., 2016                      |
| Isofloridoside            | heteroside             | Kremer et al., 1982                           |
| 1-Monohexadecanoic        | Long chain aliphatic alcohol | Santos et al., 2015                          |
| Tetradecan-1-ol           | Long chain aliphatic alcohol | Santos et al., 2015                          |
| Octadecan-1-ol            | Long chain aliphatic alcohol | Santos et al., 2015                          |
| 1-penten-3-ol             | short chain aliphatic alcohol | Santos et al., 2015                          |
| 2(Z)-penten-1ol           | short chain aliphatic alcohol | Santos et al., 2015                          |
| 3-methylbutanoic acid     | carboxylic acid         | Santos et al., 2015                           |
| 2-methylbutanal           | aldehyde                | Santos et al., 2015                           |
| 1-octen-3-ol              | short chain aliphatic alcohol | Santos et al., 2015                          |
| 2,6-dimethylpyrazine      | short chain ketone      | Santos et al., 2015                           |
| 2-propanone               | short chain ketone      | Santos et al., 2015                           |
| acetic acid, anhydride    | carboxylic acid         | Santos et al., 2015                           |
| 2,2,3-trimethylpentane    | alcane                  | Santos et al., 2015                           |
| tetradecane               | alcane                  | Santos et al., 2015                           |

* incorporated in GSMN after Pathmodel
| Analysed compounds      | Molecular weight (g.mol⁻¹) | RT (min) | m/z [M+H]+ (TMS) |
|-------------------------|----------------------------|----------|------------------|
| brassicasterol          | 398.64                     | 25.5     | 470              |
| campesterol             | 400.68                     | 26.6     | 472              |
| 5α-cholestane           | 372.67                     | 20.3     | 372              |
| cholesterol             | 386.65                     | 24.7     | 458              |
| cycloartanol            | 428.75                     | 30.5     | 500              |
| cycloartenol            | 426.72                     | 29.0     | 498              |
| cycloeucalenol          | 426.73                     | 30.4     | 498              |
| 7-dehydrocholesterol    | 384.63                     | 25.6     | 456              |
| desmosterol             | 384.64                     | 25.5     | 456              |
| ergosterol              | 396.65                     | 26.2     | 468              |
| fucosterol              | 412.69                     | 28.2     | 484              |
| lanosterol              | 426.39                     | 27.8     | 498              |
| lathosterol             | 386.65                     | 25.8     | 458              |
| β-sitosterol            | 414.39                     | 28.0     | 486              |
| squalene                | 410.72                     | 19.8     | 482              |
| stigmasterol            | 412.66                     | 27.0     | 484              |
| zymosterol              | 384.64                     | 25.9     | 456              |

Table S3, related to Figure 5. Retention times and m/z ratio for analytical standards of sterols on a 7890-5975C Agilent GC-MS.
Figure S1, related to Figure 5. Identification of squalene in *C. crispus*. a) Total Ion Chromatogram (TIC) from *C. crispus* extract. b) MS spectrum of squalene in *C. crispus* extract. c) MS spectrum of the squalene analytical standard. Main fragmentation peaks identical in both spectra are highlighted in red circles.
Figure S2, related to Figure 5. Control for technical detectability of cycloartenol in spiked *Chondrus crispus* extract.
a) TIC from *Chondrus crispus* extract incubated with cycloartenol. b) MS spectrum of cycloartenol standard incorporated in *C. crispus* extract. c) MS spectrum of cycloartenol standard alone. Main fragmentation peaks identical in both spectra are highlighted in red circles.
Table S4, related to Figures 2 and 4. MAAs composition in *Chondrus crispus* determined by LC-UV-HRMS. Extracted Ion Chromatogramm (EIC) of selected MAAs were obtained in positive mode; UV Absorbance was recorded at 330 nm (LOD = Limit Of Detection).
| Name of source reaction   | Metacyc ID                | Molecular transformation         | Biosynthesis pathway |
|--------------------------|---------------------------|----------------------------------|----------------------|
| rxn_4282                 | RXN-4282                  | delta24_25_reduction              | Sterols              |
| c24_c29_demethylation    | RXN-20433, RXN20434, RXN20435 | c24_c29_demethylation            | Sterols              |
| rxn_20436                | RXN-20436                 | cyclopropylsterol isomerisation   | Sterols              |
| rxn_20438                | RXN-20438                 | c14_demethylation                | Sterols              |
| rxn_20439                | RXN-20439                 | c14_reduction                     | Sterols              |
| rxn_4286                 | RXN-4286                  | c8_isomerisation                 | Sterols              |
| c24_c28_demethylation    | RXN-20440, RXN20441, RXN20442 | c24_c28_demethylation            | Sterols              |
| rxn_1_14_21_6            | 1.14.21.6-RXN             | c5_desaturation                   | Sterols              |
| rxn66_323                | RXN66-323                 | delta24_25_reduction              | Sterols              |
| rxn66_28                 | RXN66-28                  | delta24_25_reduction              | Sterols              |
| rxn_4021                 | RXN-4021                  | c24_methylation                  | Sterols              |
| rxn_2_1_1_143            | 2.1.1.143-RXN             | c24'_methyltransfer              | Sterols              |
| rxn_20131                | RXN-20131                 | delta24_24'_reduction            | Sterols              |
| rxn_4243                 | RXN-4243                  | c22_desaturation                  | Sterols              |
| c22_desaturation         | RXN-4242 or RXN-8352      | c22_desaturation                  | Sterols              |
| mysa                     | RXN-17372                 | cyclisation                       | MAA                  |
| rxn_17366                | RXN-17366                 | methyl_transfer                   | MAA                  |
| rxn_17370                | RXN-17370                 | non-enzymatic tautomerization     | MAA                  |
| rxn_17896                | RXN-17896                 | methyl_transfer                   | MAA                  |
| aminoacid_C_1_transfer   | RXN-17368                 | aminoacid_c1_transfer            | MAA                  |
| aminoacid_C_3_transfer_serine | RXN-17367              | aminoacid_c3_transfer            | MAA                  |
| aminoacid_C_3_transfer_threonine | -                      | aminoacid_c3_transfer            | MAA                  |
| hydrogenation             | -                         | hydrogenation                     | MAA                  |
| demethylation             | -                         | demethylation                     | MAA                  |
| dehydration              | -                         | dehydration                       | MAA                  |
| decarboxylation_1         | -                         | decarboxylation                   | MAA                  |
| decarboxylation_2         | -                         | decarboxylation                   | MAA                  |
| hydrolysis                | -                         | hydrolysis                        | MAA                  |

Table S5, related to Figure 3. List of molecular transformations inferred in Pathmodel and associated source reactions.
| Steps                        | Yeast     | Human    | Arabidopsis | C. crispus          |
|------------------------------|-----------|----------|-------------|---------------------|
| squalene monoxygenation      | ERG1      | SQLE     | SQE1-7      | scaffolds 90*, 20*, 57* |
| oxydosqualene cyclisation    | ERG7      | LSS      | CAS         |                     |
|                               |           |          |             | CHC_T00008265001    |
| C-14 demethylation           | ERG11     | CYP51A1  | CYP51G1     |                     |
|                               |           |          |             | CYP51G1 (CHC_T00009303001) |
| C-14 reduction               | ERG24     | TM7SF2   | FK          |                     |
|                               |           |          |             | CHC_T00003466001    |
| C-14 reduction               | ERG25     | SC4MOL   | SMO1, SMO2  |                     |
| C-4 demethylation            |           |          |             | CHC_T00010320001, scaffold212* |
| delta-8, delta-7 isomerisation| ERG2    | EBP      | HYD1        |                     |
|                               |           |          |             | CHC_T00001257001    |
| C-5 desaturation             | ERG3      | SC5DL    | STE1        |                     |
| C24 or C24' methylation      | ERG6      | -        | SMT1, SMT2  |                     |
| delta-7 reduction            |           | DHCR7    | DWF5        |                     |
|                               |           |          |             | CHC_T00006492-3001* |
| delta-24 reduction           | ERG4      | DHCR24   | DWF1/SSR    |                     |
|                               |           |          |             | CHC_T00002789001    |
| C-22 desaturation            | ERG5      | -        | CYP710      |                     |
|                               |           |          |             | CYP805A1-C1 or CYP808A1-H1 |
| cyclopropylsterol isomerisation|         | -        | CPI1        |                     |
|                               |           |          |             | CHC_T00002985001    |

**Table S6, related to Figure 5. Comparative Genomic Analysis of Sterol Synthesis Enzymes.** Dark blue indicates orthologous sequences, light blue indicates paralogous ones, and green indicates yeast enzymes non orthologous to animal or plant sequences but known to perform the same enzymatic reaction. Five corrected sequences and new predictions are indicated with an asterisk (*) and provided in Dataset S1.
**Figure S3, related to Figure 5. Maximum-likelihood tree of eukaryotic side-chain reductases.** In green: protein sequences from green plants (streptophytes). The green dot indicates lineage-specific duplication in solanaceans. In red: protein sequences from red algae. In blue: protein sequences from opisthokonts (vertebrates + choanoflagellates). In brown: sequences from oomycete stramenopiles. Likelihood-ratio test values above 0.90 are indicated. Those above 0.97 are considered significant.
Dataset S1, related to Figure 5. New or edited protein sequences associated to the sterol synthesis pathway in Chondrus crispus.

> scaffold90:7511-6165(-) putative squalene epoxidase
RDGRRLCVERQLYAPSGALCAPPRIVGELLQPGGYDALCRLGLADALLIDDAQVIRGYALFLGPRAERLPYHQPGPDPDPAARPQPEGRAFHNGRFLKRLREIARAHNPVT
TLVEGNVLALLERDGAVGVGRYATRGKAAAAHVTYVSNFHGLVHLVPALPFPNHGHVVLADPCPVLFYPISATEVR
CLVDIPSTYAGDAAEYILHTVYPQVPPPLRAPLATAVRERRSKMMPNRPAPA
HVVPAGVLLGDADFNMRHPLTGGMTVALTDVELLRELLAPVDPDSDAPAVAAKLQLFYER
RKPMSTTINILANALYTLCATDDPALRDMRAACLDYLAKGGRMTHDPIAMMLGGLKPQRH
LLLHFFAVALEYGCGLMPFPFTPRLVRAWSIFRASFNIIKPLANAEGFWPLSWLPLNSL

> scaffold20:461442-460650(-) putative squalene epoxidase
LCRLGLADALLHIDAQVIRGYALFLGPRAERLPYHQPGPDPDPAARPQPEG
RAFHNGRFLKRLREIARAHNPVTLEIGNVLALLERDGAVV
GVRYATRGKAAAAHVTYVSNFHGLVHLVPALPFPNHGHVVLAPCVPVLFPISATEVRCLVDL
YILHTVYPQVPPSLRAPLATTVRERRSKMMPNRPAPAHVVPAGVLLGDADFNRHPLTG
GGMTVALTDVELLRGLLAP
>scaffold57:152407-364140(+) putative squalene epoxidase
RFAGPEHPSCGLKPQRHLLLAAHFFAAVALYGCNKLMPFPFTPARPVRGAWRAISFRAFSFNFK
PLANAEGFWPLSWLPLN
LCRLGLADALLDIDAQVIRGYALFLGFAERLPY
LCRLGLADALLHIDAQVIRGYALFLGFAERLPYHQPAGPDPDAARPQPQEG
RAFHNCRFLKRLREIARAHPNVTLEIGNVLALLERGAVV
GVRYATRGNDAAATAHAGLTIACDGCGSALRKRRAAAHHTVTVSNGHGLVLHPALPFPNH
GHVVLAHCPVLFYPISATEAVRCVLVD
WSTYAGDAEYILHTVVFQVPPSLRAPLATAVRERRSRRKMPMRVMPAPAHVVPVGAVLLGD
AFNMRHPLTGGMTVALDTVELLRGGLAP

>scaffold212:177405-176674(−) putative C-4 sterol methyl oxidase
WDLLCRHRAYPMFVVGCFASQMLAGYFLGCAFPVLLDLARSTPFRKIQPGKYAPRAV
FAAAAAMLRSFATVVLPLLAAGGLFIERVGGRSDAPFSPRVPVLLQVAYFLVEDFLNYW
VHRALHLRPYTRVHSVVHEYDAPFAVAAYAPVEVVLQALPTFAOAPLMLGPHLTYLCV
WQLFRNWEAIDIHSGYDHAWGLASVLPVYAGPEHHDFHHFLHSNGFASVFTWCDWAYGTD
LAYER

>CHC_T00006492-3001 fusion of adjacent protein predictions CHC_T00006492001 and CHC_T00006493001; putative sterol delta-7 reductase
MLGIAAWKGFIRYGLLYDHFGEVLAFLGKALFVTLNYFRGIYFTPNTSDSFGIVWDWMDWGETLHP
EIFGVSLQVLNCRFALMGWSVAVIAFACKREQYQGYSNMLVSVVLQVYIFKFFVWEAGYFNSVSLD
HSVCLFWIYLRLPY
MVGVAICCCNYWTDKQREVFRATNGQVTIWGQPVSIEAQYVTGDKKRRSLASGWGGRHDYNVFE
IALTFCWSVPAGGTGVVIPHYVVMFLTILLTDRAVRDERVCSEKYGGYEEYCLRPYKMIPGVY
Figure S4, related to Figure 3. Architecture of Pathmodel scripts. In green: Data files, either input or result files. In red: ASP scripts. In blue: Python scripts. The black line shows the wrapping of all the scripts inside by pathmodel_wrapper.py.
Figure S5, related to Figure 3. Comparisons of atom mapping using MetaCyc, CLCA and IUPAC. A. Biochemical origin of G3P and DHAP generates numbering inconsistencies that have to be solved by carbon atom renaming. B. Metacyc numbering does not label identically atoms from two reactions involving the same molecular transformation. C. CLCA numbering allows comparisons of reactions but not simultaneous atom mapping. D. IUPAC numbering allows both simultaneous atom mapping and comparisons of reactions.