iScience, Volume *23*

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 CO_2 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

-80°C.

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 (5*α*-cholest-7-en-3*β*-ol), squalene, campesterol, brassicasterol, desmosterol, lanosterol, fucosterol, cycloartenol, 5*α*-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 (5*α*-cholest-7-en-3*β*-ol), squalene, campesterol, brassicasterol, desmosterol, lanosterol, fucosterol, cycloartenol and 5*α*-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 N_2, and resuspended with 120 µL of (trimethylsilyl)diazomethane, 50 µL of methanol:toluene (2:1 (v/v)) and 5 µL of 5*α*-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 N_2, 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 N_2, 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 v*alue varying less than ± 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).

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. reinhardti*. 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, ... , b_m *not* b_{m+1}, ... , *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, … , b_m are true and it cannot be proved that b_{m+1}, … , 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 *diffAtomBeforeReaction, diffBondBeforeReaction*, *diffAtomAfterReaction and 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 *siteBeforeReaction* and *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 *siteBeforeReaction* and *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	5Z8Z11Z14Z17Z-EICOSAPENTAENOATE*	Tasende et al., 2000 ; Banskota et al., 2014 ; Robertson et al., 2015 ; Belghit et al., 2017
Octanedioic acid	fatty acid	CPD0-1264*	Santos et al., 2015
Nonanedioic acid	fatty acid	CPD0-1265*	Santos et al., 2015
Cycloartenol	sterol	CYCLOARTENOL*	Saito and Idler, 1966; Alcaide et al., 1968
Cholesterol	sterol	CHOLESTEROL*	Saito and Idler, 1966; Tasende et al., 2000 ; Santos et al., 2015
7-Dehydrocholesterol	sterol	7-DEHYDROCHOLESTEROL*	Tasende et al., 2000
Brassicasterol	sterol	BRASSICASTEROL*	Saito 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*	Saito and Idler, 1966; Tasende et al., 2000 ; Santos et al., 2015
Stigmasterol	sterol	STIGMASTEROL*	Tasende et al., 2000
15-keto-prostaglandin E2	oxylipin	HYDROXY-915-DIOXOPROSTA-13-ENOATE*	Gaquerel et al., 2007
lutein	carotenoid	LUTEIN*	Banskota et al., 2014
Chlorophyll a	tetrapyrrole	CHLOROPHYLL-A	Melo et al., 2015 ; Robertson et al., 2015
all-trans-beta-carotene	carotenoid	CPD1F-129	Robertson et al., 2015
9-cis-betacarotene	carotenoid	CPD-14646	Robertson et al., 2015 ; Belghit et al., 2017
zeaxanthin	carotenoid	CPD1F-130	Robertson et al., 2015
2,6,6-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

Table S1 related to Fig 6. Database metabolites.

Cystine	aminoacid	CYSTINE	Young et al., 1958
Glutamic acid	aminoacid	GLT	Young et al., 1958 ; Belghit et al., 2017
Glycine	aminoacid	GLY	Young et al., 1958 ; Belghit et al., 2017
Histidine	aminoacid	HIS	Young et al., 1958 ; Belghit et al., 2017
Isoleucine	aminoacid	ILE	Young et al., 1958 ; Belghit et al., 2017
Leucine	aminoacid	LEU	Young et al., 1958 ; Belghit et al., 2017
Lysine	aminoacid	LYS	Young et al., 1958 ; Belghit et al., 2017
Methionine	aminoacid	MET	Young et al., 1958 ; Belghit et al., 2017
Ornithine	aminoacid	L-ORNITHINE	Young et al., 1958 ; Belghit et al., 2017
Phenylalanine	aminoacid	PHE	Young et al., 1958
Proline	aminoacid	PRO	Young et al., 1958 ; Belghit et al., 2017
Serine	aminoacid	SER	Young et al., 1958 ; Belghit et al., 2017
Threonine	aminoacid	THR	Young et al., 1958 ; Belghit et al., 2017
Tyrosine	aminoacid	TYR	Young et al., 1958 ; Belghit et al., 2017
Valine	aminoacid	VAL	Young et al., 1958 ; Belghit et al., 2017
Shinorine	Mycosporine-like aminoacid	CPD-18778	Kräbs et al., 2004
UDP-α-D-galactose	nucleotide sugar	CPD-14553	Collén et al., 2014
D-galactosyl-1,2-diacylglycerol	galactolipid	D-Galactosyl-12-diacyl-glycerols	Banskota et al., 2014
ι-carrageenose	carrageenan	Iota-Carrageenan*	Matsuhiro et al., 1992
ν-carrageenan	carrageenan	Nu-Carrageenan*	Matsuhiro et al., 1992
Glycerol	polyol	GLYCEROL	Santos et al., 2015
Heptadecane	alcane	HEPTADECANE-CPD	Santos et al., 2015
6,10,14-Trimethyl-2-pentadecanone	methylketone	CPD-7875	Santos et al., 2015
Hexadecan-1-ol	Long chain aliphatic alcohol	CPD-348	Santos et al., 2015
9-Octadecen-1-ol	Long chain aliphatic alcohol	CPD-7873	Santos et al., 2015
Docosan-1-ol	Long chain aliphatic alcohol	CPD-7845	Santos et al., 2015
Octacosan-1-ol	Long chain aliphatic alcohol	CPD-7872*	Santos et al., 2015
acetaldehyde	aldehyde	ACETALD	Pina et al., 2014
2-methypropanal	aldehyde	BUTANAL	Pina et al., 2014
Butanal	aldehyde	CPD-7031	Pina et al., 2014
3-methybutanal	aldehyde	METHYLBUT-CPD	Pina et al., 2014
Pentanal	aldehyde	CPD-9053*	Pina et al., 2014
Hexanal	aldehyde	HEXANAL	Pina et al., 2014
Benzaldehyde	aldehyde	BENZALDEHYDE	Pina et al., 2014
Ethanol	short chain aliphatic alcohol	ETOH	Pina et al., 2014
1-butanol	short chain aliphatic alcohol	BUTANOL	Pina et al., 2014
1-pentanol	short chain aliphatic alcohol	PENTANOL*	Pina et al., 2014
2-butanone	short chain ketone	ACETONE	Pina et al., 2014
3,5-octadien-2-one	short chain ketone	MEK	Pina et al., 2014

Table S1 related to Fig 6. Database metabolites.

dichloromethane	halocarbon	CPD-681	Pina et al., 2014
chloroform	halocarbon	CPD-843*	Pina et al., 2014
glycerate	carboxylic acid	GLYCERATE	Belghit et al., 2017
2-methylpropanoic acid	carboxylic acid	ACET	Pina et al., 2014
2-methylbutanoic acid	carboxylic acid	ISOBUTYRATE	Pina et al., 2014
hexane	alcane	CPD-9288*	Pina et al., 2014
2,2,4-trimethylpentane	alcane	CPD-19039*	Pina et al., 2014

* non predicted in initial GSMN reconstruction

Table S2 related to Fig 6. Orphan metabolites.

Usual name	Category	References
Heneicosanoic acid	21:0 fatty acid	Santos et al., 2015
N/A	15:1 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	Belghit 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	Belghit et al., 2017
Octadeca-9-enoic acid	fatty acid	Santos et al., 2015
22-Dehydrocholesterol*	sterol	Tasende et al., 2000
11-hydroxy-octadecadienoic acid (11-HODE)	oxylipin	Gaquerel et al., 2007
13-hydroxy-9Z,11E-octadecadienoic acid (13-HODE)	oxylipin	Gaquerel et al., 2007; Belghit et al., 2017
13S-hydroxy-9Z,11E,15Z-octadecatrienoic acid (13-HOTrE)	oxylipin	Belghit et al., 2017
13-oxo-9Z,11E-octadecadienoic acid (13-oxo-ODE)	oxylipin	Gaquerel et al., 2007
13-hydroxyeicosatrienoic acid (13-HETrE)	oxylipin	Gaquerel et al., 2007
13-hydroxyeicosatetraenoic acid (13-HETE)	oxylipin	Gaquerel et al., 2007
13-hydroxyeicosapentaenoic acid (13-HEPE)	oxylipin	Gaquerel et al., 2007
15-hydroxydocosahexaenoic acid (15-HDHE)	oxylipin	Gaquerel et al., 2007
11-hydroxyoctadecadienoic acid (11-HETE)	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 (QDG2)	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- octadecatetranoyl)-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-eicosapentaenoyl-2-O-palmitoyl-3-O-(β-D-galactopyranosyl-6-1α-D–galactopyranosyl)-glycerol	galactolipid	Banskota et al., 2014
(2S)-1-O-arachidonoyl-2-O-palmitoyl-3-O-(β-D-galactopyranosyl-6-1α-D-galactopyranosyl)-glycerol	galactolipid	Banskota et al., 2014
diphosphatidylglycerol, 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; Guihéneuf et al., 2018

Table S2 related to Fig 6. Orphan metabolites.

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-Monohexadecanoin	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	Pina et al., 2014
2(Z)-penten-1ol	short chain aliphatic alcohol	Pina et al., 2014
3-methylbutanoic acid	carboxylic acid	Pina et al., 2014
2-methylbutanal	aldehyde	Pina et al., 2014
1-octen-3-ol	short chain aliphatic alcohol	Pina et al., 2014
2,6-dimethylpyrazine	short chain ketone	Pina et al., 2014
2-propanone	short chain ketone	Pina et al., 2014
acetic acid, anhydride	carboxylic acid	Pina et al., 2014
2,2,3-trimethylpentane	alcane	Pina et al., 2014
tetradecane	alcane	Pina et al., 2014

* 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.

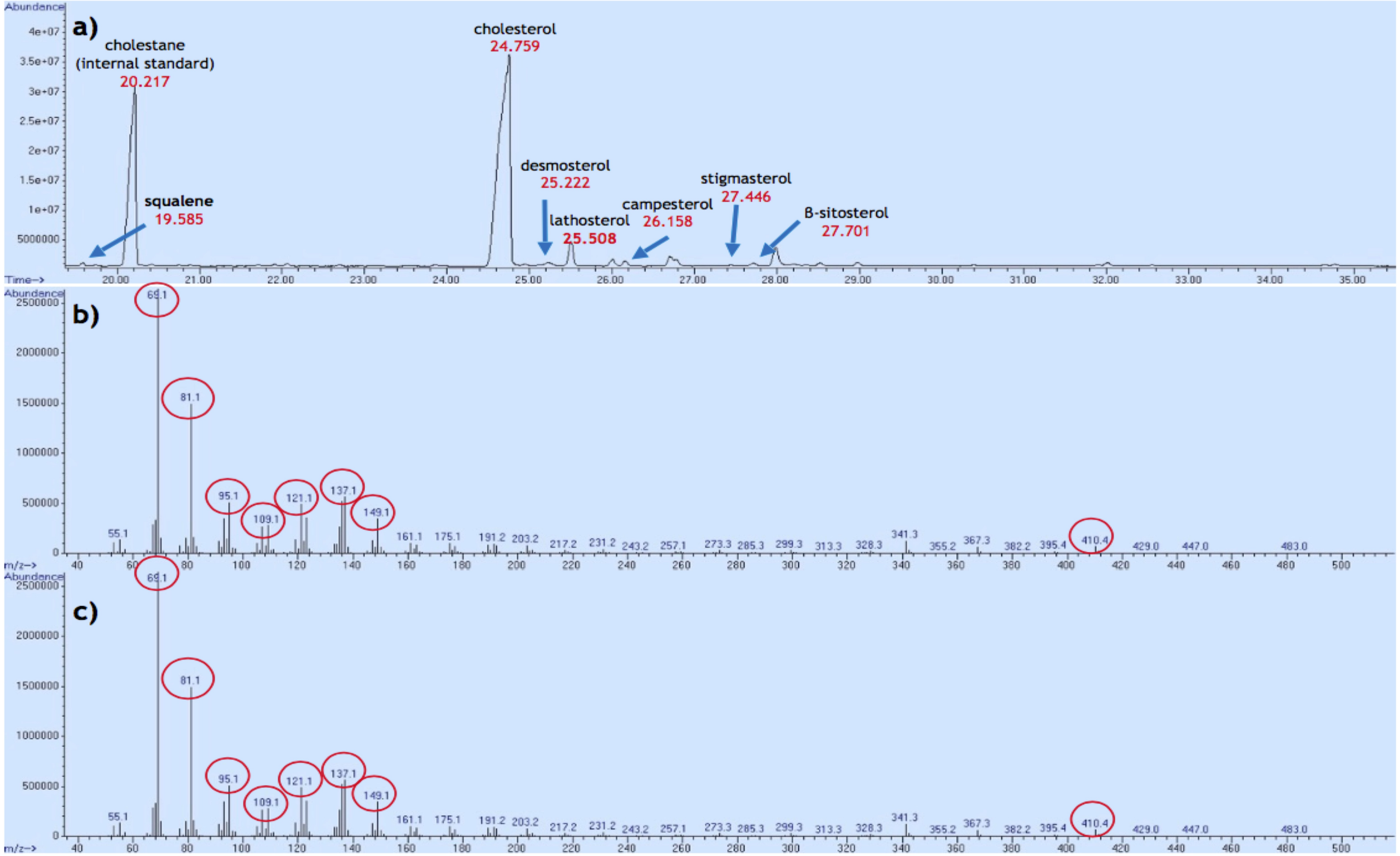

Figure S1, related to Figure 5. Identification of squalene in *C. crispus*. a) Total Ion Chromatogramm (TIC) from *C. crispus* extract. b) MS spectrum of squalene in *C. crispus* extract. c) MS spectrum of the squalene analytical standard. Main fragmentation peaks identical in both spectra are highlighted in red circles.

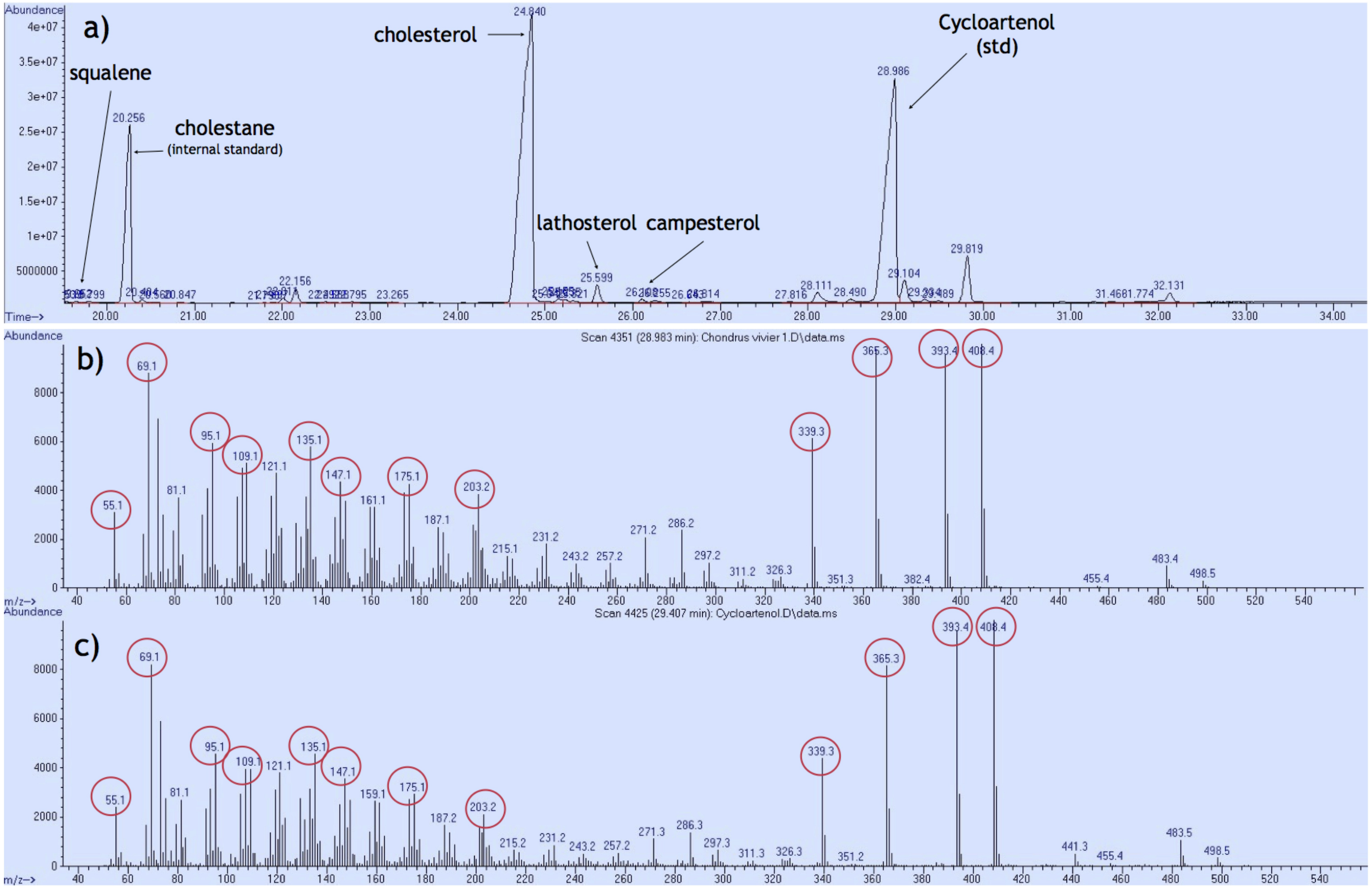

Figure S2, related to Figure 5. Control for technical detectability of cycloartenol in spiked *Chondrus crispus* extract.
a) TIC from *Chondrus crispus* extract incubated with cycloartenol. b) MS spectrum of cycloartenol standard incorporated in *C. crispus* extract. c) MS spectrum of cycloartenol standard alone. Main fragmentation peaks identical in both spectra are highlighted in red circles.

MAAs	Palythine	Mycosporine-glycine	MAA1	Isujirene/Palythene	Asterina-330	Palythinol or MAA2	Shinorine	Porphyra-334
Rt (min.)	8.3	20.0	10.8	19.3	8.7	10.1	18.5	19.5
m/z [M+H]+ observed	245.1090	246.0932	271.1241	285.1401	289.1349	303.1497	333.1245	347.1399
m/z calculated	245.1132	246.0972	271.1288	285.1445	289.1394	303.1551	333.1292	347.1449
EIC (Intens. x108)								
C. crispus (April)	16118542	707375	3254803	209911	5116637	26945	3129533	353130
C. crispus (July)	12600749	85700	928894	36714	3788544	18560	394887	11021
C. crispus (August)	16469850	219296	857212	238033	5653618	32998	1063642	83569
C. crispus (Sept.)	11230824	56477	2546286	77636	2917730	< LOD	5199737	33580
UV (mAU)								
C. crispus (April)	20420	31,525	1541	< LOD	6996	< LOD	2299	117
C. crispus (July)	14578	171,83	1487	< LOD	7106	327,43	335	< LOD
C. crispus (August)	19005	242,7	2143	< LOD	9927	707,57	1245	248
C. crispus (Sept.)	12768	< LOD	989	< LOD	6367	< LOD	5128	136

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	delta7reduction	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-4 demethylation	ERG25	SC4MOL	SMO1, SMO2	CHC_T00010320001, scaffold212*
delta-8, delta-7 isomerisation	ERG2	EBP	HYD1	CHC_T00001257001
C-5 desaturation	ERG3	SC5DL	STE1	CHC_T00006481001
C24 or C24' methylation	ERG6	-	SMT1, SMT2	CHC_T00009101001, CHC_T00000837001
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.

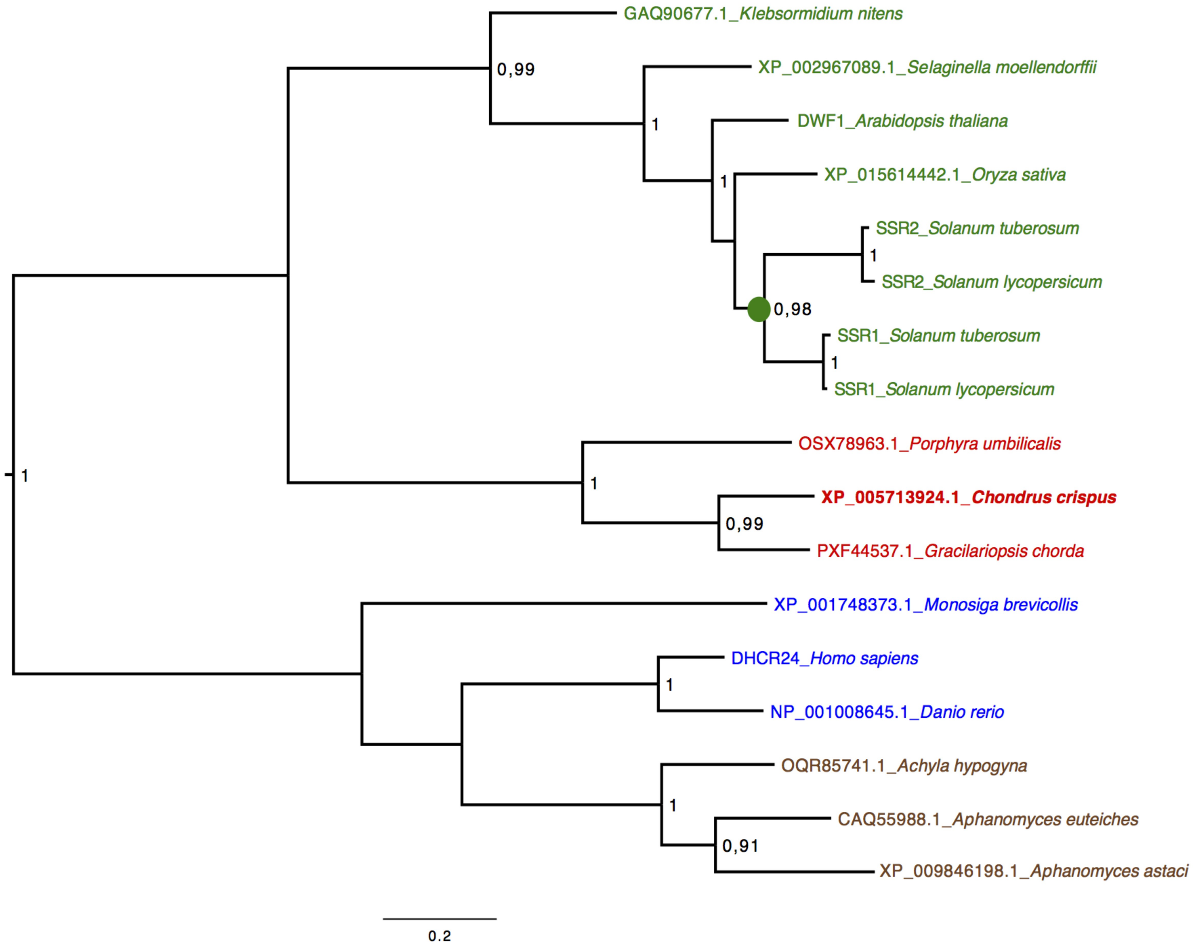
GAQ90677.1_Klebsormidium nitens
0,99
XP_002967089.1_Selaginella moellendorffii
1
DWF1_Arabidopsis thaliana
1
XP_015614442.1_Oryza sativa
SSR2_Solanum tuberosum
1
SSR2_Solanum lycopersicum
0,98
SSR1_Solanum tuberosum
1
SSR1_Solanum lycopersicum
OSX78963.1_Porphyra umbilicalis
1
1
XP_005713924.1_Chondrus crispus
0,99
PXF44537.1_Gracilariopsis chorda
XP_001748373.1_Monosiga brevicollis
DHCR24_Homo sapiens
1
NP_001008645.1_Danio rerio
OQR85741.1_Achyla hypogyna
1
CAQ55988.1_Aphanomyces euteiches
0,91
XP_009846198.1_Aphanomyces astaci
0.2

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
RDGRRVLCVERQLYAPSGALCAPPRIVGELLQPGGYDALCRLGLADALLDIDAQVIRGYA
LFLGPRAERLPYHQPGGPDPDPDPAARPQPEGRAFHNGRFLKRLREIARAHPNV
TLVEGNVLALLERDGAVVGVRYATRGNKAATAHAGLTIAC
DGCGSALRKRAAAHHHVTVYSNFHGLVLHVPALPFPNHGHVVLADPCPVLFYPISATEVR
CLVDIPSTYAGDAAEYILHTVVPQVPPPLRAPLATAVRERRSKMMPNRVMPAPA
HVVPGAVLLGDAFNMRHPLTGGGMTVALTDVELLRELLAPVPDLSDAPAVAAKLQLFYER
RKPMSTTINILANALYTLFCATDDPALRDMRAACLDYLAKGGRMTHDPIAMLGGLKPQRH
LLLAHFFAVALYGCGKALMPFPTPARLVRAWSIFRASFNIIKPLANAEGFWPLSWLPLNSL
```

```
>scaffold20:461442-460650(-) putative squalene epoxidase
LCRLGLADALLHIDAQVIRGYALFLGPRAERLPYHQPGEPDPDPAARPQPEG
RAFHNGRFLKRLREIARAHPNVTLIEGNVLALLERDGAVV
GVRYATRGNKAATAHAGLTIACDGCGSALRKRAAAHHHVTVYSNFHGLVLHVPALPFPNH
GHVVLAHPCPVLFYPISATEVRCLVDL
YILHTVVPQVPPSLRAPLATTVRERRSKMMPNRVMPAPAHVVPGAVLLGDAFNMRHPLTG
GGMTVALTDVELLRGLLAP
```

```
>scaffold57:152407-364140(+) putative squalene epoxidase
RFAGPEHPSCGLKPQRHLLLAHFFAVALYGCGKALMPFPTPARPVRAWSIFRASFNFIK
PLANAEGFWPLSWLPLN
LCRLGLADALLDIDAQVIRGYALFLGPRAERLPY
LCRLGLADALLHIDAQVIRGYALFLGPRAERLPYHQPGGPDPDPAARPQPEG
RAFHNCRFLKRLREIARAHPNVTLIEGNVLALLERDGAVV
GVRYATRGNKAATAHAGLTIACDGCGSALRKRAAAHHHVTVYSNFHGLVLHVPALPFPNH
GHVVLAHPCPVLFYPISATEVRCLVDL
WSTYAGDAAEYILHTVVPQVPPSLRAPLATAVRERRSKMMPNRVMPAPAHVVPGAVLLGD
AFNMRHPLTGGGMTVALTDVELLRGLLAP

>scaffold212:177405-176674(-) putative C-4 sterol methyl oxidase
WDLLCRHTRAYPMFVVGCFASQLAGYFLGCAPFVLLDALRARSTPFRKIQPGKYAPRRAV
FAAAAAMLRSFATVVLPLLAAGGLFIERVGISRDAPFPSPRVVLLQVAYFFLVEDFLNYW
VHRALHLPWLYTRVHSVHHEYDAPFAVVAAYAHPVEVVLQALPTFAGPLMLGPHLYTLCV
WQLFRNWEAIDIHSGYDHAWGLASVLPWYAGPEHHDFHHFLHSGNFASVFTWCDWAYGTD
LAYE

>CHC_T00006492-3001 fusion of adjacent protein predictions CHC_T00006492001 and CHC_T00006493001;
putative sterol delta-7 reductase
MLGIAAWKGFIRYGLLYDHFGEVLAFLGKFALVVTVLLYFRGIYFPTNSDSGTTSFGIVWDMWHGTELHP
EIFGVSLKQLVNCRFALMGWSVAIVAFACKQREQYGYVSNSMLVSVVLQLVYIFKFFVWEAGYFNSVSLD
HSHVCLFWIYLRPLY
MVGVGAICCNYWTDKQREVFRATNGQVTIWGQKPVSIEAQYVTGDGKKRRSLLLASGWWGVSRHVNYVFE
IALTFCWSVPAGGTGVIPYVYVMFLTILLTDRAYRDEVRCSEKYGKYYEEYCRLVPYKMIPGVY
```

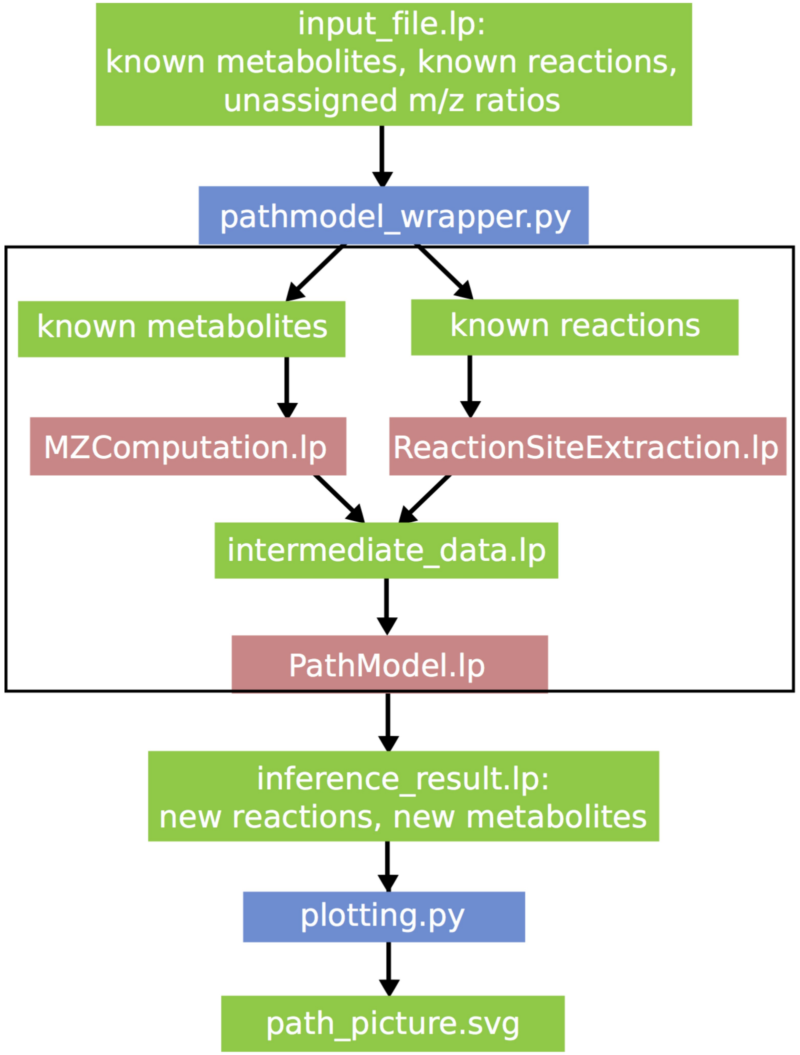

Figure S4, related to Figure 3. Architecture of Pathmodel scripts. In green: Data files, either input or result files. In red: ASP scripts. In blue: Python scripts. The black line shows the wrapping of all the scripts inside by pathmodel_wrapping.py.

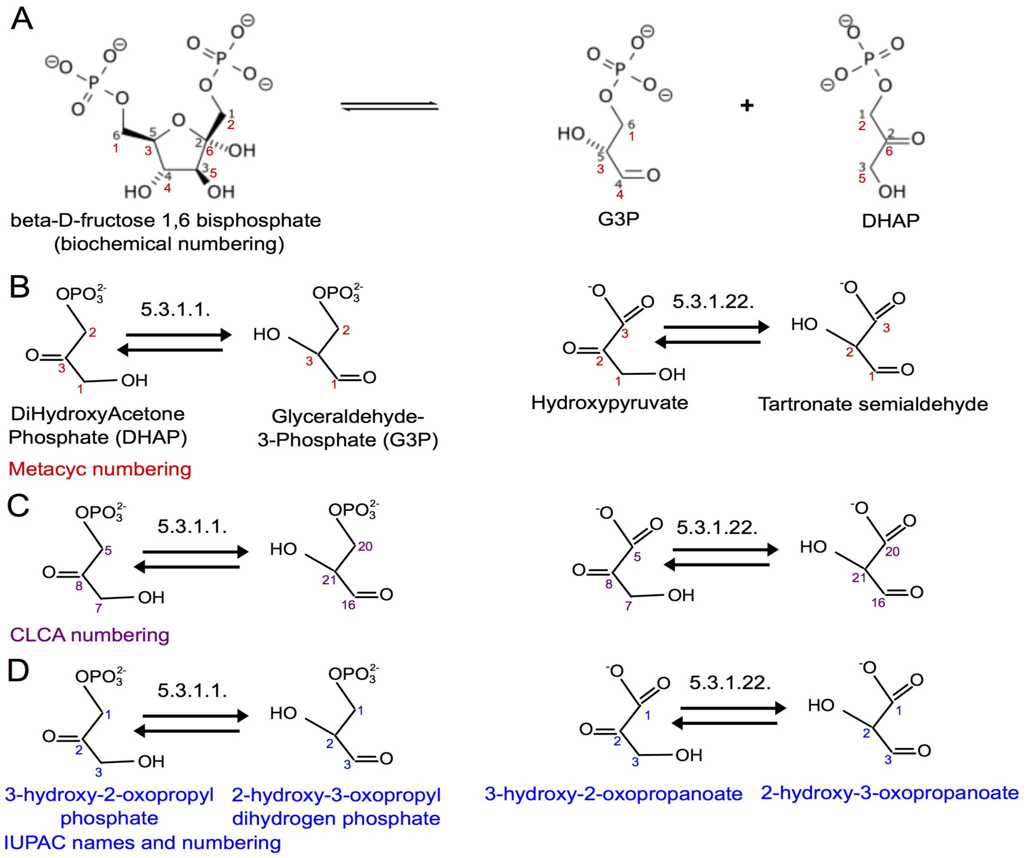
A
beta-D-fructose 1,6 bisphosphate
(biochemical numbering)
G3P
DHAP
B
5.3.1.1.
5.3.1.22.
DiHydroxyAcetone Phosphate (DHAP)
Glyceraldehyde-3-Phosphate (G3P)
Hydroxypyruvate
Tartronate semialdehyde
Metacyc numbering
C
5.3.1.1.
5.3.1.22.
CLCA numbering
D
5.3.1.1.
5.3.1.22.
3-hydroxy-2-oxopropyl phosphate
2-hydroxy-3-oxopropyl dihydrogen phosphate
3-hydroxy-2-oxopropanoate
2-hydroxy-3-oxopropanoate
IUPAC names and numbering

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.