Inferring biochemical reactions and metabolite structures to cope with metabolic pathway drift

Arnaud Belcour[1¶], Jean Girard[2¶], Méziane Aite[1¶], Ludovic Delage[2¶], 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], Gabriel V. Markov[2]*

[1]Univ Rennes, Inria, CNRS, IRISA, Equipe Dyliss, Rennes, France

[2]Sorbonne Université, CNRS, Integrative Biology of Marine Models (LBI2M), Station Biologique de Roscoff (SBR), 29680 Roscoff, France,

[3]LBCM, IUEM, University of Bretagne-Sud, Lorient, France

[4]Sorbonne Université, CNRS, Plateforme METABOMER-Corsaire (FR2424), Station Biologique de Roscoff, Roscoff, France

[5]Sorbonne Université, CNRS, Plateforme ABiMS (FR2424), Station Biologique de Roscoff, Roscoff, France

*Corresponding author

Email: Gabriel.Markov@sb-roscoff.fr

[¶] These authors contributed equally to this work.

Abstract

Inferring genome-scale metabolic networks in emerging model organisms is challenging because of incomplete biochemical knowledge and incomplete conservation of biochemical pathways during evolution. This limits the possibility to automatically transfer knowledge from well-established model organisms. Therefore, specific bioinformatic tools are necessary to infer new biochemical reactions and new metabolic structures that can be checked experimentally. Using an integrative approach combining both genomic and metabolomic data in the red algal model *Chondrus crispus*, we show that, even metabolic pathways considered as conserved, like sterol or mycosporine-like amino acids (MAA) synthesis pathways, undergo substantial turnover. This phenomenon, which we formally define 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 new methodological approach to formalize the logical reasoning necessary to infer new reactions and new molecular structures, based on previous biochemical knowledge. We use this approach to infer previously unknown reactions in the sterol and MAA pathways.

Author summary

Genome-scale metabolic models describe our current understanding of all metabolic pathways occuring in a given organism. For emerging model species, where few biochemical data are available about really occurring enzymatic activities, such metabolic models are mainly based on transferring knowledge from other more studied species, based on the assumption that the same genes have the same function in the compared species. However, integration of metabolomic data into genome-scale metabolic models leads to situations where gaps in pathways cannot be filled by known enzymatic reactions from existing databases. This is due to structural variation in metabolic pathways accross evolutionary time. In such cases, it is necessary to use complementary approaches to infer new reactions and new metabolic intermediates using logical reasoning, based on available partial biochemical knowledge. Here we present a proof of concept that this is feasible and leads to hypotheses that are precise enough to be a starting point for new experimental work.

Introduction

Reconstruction of genome-scale metabolic networks (GSMs) is a useful and powerful way to integrate data about the metabolism of model organisms due to the increasing availability of genome data [1]. In parallel, metabolomics has maturated as a separate research field, and both are now converging, with the proposal to focus on model organism metabolomes [2]. However, integrating genomic and metabolomic data remains challenging, partly because not all metabolites are indexed in the databases used for GSM reconstruction. For the set of core models prioritized for such approaches, a list of experimentally identified metabolites is not always maintained in a publicly available database [3], and this issue become even more problematic for emerging model species, for which a genome is or will be sooner sequenced, but where the community collecting experimental data is rather limited. Macroalgae belong to this second group of emerging models, which is experiencing drastic changes in research practice due to the availability of high-throughput omics tools [4]. Despite extensive discussion on quality criteria in the field of genome-scale metabolic model reconstruction [5], one missing piece of information is the proportion of metabolites incorporated into the genome-scale metabolic model that are actually described in the literature. Literature data are acknowledged as an important source of knowledge to incorporate into GSMs [6], but current databases tend to point towards bibliographical references concerning pathways or single reactions rather than providing information about the presence of metabolites. A recent survey on 391 metabolites from 21 red, brown, and green macroalgae showed that only 184 of those metabolites were indexed into the PlantCyc database [7]. As a response to this, the metabolomic community is organizing the automation of the taxonomic assignation of metabolites [8].

Integrating data on metabolite presence/absence into GSMs is especially important when working on emerging model organisms that are phylogenetically distant from well-established models, because there are many ways to generate variations during evolutionary time, even within metabolic pathways that may appear to be conserved at first glance. Indeed, even for the well-studied human pathogenic bacterium *Mycobacterium tuberculosis,* high throughput metabolomic screens revealed an unexpected diversity of reactions in central carbon metabolism [9]. Evolutionary models have already been developed to explain the arising of new pathways, with most experimental validations being focused so far at the level of individual enzyme activities [10]. The complementary question, how much conserved

pathways remain stable in terms of enzymes, has not yet been addressed in a systematic way. However, very similar issues have been tackled in other subfields of evolutionary biology, and can thus be exported to the field of metabolic pathway evolution.

Developmental system drift has been evidenced some decades ago in the field of animal comparative biology, to explain how morphologically similar structures can be maintained even if there are substantial variations in the molecular mechanisms underlying their formation [11]. The concept was more recently extended to plants, where such cases have been observed in leaf development [12]. It was later exported to the fields of protein evolution [13] and gene expression evolution [14]. We hypothesize that this evolutionary concept also adequately explains the strict conservation of metabolic pathways due to enzymatic replacement by non-orthologous displacement of genes encoding enzymes with identical biochemical function ([15]; Figure 1). A second possible mechanism for metabolic pathway drift, that has the potential to generate observable biochemical diversity in pathways is change in enzyme order, which leads to new biosynthetic intermediates without other changes than their order of intervention (Figure 1). To the best of our knowledge, this second possibility has never been formulated in theory, maybe due to difficulties envisioning an experimental setup to test it.

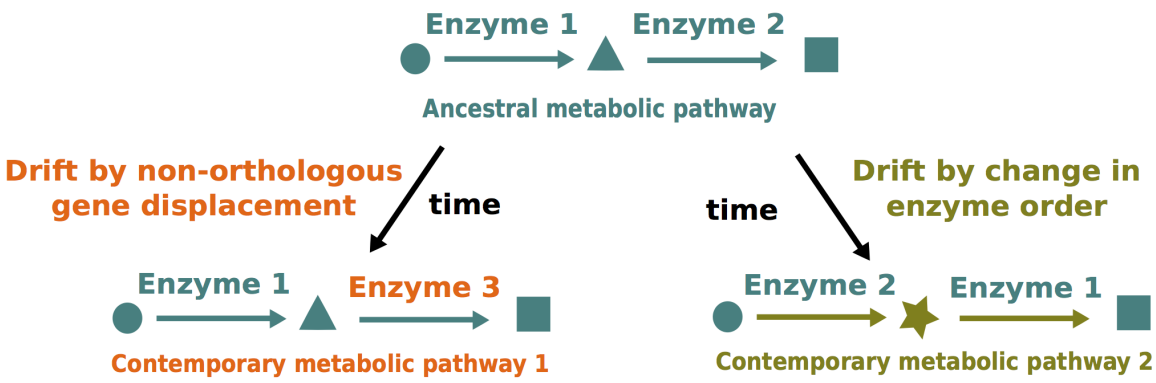


Fig 1. Two possible elementary mechanisms for metabolic pathway drift. Starting from an ancestral pathway (in teal, upper part), changes can occur either by non-orthologous gene displacement (in orange, left side) or change in enzyme order, leading to new metabolites (in olive green, right side).

118 Regarding non-orthologous gene displacement, classical comparative genomic approaches
119 can generate hypotheses that can be experimentally checked using targeted metabolic
120 profiling combined with enzyme inactivation by CRISPR-Cas9 [16]. However, in case of drift
121 by change in enzyme order, an additional theoretical step is necessary to formally infer the
122 structure of new intermediary metabolites and new enzymatic reactions before experimental
123 validation. For such approaches, most of the time, the new reaction has never been observed
124 in any organism, so that an approach purely based on a search in a database of known
125 reactions is doomed to failure [17]. It is necessary to introduce a knowledge-based approach
126 that implements reasoning in the manner of a biochemist. Such strategies have already been
127 used for designing experimental setup for the analysis of auxotrophic mutants in yeast [18] or
128 for synthetic biology [19]. To be successful, approaches of scientific discovery based on
129 artificial intelligence techniques necessitate close and iterative interactions between chemists,
130 biologists and bioinformaticians [20-22]. This has already led to promising results in the field
131 of drug screening for neglected tropical diseases [23]. To further test the hypothesis of
132 metabolic pathway drift, we decided to combine GSM reconstruction and metabolic profiling,
133 the latter based on a bibliographic survey and mass spectrometry analyses, in an emerging
134 model, the red alga *Chondrus crispus*.

135 *C. crispus* is a red seaweed that has been subject to biological studies for more than two
136 centuries [24]. Its genome was sequenced, and annotation was performed with a focus on
137 metabolic features [25]. A non-exhaustive bibliographic search enabled us to find 15 papers
138 mentioning the identification of metabolites from *C. crispus* by various methods of chemical
139 profiling. Nine of them were specifically focused on *C. crispus* [26-34], whereas the six
140 others were comparative studies between several algae [35-40]. We selected these papers as a
141 test case for incorporating the bibliographic knowledge into a GSM. Additionally, we decided
142 to acquire additional experimental data regarding two pathways chosen for their
143 complementary interest: sterols and mycosporine-like amino-acids (MAAs). The sterol
144 pathway is well investigated at the comparative genomics level [41] and consists mainly of
145 oxidoreductions on a known skeleton, the sterane, consisting of three hexacarbon rings on one
146 pentacarbon ring [42]. Analytical standards are available for different molecules, enabling
147 level 1 metabolite identification by mass spectrometry, according to the metabolomics
148 standard initiative [43]. MAA synthesis involves combination of different building blocks,
149 and analytical standards are lacking for this class of compounds, limiting metabolite

150 identification to level 2 in best cases [43]. Here, using a logical representation of molecules
151 and reactions, we propose the development of an analogy reasoning model and the use of a
152 generic solver to produce all possible inferences. Integrating this with a global analysis of the
153 genome-scale metabolic network, with targeted experimental profiling, and with comparative
154 genomic analysis, we are able to propose an exhaustive model for two metabolic pathways in
155 *C. crispus*, structurally shaped by metabolic pathway drift.

156

157 ## Results

158

159 **Chemical identification of main sterols in *C. crispus*, but not of some plant-like**
160 **biosynthetic intermediates by targeted GC-MS profiling**

161

162 Results of targeted profiling of 15 sterols plus one immediate precursor (squalene) are
163 summed up in Table 1.

Analysed compounds	Chemical formula	Found in this study	Previous evidence
brassicasterol	C28H46O	yes	[44] (GC-MS), [29] (TLC, GC-MS)
campesterol	C28H48O	yes	[29] (TLC, GC-MS)
cholesterol	C27H46O	yes	[44] (TLC, GC-MS), [29] (TLC, GC-MS)
cycloartanol	C30H52O	no	not reported
cycloartenol	C30H50O	no	[44] (TLC), [35] (TLC)
cycloeucalenol	C30H50O	no	not reported
7-dehydrocholesterol	C27H44O	yes	[29] (TLC, GC-MS)
desmosterol	C27H44O	yes	[44] (TLC, GC-MS), [45] (GC-MS)
ergosterol	C28H44O	no	not reported
fucosterol	C29H48O	no	not reported
lanosterol	C30H50O	no	[44] (TLC)
lathosterol	C27H46O	yes	[45] (GC-MS)

β-sitosterol	C29H50O	yes	[44] (GC-MS), [29] (TLC, GC-MS)
squalene	C30H50O	yes	not reported
stigmasterol	C29H48O	yes	[44] (GC-MS), [29] (TLC, GC-MS)
zymosterol	C27H44O	no	not reported

Table 1. 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 S1 Table.

In addition to confirm the presence of eight previously identified sterols (brassicasterol, campesterol, cholesterol, 7-dehydrocholesterol, desmosterol, lathosterol, β-sitosterol and stigmasterol), we identified here for the first time one immediate precursor, squalene (S1 Fig) in *C. crispus*. However, we did not find evidence for some other putative intermediates (cycloartanol, cycloeualcenol, ergosterol, fucosterol and zymosterol) that may have been present based on the knowledge of sterol synthesis pathway in other eucaryotes [41, 46]. We also did not find cycloartenol in *C. crispus* extracts despite the fact that we are able to identify the cycloartenol standard when added in algal extract (S2 Fig). Cycloartenol has been reported a long time ago in *C. crispus* extracts from Roscoff using another analytical technique, thin-layer chromatography (TLC) [35]. However, the same group was unable to isolate cycloartenol from another red alga, *Rytiphlea tinctoria*, using GC-MS [47]. Independently, Saito and Idler [44] isolated lanosterol instead of cycloartenol from *C. crispus* using TLC, but also failed to find lanosterol using GC-MS. More recently, 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, but the authors did not report cycloartenol identification in the whole alga by GC-MS, as they did for cholesterol (Calegario et al., 2016). Even if undetectable using GC-MS, another indirect argument for cycloartenol as a biosynthetic intermediate is the presence of a compound with a cyclopropyl ring in another florideophyte red alga, *Tricleocarpa fragilis* [48]. The cyclopropyl ring on sterols is usually made by oxidosqualene cyclisation, and the only described product of this reaction is cycloartenol, so we consider more parsimonious to hypothesize that cycloartenol is below the detection limit rather than considering that this step is performed by an unknown intermediate.

190 **An unknown compound among most abundant MAAs in *C. crispus***

191 Results of LC-MS targeted profiling of mycosporin-like aminoacids are summed up in Table
192 2.

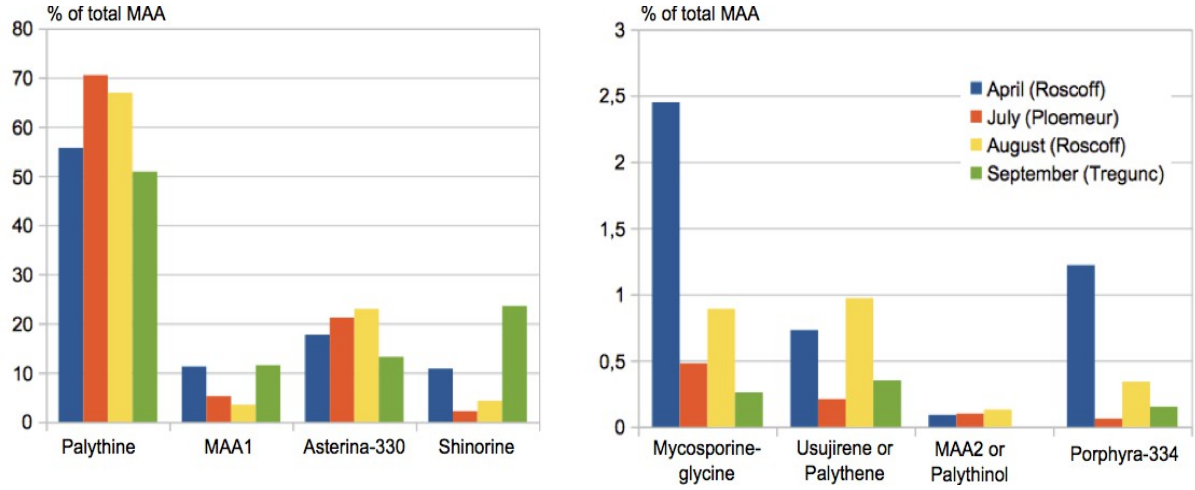
Analysed compounds	Chemical formula	Found in this study	Previous evidence
Asterina-330	C12H20N2O6	yes	[49] (LC-MS-MS); [50] (LC-MS)
MAA1	compatible with m/z 271.1241	yes	not reported
MAA2	compatible with m/z=302,3117	yes	not reported
Mycosporin-glycine	C10H15NO6	yes	not reported
Palythine	C10H16N2O5	yes	[51] (UV+LC-MS); [49] (LC-MS-MS); [50] (LC-MS)
Usujirene/Palythene	C27H46O	yes	[51] (UV+LC-MS)
Palythinol	C30H52O (m/z=302,3117)	no	[51] (UV+LC-MS), [49] (LC-MS-MS)
Porphyra-334	C30H50O	yes	[49] (LC-MS-MS)
Shinorine	C30H50O	yes	[51] (UV+LC-MS), [49] (LC-MS-MS)

193

194 **Table 2. List of mycosporin-like amino-acids identified in this study, and comparisons with**
195 **previous ones.** For each compound, analytical parameters (RT, mz and UV absorption parameters)
196 are given in S2 Table.

197 Using LC-MS profiling, we confirmed, consistently with previous studies (see references in
198 Table 2), the presence of six mycosporine-like aminoacids in *C. crispus*: asterina-330,
199 palythene, palythine, palythinol, porphyra-334 and shinorine. Additionally, we identified
200 mycosporine-glycine for the first time in *C. crispus*, and also found a peak at m/z=271.1 that
201 does not match with any already identified candidate MAA, that we named it MAA1 in Table
202 2. We also decided not to assign the peak at m/z=302,3117 to palytinol, as done previously
203 [49, 51], based on logical reasoning about this part of the pathway (see below). That is the
204 reason why an other unknown compound, MAA2, appears in the table. The relative
205 abundance of MAAs seems to vary according to the sampling dates (Fig 2).

206



207 **Fig 2. Composition and seasonal variation *(MS quantification)* of MAAs in *C. crispus.***

208 These results should be interpreted carefully because MAAs are known to react differently to
209 MS ionization. Furthermore, under UV, the molar extinction coefficients are different. This
210 allows only semi-quantitative measurements. However, our results are consistent with an
211 independent report of MAA variation in the Galway Bay, Ireland [50]. In both cases,
212 palythine was the most abundant compound. Depending on localisation, and time, then
213 shinorine and asterina-330 were the most abundant compounds, and porphyra-334 was very
214 scarce. The unknown compound at m/z 271.1241, which was here labelled MAA1, is the
215 fourth most abundant MAA in Brittany samples from *C. crispus*.

216 Our new metabolite profiling data on sterols and MAAs were pooled with results from other
217 studies, retrieved by bibliographic search, in order to obtain a set of metabolite targets that
218 was used to constrain the genome-scale reconstruction of the *C. crispus* metabolic network
219 (S3-S4 Tables).

220

221

222 **Several non-orthologous genes encode best candidate enzymes for performing conserved**
223 **reactions in the sterol synthesis pathway**

224 In order to enable comparisons with the automated genome-scale reconstruction, and to
225 facilitate integration with metabolic profiling data, we carried out a comparative genomic
226 analysis of the enzymes involved in the sterol synthesis pathways. Results are summed up in
227 Table 3.

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 3. Comparative genomic analysis of sterol synthesis enzymes. In the first column, the color code for enzymatic steps follows Desmond and Gribaldo [41]. In the four other column, 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 S1 Dataset.

In line with previous analyses on these gene families in eukaryotes [41] or more specifically in green plants [46], 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

239 epoxidase and C-4 demethylase) or were lost and may have been replaced by distant paralogs
240 (C24 and C24' methylases and C22 desaturases). In one case, we found no homolog of known
241 plant or animal enzymes performing delta-7/delta-8 isomerisation at all in the *C. crispus*
242 genome, but we found a 1:1 ortholog of ERG2, the gene that secondarily took up this function
243 in yeast [41]. We consider this gene the best candidate to test among known gene families, but
244 it is also possible that this reaction is performed by an enzyme encoded by a taxonomically-
245 restricted orphan gene, which have been shown to have some biological roles in other lineages
246 [52]. Actually, this is likely the case in the sterol synthesis pathway in some diatoms, where
247 the epoxisqualene cyclase, otherwise conserved in eucaryotes, was secondarily lost and
248 replaced by another yet unidentified enzyme [53].

249 We did not carry out a similar genomic analysis for MAAs, because it was already fully done
250 during the annotation of the *C. crispus* genome [25], and was recently put in a comparative
251 perspective following the annotation of the MAA genes in an other red alga, *Porphyra*
252 *umbilicalis* [54].

253

254 **Integration of genome-scale reconstruction and targeted chemical profiling highlights**
255 **the need for *ab initio* inferences to fill knowledge gaps**

256 A global overview of the procedure used to build an integrated metabolic network model for
257 *C. crispus* is shown in Fig 3. The network is browsable at:

258 http://gem-aureme.irisa.fr/ccrgem/index.php/Main_Page

259

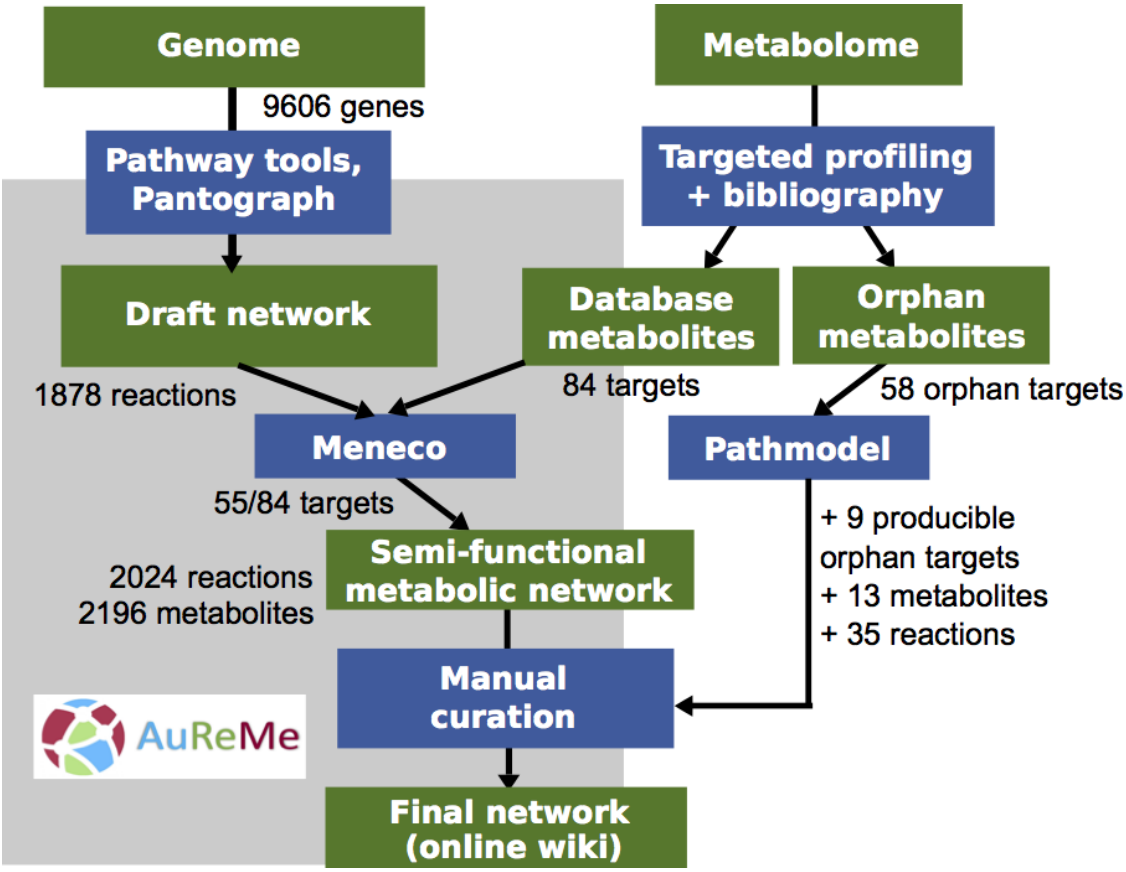


Fig 3. Reconstruction scheme for the genome-scale metabolic network of *C. crispus*. Green boxes indicate starting data and resulting new knowledge. Blue boxes indicate the tools that were used to analyse and integrate genome and metabolome data. The part overshadowed in grey indicates tools that are already integrated in the AuReMe workflow [55].

The final network contains 595 reactions coming directly from genome annotation through PathwayTools, 383 reactions coming from orthology with *Arabidopsis thaliana*, 1361 reactions coming from orthology with *Galdieria sulphuraria*, and 1161 reactions coming from orthology with *Ectocarpus siliculosus*. The total number of reactions in the fused network (2024) is in the same range as in the networks of two other macroalgae, *E. siliculosus* (1977) and *E. subulatus* (2074), reconstructed also using the AuReMe toolbox (Table 4).

275

Species	Reactions	Enzymes	Metabolites	Pathways	Reference
C. crispus	2024	2006	2196	1108	This study
E. siliculosus	1977	2281	2132	1101	[55]
E. subulatus	2074	2445	2173	1083	[56]
A. thaliana	1567	1419	1748	796	[57]
C. reinhardtii	3083	1355	1133	522	[58]

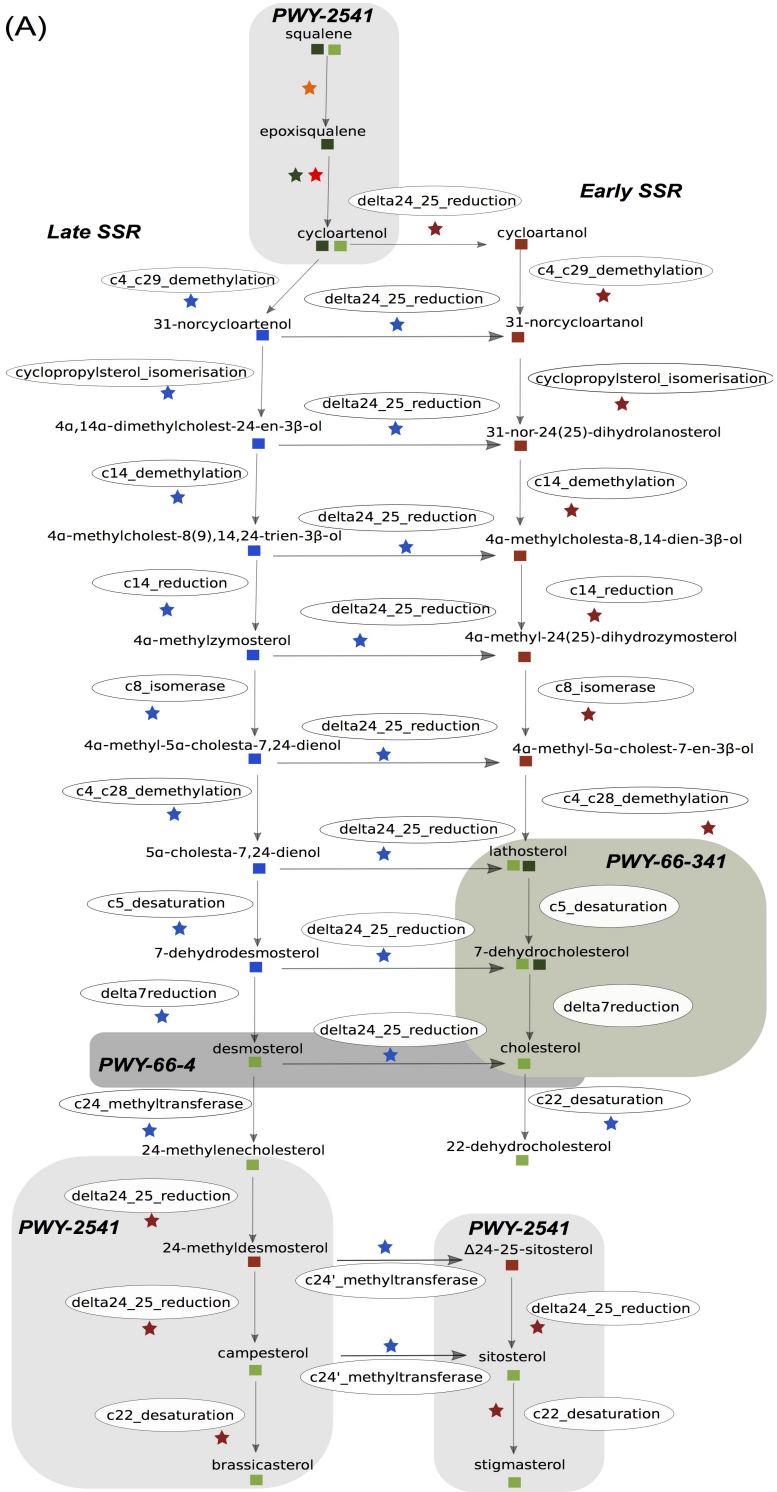
276

277 **Table 4. Comparison of global features of genome-scale metabolic networks from macroalgae**
278 **and other chlorophyllian eucaryotes**

279 Detailed manual comparisons of the networks from *E. siliculosus* and *E. subulatus* have
280 shown that all the differences between them are due to technical biases during the
281 reconstruction process [56]. The *C. crispus* network once again illustrates the high sensitivity
282 of the results to the quality of the input data. Due to differences in the annotation level,
283 annotation-based reconstruction gave different results between the two *Ectocarpus* species
284 (1661 or 1779 predicted reactions) and in *C. crispus* (595 predicted reactions), while
285 orthology-based transfer of central metabolism reactions from *Arabidopsis thaliana* led to a
286 similar number of reactions in all three algae (383 in *C. crispus*, 440 in *E. siliculosus* and 421
287 in *E. subulatus*). More than half of the reactions (1361 out of 2024) were transferred based on
288 orthology from the red microalga *Galdiera sulphuraria*, which was selected after inspection
289 of the automatically reconstructed annotation-based network available in the MetaCyc
290 database. This illustrates the usefulness of the AuReMe pipeline to efficiently correct for
291 annotation biases using orthology information. Interestingly, orthology-based transfer of
292 reactions from *E. siliculosus* to *C. crispus* was more successful than orthology transfer from
293 *A. thaliana* (1161 versus 383 reactions), despite the fact that *A. thaliana* is more closely
294 related to *C. crispus* than to *E. siliculosus*. This clearly shows how technical issues interfere
295 with biology : the higher number of transferred reactions from *E. siliculosus* is linked with the
296 fact that, since both networks being already incorporated in the AuReMe pipeline through the
297 PADMet format, correspondences of reactions IDs were easier than with the *Arabidopsis*
298 network which was reconstructed using a different workflow [57]. Variability comes also

299 from the level of database completeness: the increase of database completeness with time is
300 striking when comparing the reaction numbers between *A. thaliana* [57] and *C. reinhardtii*
301 [58].

302 Another important comparison level is the number of metabolites for which the presence in *C.*
303 *crispus* is experimentally proven. 84 metabolites from *C. crispus* were indexed in the
304 MetaCyc database and could thus be used as targets, that should be present in the final
305 network. Using the gap-filling program Meneco, we managed to incorporate 55 of them,
306 which is higher than the two *Ectocarpus* species (50 targets), but still only represents two
307 thirds of all targets. In addition, there were 58 orphan metabolites that could not be
308 incorporated automatically, because they were not yet indexed in MetaCyc. This prompted us
309 to develop Pathmodel, a new tool that enables to infer new metabolic reactions and new
310 molecules to connect orphan metabolites with the main network. This tool was tested on the
311 sterol and mycosporine-like amino-acid synthesis pathways because they were suitable to
312 address complementary issues. The sterol pathway raised the problem of connecting and
313 integrating various portions of known sterol synthesis pathways from animals and plants (Fig
314 4A) while the MAA pathway raised the problem of integrating unannotated compounds that
315 were identified uniquely based on their m/z ratio (Fig 4B).



317 **Fig 4. Overview of pathways reconstructed with Pathmodel using multiple heterogenous data**
318 **and analogical reasoning.** (A) Sterol biosynthesis pathway. Portions that are already in the MetaCyc
319 database are highlighted with grey boxes. PWY-2541: plant sterol biosynthesis pathway. PWY-66-
320 34: animal modified Kandutsch-Russell pathway. PWY-66-4: animal Bloch pathway. See part (B)
321 fore the detailed figure legend.

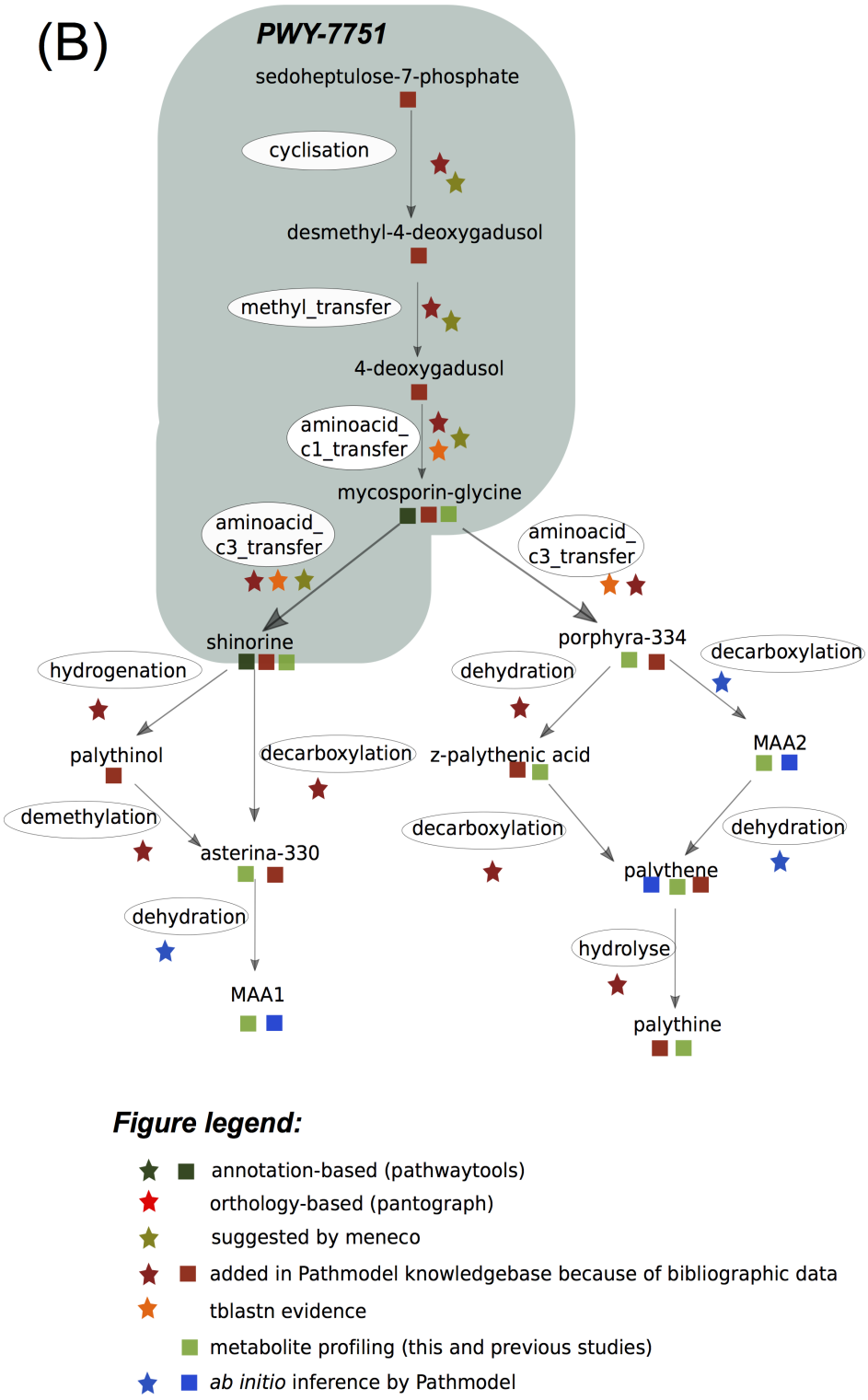


Figure 4 (Continued). (B) MAA Biosyntesis pathway. PWY-7751: shinorine biosynthesis pathway. The figure legend details the various data sources integrated to infer the pathways. Stars indicate reactions, squares indicate molecules.

325 The metabolites present in *C. crispus* only partially fitted with standard pathways indexed in
326 the MetaCyc database, for different reasons. Regarding the sterol synthesis pathways, they
327 belong to three different pathways: cycloartenol, 24-epicampesterol, brassicasterol, sitosterol
328 and stigmasterol belong to the canonical plant sterol biosynthesis pathway (PWY-2541; [59]),
329 whereas lathosterol, 7-dehydrocholesterol belong to the animal modified Kandutsch-Russell
330 pathway (PWY-66-341, [60]) and desmosterol belongs to the animal Bloch pathway (PWY-
331 66-4, [60]). Both animal pathways result in the production of cholesterol at their end.
332 Additionally, 22-dehydrocholesterol is not a part of any of those pathways. Regarding MAAs,
333 mycosporin-glycin and shinorine belong to the shinorine biosynthesis pathway (PWY-7751),
334 corresponding to the best understood part of the pathway [61], but all other compounds
335 identified in *C. crispus* are absent from the MetaCyc database. Morover, the query of public
336 chemical structure databases cannot help in assigning a tentative structure to the peak
337 corresponding to MAA1. All those limitations explain why we selected those two pathways as
338 case studies to develop the Pathmodel method to infer *ab initio* new reactions and new
339 metabolites.

340

341 **Inferring new metabolic reactions using Answer Set Programming**

342 The Pathmodel method takes as input a knowledge base including a set of known metabolites,
343 a set of observed mass-to-charge (m/z) ratios for unknown metabolites, and a set of known
344 enzymatic reactions. For each pair of metabolites which are not linked by a reaction in the
345 knowledge base, the method checks whether a type of known reaction can occur between
346 them, and further derives from known reactions new candidate metabolites corresponding to
347 observed unassigned mass-to-charge ratios. This is the basis for the selection of new reaction
348 occurrences and/or new metabolites, using either deductive or analogical reasoning (Fig 5).

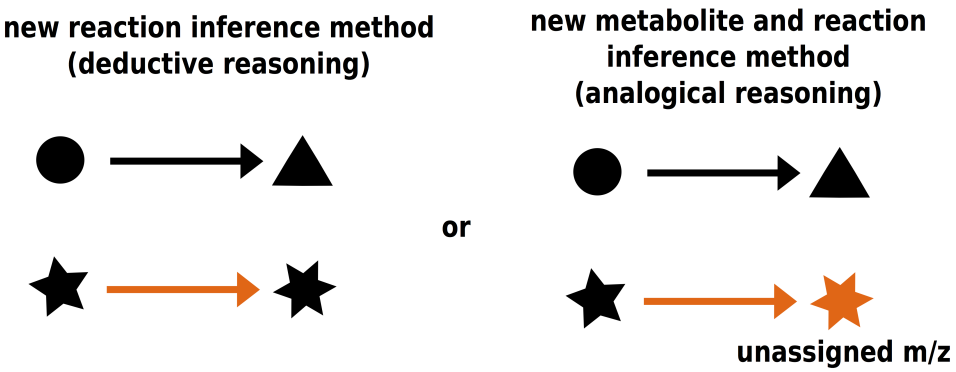


Fig 5. The two reasoning methods implemented in Pathmodel. Input data encoded in the knowledge base are in black, newly inferred reactions and metabolites structures are in orange.

Molecules are modeled by a set of logical predicates *atoms* (identified by a number and atom types) and *bonds* (identified by atom numbers and bond type), as highlighted in orange on Fig 6.

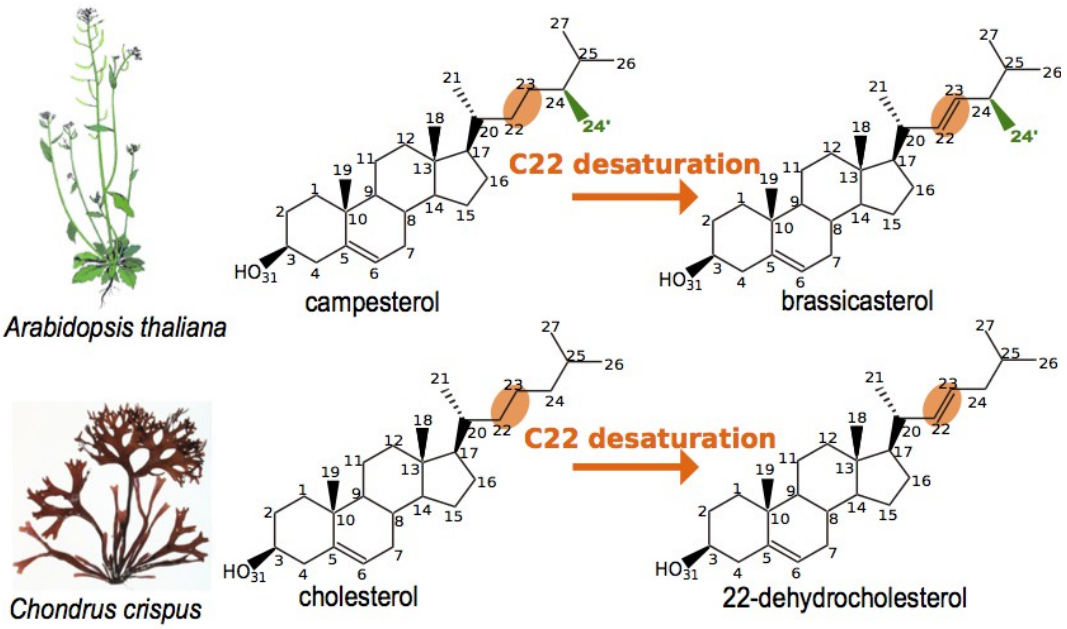


Fig 6. Detailed encoding of metabolites and reactions in Pathmodel. In black, molecules structures with carbon and oxygen atoms labelled. For example, carbon 22 from brassicasterol is encoded by the predicate *atom("brassicasterol",22,carb)*. In orange, position of the bond between atoms submitted to the chemical reaction, encoded by the predicate *bond ("brassicasterol",double,22,23)*. In green: the C24 methyl group that makes the difference between molecules from *Arabidopsis thaliana* and from *Chondrus crispus*.

In order to perform this reasoning, the program needs some preprocessing steps (S4 Fig). For each newly inferred molecule, the theoretical m/z ratio is determined by logical rules, which was encoded in the program MZComputation.lp. First, the number of hydrogens for each atom of a molecule was deduced (predicate *numberHydrogens)* from the total number of bonds in which the atom is involved and from the valence of the atom. Then the number of each atom species (hydrogens, carbons, …) is determined for each molecule (predicate *moleculeComposition*). Finally, the m/z ratios are derived from the molecular composition (predicate *moleculeMZ*), using the following formula:

```
 moleculeMZ(MoleculeName, MassCarbon*NumberCarbon +
MassHydrogen*NumberHydrogen + MassOxygen*NumberOxygen +
MassNitrogen*NumberNitrogen + MassPhosphorus*NumberPhosphorus):-
moleculeComposition(MoleculeName, NumberCarbon, NumberHydrogen,
NumberOxygen, NumberNitrogen, NumberPhosphorus).
```

In this formula, atomic weights are encoded following the latest IUPAC Technical Report [62], truncated after the fourth decimal and multiplied by 1000 because the ASP syntax does not allow the use of decimals.

The predicate *reaction* models the link between two molecules (a reactant and a product, e.g. *reaction(c22_desaturation,"24-epicampesterol", "brassicasterol")*). By comparing reactants and products, the program ReactionSiteExtraction.lp characterizes two structures of the reaction site containing atoms and bonds involved in the reaction: one structure describes the reaction site before the reaction (Figure 6, simple bond between atoms 22 and 23 in campesterol) and the other describes the reaction site after the reaction (Fig 6, double bond 22-23 in brassicasterol). 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. These are the bases for the selection of new potential reactants or products and the inference by a reasoning component of new reaction occurrences or new metabolites, using either deductive or analogical reasoning in the PathModel.lp program.

By deductive reasoning, the reference molecule pair of each reaction (Fig 6, campesterol and brassicasterol) is compared to the structures of a potential reactant-product pair sharing a common chemical structure (Fig 6, cholesterol and 22-dehydrocholesterol, sharing a sterane skeleton) with the predicate *deductiveReasoningInference*. The presence of the reaction site in the two putative molecules is checked by using the predicates *siteBeforeReaction* and *siteAfterReaction*. Furthermore, if the product and the reactant have the same overall structure, except for the reaction site (see Fig 6, bond between atoms 22 and 23), the program will infer that the reaction actually occurs between the reactant and the product (Fig 6, desaturation between cholesterol and 22-dehydrocholesterol). To constraint further the number of possible pathways, a predicate *absentmolecules* was added to avoid pathways going through compounds for which targeted profiling with analytical standards gives strong evidence for real absence (here ergosterol, fucosterol and zymosterol).

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 observed m/z ratio (predicate *possibleMetabolite*) and checked if they share the same structure as a known molecule (predicate *alreadyKnownMolecule*).

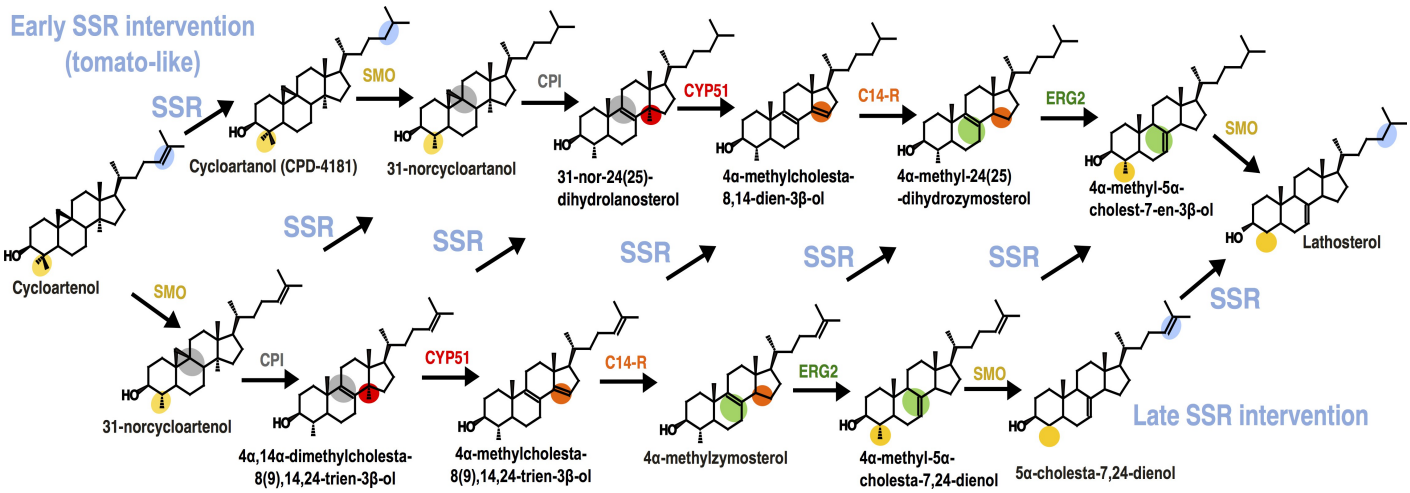
Given a source molecule and a target molecule, the program will take several inference steps iteratively applying either analogical or deductive reasoning modes. To connect the source and the target molecules along a pathway, Pathmodel infers missing reactions and metabolites using a minimal number of reactions.

422 # Discussion

423

424 **Multiple alternative pathways for sterol synthesis**

425 Based on the available genomic and metabolomic data, we can propose two alternative
426 pathways from cycloartenol to cholesterol, depending on when the side-chain reductase (SSR)
427 enzyme is acting (Fig 7).



429 **Fig 7. Alternative lathosterol synthesis pathways from cycloartenol in *C. crispus*.** Enzyme names
430 refer to terrestrial plants, escept for ERG2 that refers to yeast, and are explained in Table 3.

431 The « early SSR » pathway is based on the model previously published for tomato [46]. The
432 reactions were manually incorporated in the PathModel knowledge base because they are not
433 yet available in MetaCyc. If *C. crispus* uses this pathway, the metabolic intermediates would
434 be identical to tomato, but there would be an important difference concerning the enzymes.
435 Indeed, the genes encoding SSR are duplicated in Solanaceae (tomato and potato) but not in
436 other plants, in the *C. crispus* genome or in any red algal genome analyzed so far (Supp
437 Figure 3). In Solanacea, SSR2 acts on cycloartenol whereas SSR1 acts late in the phytosterol
438 synthesis pathway, as does the unduplicated SSR from non-solanaceous plants. Moreover,
439 SSR is known to be catalytically promiscuous, and in humans the unique SSR enzyme is able
440 to act either late or early [60]. Therefore, our data suggest that the single SSR is also flexible
441 in *Chondrus* as it is in humans, enabling the existence of multiple synthesis pathways leading
442 to cholesterol. A high level of reticulation with multiple alternative routes in the plant sterol

443 synthesis pathways has already been suggested [59], although only the main pathway has
444 been incorporated in the knowledge base (see for example PWY-2541 in MetaCyc).
445 Consistently, Pathmodel suggested that SSR could act on all possible intermediates. However,
446 flux analyses in mouse have shown that among all theoretical possibilities, two distinct
447 pathways, whose relative abundance vary accross tissue, are sufficient to enable refined and
448 partially distinct regulations [60].

449 Another major difference with land plants concerns the position of the sterol
450 methyltransferases, that are necessary to produce methylated sterols like campesterol or
451 brassicasterol. In the standard model for land plants, a first methylation occurs directly on
452 cycloartenol whereas the second one occurs later on 24-methylenelophenol [59]. Although
453 this possibility cannot be fully ruled out concerning *C. crispus* with present data, various
454 pieces of evidence points toward the necessity to consider alternative pathways. First, we did
455 not find any evidence for the presence of cycloeucalenol or fucosterol, which are common
456 synthetic intermediates in the plant pathway. Second, another methylated sterol, 24-
457 methylenecholesterol, was identified previously in *C. crispus* [29]. In line with this, and
458 building on other reports about methyltransferase catalytic promiscuity accross land plants
459 and green algae, [63, 64], Pathmodel inferred an alternative synthesis pathway for methylated
460 sterols through C24-methylation on desmosterol (Fig 4). This option highly reduces the
461 number of non-identified methylated intermediates, limiting them to 24-methydesmosterol
462 and Δ24-25-sitosterol. It seems also more relevant from a quantitative viewpoint, because this
463 late methylation step would enable the production of methylated sterols using the late SSR
464 pathway, which is also in agreement with the formation of cholesterol as the main sterol.

465

466 **New candidate enzymes for decarboxylation and dehydration lead to a more consistent**
467 **model for MAA synthesis pathway in *C. crispus***

468 The upstream part of the MAA synthesis pathway in *C. crispus*, down to shinorine and
469 porphyra-334, follows the current consensus. For this part, candidate enzymes were already
470 proposed [54], and this knowledge is already partially incorporated in the MetaCyc database
471 (PWY-7751 on Fig 4). Here we added further experimental support to the presence of this
472 part of the pathway, performing the first identification of mycosporine-glycine in *C. crispus*
473 (Fig 4 and Fig 8). We also encoded in the Pathmodel knowledge base an extended version of

the aminoacid C3-transfer reaction (RXN-17371 in MetaCyc) to incorporate the already formulated hypothesis, based on structural comparisons between molecules, that MysD can also perform the aminoacid C-3 transfer of threonine, thus leading to porphyra-334 (Fig 8, reaction in red, redrawn from [54]). For the more downstream part of the pathway, some reactions have been proposed, such as decarboxylation of shinorine to asterina-330, but without association with a specific enzyme family [65]. Encoding this literature-based information and constraining the Pathmodel output to find a pathway leading to a molecule structure compatible with the mesured m/z ratio for MAA1, it was possible to infer the hypothetic structure shown on Fig 8. The reaction leading from asterina-330 to MAA1 would be a dehydration (in purple), the same kind of reaction that is already observed between other MAAs such as porphyra-334 and Z-palythenic acid, a compound not identified in *C. crispus* (Fig 4).

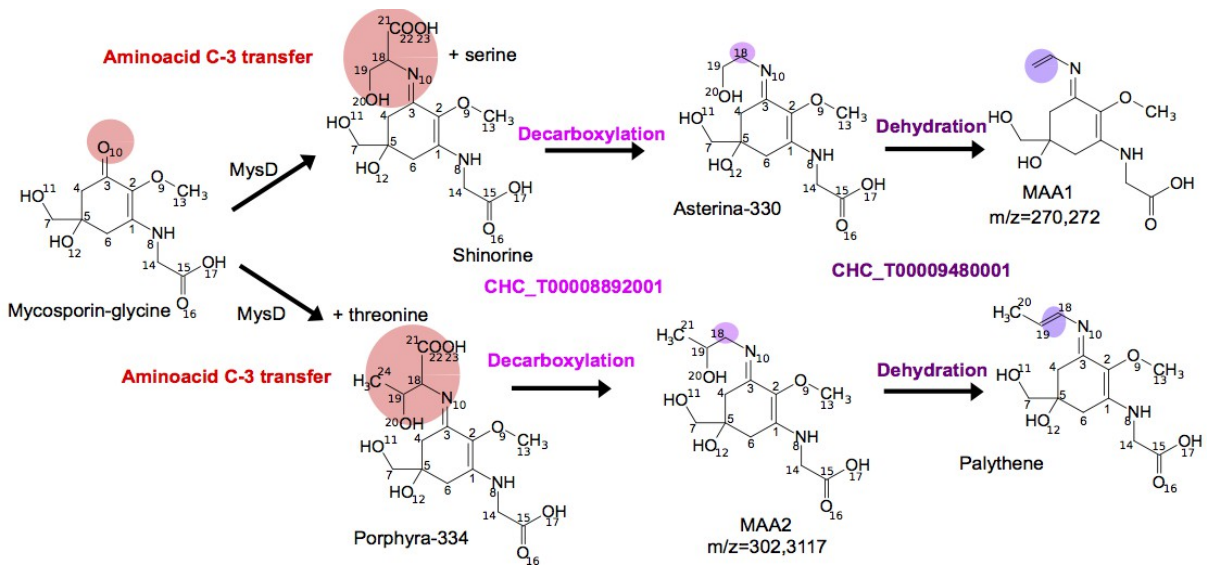


Fig 8. New candidate reactions, enzymes and metabolites in the downstream part of the MAA biosynthesis pathway in *C. crispus*. Structure of MAAs and their precursors are drawn with carbon and oxygen atom labelling corresponding to the numeration used in Pathmodel. The two newly inferred reactions are serine/threonine decarboxylation, in pink, and serine/threonine dehydration, in purple.

For both decarboxylation and dehydration reactions, no candidate enzymes were mentioned so far in the literature related to MAA biosynthesis pathways. We thus performed a simple

505 semantic search on a draft version of the GSM from *C. crispus*, to identify other enzymes that
506 may perform those reactions on a serine coupled with other chemical building blocks. Serine
507 decarboxylation indeed occurs in phospholipid metabolism and was inferred in the *Chondrus*
508 GSM based on orthology with *Galdieria sulphuraria*. The candidate gene is
509 CHC_T00008892001. Interestingly, there is some evidence of catalytic promiscuity for this
510 enzyme, enabling it to also decarboxylate a threonine residue. So far, biochemical data in
511 mammalian cell cultures indicate that phosphatidylthreonine decarboxylation by
512 phosphatidylserine occurs, but with a weak activity [66]. The *in vivo* occurrence and
513 biosynthetic origin of phosphatidylthreonine were only recently demonstrated using HPLC-
514 MS/MS in the apicomplexan parasite *Toxoplasma gondii* where it is produced by a
515 phosphatidylthreonine synthase coming from an ancient gene duplication of a
516 phosphatidylserine synthase specifically in the lineage encompassing stramenopiles,
517 alveolates and rhizarians [67]. Therefore, we hypothesize that the enzyme may be
518 promiscuous in *C. crispus* and may also perform serine/threonine decarboxylation on a
519 serine/threonine linked to a mycosporin-glycin instead or in addition to performing this on a
520 phospholipid.

521 Following the same rationale, we performed a semantic search for a serine/threonine
522 dehydratase, and found the enzyme encoded by CHC_T00009480001. This enzyme was
523 predicted based on Pathway Tools to be involved either in degradation of glycine betaine,
524 purine nucleobases, or L-serine, and is a member of the Pyridoxal-phosphate dependent
525 enzyme family, that contains both the human serine dehydratase (EC:4.3.1.17; P20132) and
526 the *E. coli* threonine dehydratase (EC:4.3.1.19; P04968) signatures. Common ancestry for
527 serine and threonine dehydratases has been proposed long ago [68]. However, it should be
528 noted that enzymatic promiscuity is very high among pyridoxal-phosphate dependent
529 enzymes [69] and that hydrolases represent only a minor fraction of their overall described
530 biochemical activities [70].

531 Inferring a single pair of enzymes to decarboxylate shinorine and porphyra-334 and further
532 dehydrate their derivatives was also parsimonious in respect to the absence of a peak
533 corresponding to Z-palythenic acid in *C. crispus* extracts, which does not support dehydration
534 occuring before decarboxylation, as proposed in other species (Fig 4 and [65]). The structure
535 of a new intermediate was therefore inferred manually, leading to MAA2 on Figure 8.
536 Calculating its m/z ratio, we found this was identical to palythinol, a compound previously

537 considered to be present in *C. crispus* based on UV+LC-MS or LC-MS/MS data [49, 51]. We
538 then verified using Pathmodel that constraining the pathway search with a molecule having a
539 m/z ratio of 302,3177 leads to the same actual MAA2 as a proposed unique solution. Because
540 there is no synthesis-based analytical standard available for palythinol, as for all other MAAs,
541 it was useful to make this alternative hypothesis explicit. From a genomic viewpoint,
542 switching palythinol with MAA2 does not necessitate a candidate enzyme to perform
543 hydrogenation and demethylation on a MAA-like substrate (Fig. 4), and thus reduces the
544 number of unassigned enzymatic activities to candidate genes.

545

546 **Implications of possible sterol and MAA synthesis pathways in *C. crispus* on**
547 **evolutionary scenarios regarding metabolic pathway drift**

548 Our study demonstrates that data on metabolite occurrence can be explicitly incorporated into
549 the quality criteria for evaluating a GSM. Putting more emphasis on metabolites, especially
550 the missing ones, creates new methodological challenges regarding *ab initio* inferences of
551 pathways when enzymes are not yet known, and we have shown that it is now possible to
552 build new tools to specifically address those challenges. The next issue is about the scalability
553 of our approach. The Pathmodel version we present here is a working prototype that can
554 already be applied to other metabolic pathways in *C. crispus* or in other organisms where
555 genomic and metabolomic data are available. Further improvements should be done in order
556 to minimize the user's burden in manually entering molecular structures. It is not yet possible
557 to fully automate the atom numbering during metabolic reaction. Note that the five main
558 existing solutions have a success rate of 91% compared with manual mapping, which means
559 that errors would remain with such an approach [71]. We have thus proposed a graphical
560 output in order to facilitate the check of encoded molecule structures (S5-S6 Figs).

561 The Pathmodel tool was developed to support reasoning based on the metabolic pathway drift
562 hypothesis in order to automatically infer new reactions and metabolites. A first key feature of
563 the successful application of this strategy was the precision and the quality of the biochemical
564 and biological knowledge encoded in Pathmodel. Generalizing this approach to any other
565 application will similarly require interactions between chemists, biologists and computer
566 scientists. The second key feature of Pathmodel is to be focused on a selected pathway rather
567 than on a complete genome-scale metabolic network. The selection of the relevant pathway to

568 be considered - for instance from preliminary evidences extracted from metabolomics analysis
569 - is therefore a key pre-processing step to combine and filter the predictions of Pathmodel
570 with genomics and metabolomics data.

571 Whatever the actual topology of the sterol and MAA pathways in *C. crispus,* each discussed
572 hypotheses have implications regarding metabolic pathway drift. All possible sterol pathways
573 provide further strong candidates case studies for a drift by non-homologous enzyme
574 replacement, and the new pathways inferred by Pathmodel provide candidates case studies for
575 of drift by enzyme inversion. The unresolved point with the sterol pathways is that, among
576 eukaryotes, there is no consensus yet about the ancestral order of enzymatic reactions.
577 Experimental data are too disparate across the tree of life to enable firm conclusions on this.
578 In that respect, the MAA pathway is interesting, because if our hypothesis about
579 decarboxylation of porphyra-334 before dehydration is true, this would mean that an
580 enzymatic inversion took place in other lineages where porphyra-334 is first dehydrated to Z-
581 palytenic acid and then decarboxylated to palythene. Here the limit is that, to date, enzymes
582 are unknown for both reactions, so the system is not yet genomically tractable. Identifying
583 close enzymatic inversions is important, because experimental analyses on *E. coli* have shown
584 that drastic pathway rewiring by enzyme knockout or gene overexpression can led to toxic
585 intermediates [72]. Enzyme inversion would provide a milder mechanism for gradual
586 divergence of pathways. But to identify such cases we need genomic and metabolomic data
587 for more closely related model species. Such data will become available in the coming years
588 thanks to ongoing integrative sequencing and metabolomic projects.

589

590 # Material and Methods

591 **Sampling of algae**

592 For sterol analyses, samples from *Chondrus crispus* were collected from a population on the
593 shore at Roscoff, France, in front of the Station Biologique (48°43'38'' N ; 3°59'04'' W).
594 Algal cultures were maintained in 10 L flasks in a culture room at 14°C using filtered
595 seawater and aerated with filtered (0.22 µm) compressed air to avoid CO_2 depletion.
596 Photosynthetically active radiation (PAR) was provided by Philips daylight fluorescence
597 tubes at a photon flux density of 40 $\mu mol.m^{-2}.s^{-1}$ for 10 $h.d^{-1}$. The algal samples were freeze
598 dried, ground to powder using a cryogrinder and stored at -80°C.

599 For MAAs analysis, more than 50 g (wet weight) of *Chondrus crispus* were collected along
600 the Brittany coasts (France) at Ploemeur (47°42'07'' N; 3°24'31'' W) in July 2013, Roscoff
601 (48°43'38'' N; 3°59'04'' W) in April and August 2013, and Tregunc (47°50'25''N; 3°54'08''
602 W) in September 2013.

603

604 **Standards and reagents**

605 Cholesterol, stigmasterol, *β*-sitosterol, 7-dehydrocholesterol, lathosterol (5*α*-cholest-7-en-3*β*-
606 ol), squalene, campesterol, brassicasterol, desmosterol, lanosterol, fucosterol, cycloartenol,
607 5*α*-cholestane (internal standard) were acquired from Sigma-Aldrich (Saint-Quentin-Fallavier,
608 France), cycloartanol and cycloeucalenol from Chemfaces (Wuhan, China) and zymosterol
609 from Avanti Polar Lipids (Alabaster, USA). The C7-C40 Saturated Alkanes Standards were
610 acquired from Supelco (Bellefonte, USA). Reagents used for extraction, saponification, and
611 derivation steps were *n*-hexane, ethyl acetate, acetonitrile, methanol (Carlo ERBA Reagents,
612 Val de Reuil, France), (trimethylsilyl)diazomethane, toluene (Sigma-Aldrich, Saint-Quentin-
613 Fallavier, France) and N,O-bis(trimethylsilyl)trifluoroacetamide with trimethylcholorosilane
614 (BSTFA:TMCS (99:1)) (Supelco, Bellefonte, USA).

615

616 **Standard preparation**

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

624

625 **Sample preparation**

626 Dried algal samples (60 mg) were extracted with 2mL ethyl acetate by continuous agitation
627 for 1 hour at 4°C. After 10 min of centrifugation at 4 000 rpm, the solvent was removed, the
628 extracts were saponified in 3 mL of methanolic potassium hydroxide solution (1M) by 1 hour

629 incubation at 90°C. The saponification reaction was stopped by plunging samples into an ice
630 bath for 30 min minimum. The unsaponifiable fraction was extracted with 2 mL of hexane
631 and 1.2 mL of water and centrifuged at 2000 rpm for 5 min. The upper phase was collected,
632 dried under N_2, and resuspended with 120 µL of (trimethylsilyl)diazomethane, 50 µL of
633 methanol:toluene (2:1 (v/v)) and 5 µL of 5α-cholestane (1 mg.mL^{-1}) as internal standard. The
634 mixture was vortexed for 30 seconds, and heated at 37°C for 30 min. After a second
635 evaporation under N_2, 50 µL of acetonitrile and 50 µL of BSTFA:TMCS (99:1) were added to
636 the dry residue, vortexed for 30 seconds and heated at 60°C for 30 min. After final
637 evaporation under N_2, the extract was resuspended in 100 µL of hexane, transferred into a
638 sample vial and stored at -80°C until the GC-MS analysis.
639 For MAAs, one gram of dried algae was extracted twice for two hours under continuous
640 shaking with 10 mL of acetone. After 5 min of centrifugation at 3 000 rpm, acetone was
641 discarded and samples were re-extracted twice with 10 mL water/acetone (30/70, v/v) for 24
642 hours under continuous shaking at 120 rpm. Water/acetone supernatants were pooled, added
643 to one gram of silica and evaporated to dryness by rotary evaporation. Extracts were then
644 purified by silica gel chromatography column with dichloromethane/methanol mixtures and
645 MAAs were eluted with 200 mL of dichloromethane/methanol (15/85, v/v). After rotary
646 evaporation, samples were re-suspended in water/methanol (50/50, v/v) and filtrated using
647 0.45 µm syringes filter. Solution were adjusted to a final concentration of 1 mg.mL^{-1} and
648 stored at 3°C until LC-MS analysis.

649

650 **Sterol analysis by gas chromatography-mass spectrometry**
651 The sterols were analyzed on a 7890 Agilent Technologies gas chromatography coupled with
652 a 5975C Agilent Technologies mass spectrometer (GC-MS). A HP-5MS capillary GC column
653 (30 m x 0.25 mm x 0.25 µm) from J&W Scientific (CA, USA) was used for separation and
654 UHP helium was used as carrier gas at flow rate to 1 mL.min^{-1}. The temperature of the
655 injector was 280°C and the detector temperature was 315°C. After injection, the oven
656 temperature was kept at 60°C for 1 min. The temperature was increased from 60°C to 100°C
657 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
658 3°C.min^{-1} and then held at 315°C for 2 min, resulting in a total run time of 37 min.

659 Electronic impact mass spectra were measured at 70eV and an ionization temperature of
660 250°C. The mass spectra scanned from m/z 50 to m/z 500. Peaks were identified based on the
661 comparisons with the retention times and the mass spectra (S1 Table).
662

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

664 High Resolution Mass Spectrometry was carried out on a microTOF-Q II (Bruker Daltonics,
665 Germany) coupled to an Ultimate 3000 LC System (Dionex, Germany). Experiments were
666 performed on a Gemini C6-Phenyl column (250 mm x 4.6 mm x 5 µm) (Phenomenex,
667 Germany). The gradient was as follows: methanol/water (20:80, v/v) with 0.2% acid acetic for
668 two minutes to 100 % methanol with 0.2% acid acetic in 23 minutes. The UV detector was set
669 to 330 nm, flow rate was kept constant at 0.4 $mL.min^{-1}$ and column temperature set at 30°C.
670 MS spectra were recorded in positive ESI mode with a drying gas temperature of 220°C, a
671 nitrogen flow of 12 $L.min^{-1}$, a nebulizer pressure set to 60 psi, and a collision energy of 20 eV.
672 MAAs were identified by HR-MS on the basis of the detection of the pseudo-molecular ion
673 $[M+H]^+$ with a *m/z v*alue varying less than ± 0.02 Da compared to the theoretical *m/z* value. In
674 the absence of commercially available standards, relative quantification of MAAs in each
675 sample was estimated by calculating the ratio between the area under the curve of the
676 Extracted Ion Chromatogram (EIC) corresponding to the selected MAAs and the sum of the
677 areas under the curve of the EIC of all MAAs detected in the algal extract. The same
678 procedure was applied to UV detection (S2 Table).
679
680

681 **Genome-scale metabolic network reconstruction**

682 Genome-scale metabolic network reconstruction was performed using the AuReMe pipeline
683 [55]. A set of 89 targets coming from the literature was used as an input and is provided in S3
684 Table. Orphan metabolites that are experimentally supported but do not have a MetaCyc ID
685 are listed in S4 Table.

686 The process encompassed the following steps:

687 1) an annotation-based draft network was generated using the PathoLogic program from the
688 Pathway Tools suite, using the gbk file from the *C. crispus* genome annotation [25] and the
689 metabolic reaction database MetaCyc20.5 [73].

690 2) an orthology-based network was generated using the protein sequences and metabolic
691 network of *A. thaliana* (AraGEM, [57]), using the Pantograph software [74] to combine the
692 output of ortholog searches with the Inparanoid and OrthoMCL softwares.

693 3) an orthology-based network was generated using the protein sequences from the well-
694 annotated red microalga *Galdieria sulphuraria* [75] and its metabolic network reconstructed
695 using Pathway Tools. This *G. sulphuraria* annotation-based network was then used as a
696 template to generate a *C. crispus* network using Pantograph.

697 4) an orthology-based network was also generated using the protein sequences from the
698 version 2 of the annotated genome of *E. siliculosus* [76], as well as version 2 of its metabolic
699 network [55].

700 5) the four preliminary networks were merged together in the AuReMe environment, and an
701 additional gap-filling step was performed using Meneco [77], constraining the network to
702 produce the 84 metabolites from the literature that were indexed in the Metacyc database.

703

704 **Flux-balance analysis**

705 A biomass reaction was established based on the previous *E. siliculosus* data [78]. One
706 compound, L-alpha-alanine, gave negative fluxes, thus blocking biomass production. This
707 was due to the absence of the alanine dehydrogenase reaction. The corresponding enzyme
708 (CHC_T00008930001) was present in the *C. crispus* network but annotated as an NAD(P)
709 transhydrogenase. We completed the annotation through the manual curation form to enable it
710 to dehydrogenate alanine and to restore producibility of the biomass (http://gem-
711 aureme.irisa.fr/ccrgem/index.php/Manual-ala_dehy).

712

713 **Global metabolic networks comparisons**

714 In order to compare the global features of the GSM from *C. crispus* with other ones, it is
715 necessary to use the same reference database. This is the case for *E. siliculosus* and *E.*
716 *subulatus* for which the reconstructions are based on MetaCyc [73] while *A. thaliana* and *C.*
717 *reinhardtii* are respectively from KEGG [79] and BiGG [80]. To get access to MetaCyc
718 pathway information for *A. thaliana* and *C. reinhardtii,* their networks were mapped using the

719 sbml_mapping function implemented in the AuReMe workflow [55]. This function provides a
720 dictionary of corresponding reactions from a database to an other one using the MetaNetX
721 cross-reference database [81]. This dictionary was then used in AuReMe to create a new
722 genome-scale metabolic network based on the new reference database for *A. thaliana* and *C.*
723 *reinhardti*. Those new networks, who are comparable in size with the published ones (+/- 10
724 reactions and enzymes in our counts) enabled to estimate the number of pathways as defined
725 in MetaCyc for both species.
726

727 **Ab-initio inference of new metabolic reactions**

728 To enable the incorporation of the orphan metabolites that were not yet in MetaCyc into the
729 network, we developed a new method called "Pathmodel" that can infer new reactions based
730 on molecular similarity and dissimilarity. This knowledge-based approach is founded on two
731 modes of reasoning (deductive and analogical) and was implemented using a logic
732 programming approach known as Answer Set Programming (ASP) [82, 83]. It is a declarative
733 approach oriented toward combinatorial (optimization) problem-solving and knowledge
734 processing. ASP combines both a high-level modeling language with high performance
735 solving engines so that the focus is on the problem specification rather than the algorithmic
736 part. ASP expresses a problem as a set of logical rules (clauses). Problem solutions appear as
737 particular logical models (so-called stable models or answer sets) of this set. An ASP program
738 consists of rules h :- $b_1, \ldots, b_m$ *not* $b_{m+1}, \ldots,$ *not* b_n, where each b_i and h are literals and *not*
739 stands for default negation. In fact, each proposition is a predicate, encoded by a function
740 whose arguments can be constant atoms or variables over a finite domain. The rule states that
741 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,
742 … , b_m are true and it cannot be proved that b_{m+1}, … , b_n are true.
743
744 In short, the main predicates used in Pathmodel to represent molecules and reactions forming
745 a knowledge base are *bond*, *atom* and *reaction* on which several logical rules are then applied
746 to all possible reactions and potential reactants. Resulting products that do not belong to the
747 knowledge base but that correspond to an observed m/z ratio are considered as new inferred
748 metabolites and reactions. The finally encoded reactions result from iterative interactions
749 between analogical model construction, automated inference, and manual validation of

750 inferred reactions with respect to experimental results. The principles of the encoded
751 analogical reasoning are explained in the results and discussion part.

752 The source code is available in the following Gitlab repository:
753 https://gitlab.inria.fr/DYLISS/PathModel. The added reactions are listed on the following
754 pages: http://gem-aureme.irisa.fr/ccrgem/index.php/Manual-pathmodel_inference

755 http://gem-aureme.irisa.fr/ccrgem/index.php/Manual-pathmodel_inference_new_rxn

756

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

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

764

765 **Phylogenetic analyses**

766 Collected sequences were aligned using Clustal Omega [84] and alignments were checked
767 manually and edited with Seaview [85]. Phylogenetic trees were built using PHYML [86]
768 using the LG model [87] with a gamma law. The reliability of nodes was assessed by
769 likelihood-ratio test [88].

770

771 **Acknowledgments**

772 We thank Cécile Hervé for help in collecting field samples from *C. crispus,* Gaëlle Correc for
773 help in sample preparation, Karine Cahier for help during GC-MS analyses on the
774 MetaboMer-Corsaire plateform, Jeanne Got and Marie Chevallier for help in using
775 preliminary versions of AuReMe, and Clémence Frioux for help in analyzing FBA artefacts.
776 GVM is also grateful to Ralf J. Sommer for giving him the possibility to start developping

Pathmodel during a previous postdoctoral stay at the Max-Planck Institute for Developmental Biology in Tübingen.

Author contributions

Conceptualization : GVM, LD, SMD, PS, EC, JN, CB, CL, AS, JC

Data curation : GVM, JG, MA, AB, CL, LD, SMD, PS, EC, CB, CL, AS, JC

Funding acquisition : AS, SMD, CB, CL

Investigation : GVM, JG, MA, AB, CT, CM

Project administration : GVM

Software & Methodology : AB, GVM, JN

Writing – original draft : AB, JG, MA, PS, JN, AS, GVM

Writing – review & editing : AB, JG, LD, SMD, PS, CT, EC, JN, CL, CB, JC, AS, GVM

Conflict of interest

The authors have no conflict of interest to declare.

References

1. Thiele I, Palsson B. A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc. 2010; 5: 93 – 121.
2. Edison AS, Hall RD, Junot C, Karp PD, Kurland IJ, Mistrik R, et al. The time is right to focus on model organism metabolomes. Metabolites. 2016; 6.
3. Viant MR, Kurland IJ, Jones MR, Dunn WB. How close are we to complete annotation of metabolomes? Curr Opin Chem Biol. 2017; 36: 64 – 69.
4. Brodie J, Chan CX, De Clerck O, Cock JM, Coelho SM, Gachon C, et al. The algal revolution. Trends Plant Sci. 2017; 22: 726 – 738.

802 5. Ebrahim A, Almaas E, Bauer E, Bordbar A, Burgard AP, et al. Do genome-scale models need
803 exact solvers or clearer standards? Molecular Systems Biology. 2015; 11: 831.

804 6. Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, et al. Global reconstruction of
805 the human metabolic network based on genomic and bibliomic data. Proc Natl Acad Sci.
806 2007; 104: 1777 – 1782.

807 7. Belghit I, Rasinger JD, Heesch S, Biancarosa I, Liland N, Torstensen B, et al. In-depth
808 metabolic profiling of marine macroalgae confirms strong biochemical differences between
809 brown, red and green algae. Algal Res. 2017; 26: 240 – 249.

810 8. Salek RM, Conesa P, Cochrane K, Haug K, Williams M, Kale N, et al. Automated assembly
811 of species metabolomes through data submission into a public repository. Gigascience. 2017;
812 6: 1 – 4.

813 9. Rhee KY, de Carvalho LPS, Bryk R, Ehrt S, Marrero J, Park SW, et al. Central carbon
814 metabolism in Mycobacterium tuberculosis: an unexpected frontier. Trends Microbiol. 2011;
815 19: 307 – 314.

816 10. Noda-Garcia L, Liebermeister W, Tawfik DS. Metabolite-enzyme coevolution: from single
817 enzymes to metabolic pathways and networks. Annu Rev Biochem. 2018; 87: 187 – 216.

818 11. True JR, Haag ES. Developmental system drift and flexibility in evolutionary trajectories.
819 Evol Dev. 2001; 3: 109 – 119.

820 12. Townsley BT, Sinha NR. A new development: evolving concepts in leaf ontogeny. Annu Rev
821 Plant Biol. 2012; 63: 535 – 562.

822 13. Hart KM, Harms MJ, Schmidt BH, Elya C, Thornton JW, Marqusee S. Thermodynamic
823 system drift in protein evolution. PLOS Biol. 2014; 12: e1001994.

824 14. Petit C, Rey C, Lambert A, Peltier M, Pantalacci S, Sémon M. Comparing transcriptomes to
825 probe into the evolution of developmental program reveals an extensive developmental system
826 drift. Proceedings of the JOBIM conference. 2016; 118 – 120.

827 15. Koonin EV, Mushegian AR, Bork P. Non-orthologous gene displacement.
828 Trends Genet. 1996; 12: 334 – 336.

829 16. Markov GV, Meyer JM, Panda O, Artyukhin AB, Claaßen M, Witte H, et al. Functional
830 conservation and divergence of *daf-22* paralogs in *Pristionchus pacificus* dauer development.
831 Mol Biol Evol. 2016; 33: 2506 – 2514.

832 17. Carlsson L, Spjuth O, Adams S, Glen RC, Boyer S. Use of historic metabolic
833 biotransformation data as a means of anticipating metabolic sites using MetaPrint2D and
834 Bioclipse. BMC Bioinformatics. 2010; 11: 362.

835 18. King RD, Whelan KE, Jones FM, Reiser PGK, Bryant CH, Muggleton SH, et al. Functional
836 genomic hypothesis generation and experimentation by a robot scientist. Nature. 2004; 427:
837 247 – 252.

838 19. Koch M, Duigou T, Carbonell P, Faulon JL. Molecular structures enumeration and virtual
839 screening in the chemical space with RetroPath2.0. J Cheminformatics. 2017; 9: 64.

840 20. Lindsay RK, Buchanan BG, Feigenbaum EA, Lederberg J. DENDRAL: a case study of the
841 first expert system for scientific hypothesis formation. Artif Intell. 1993; 61: 209 – 261.

842 21. Kell DB, Oliver SG. Here is the evidence, now what is the hypothesis? The complementary
843 roles of inductive and hypothesis-driven science in the post-genomic era. BioEssays. 2004; 26:
844 99 – 105.

845 22. Sparkes A, Aubrey W, Byrne E, Clare A, Khan MN, Liakata M, et al. Towards robot scientists
846 for autonomous scientific discovery. Automated Experimentation. 2010; 2: 1 – 11.

847 23. Williams K, Bilsland E, Sparkes A, Aubrey W, Young M, Soldatova, et al. Cheaper faster
848 drug development validated by the repositioning of drugs against neglected tropical diseases. J
849 R Soc Interface. 2015; 12: 20141289.

850 24. Collén J, Cornish ML, Craigie J, Ficko-Blean E, Hervé C, Krueger-Hadfield SA, et al.
851 *Chondrus crispus* – A present and historical model organism for red seaweeds. Adv Bot Res.
852 2014; 71: 53 – 90.

853 25. Collén J, Porcel B, Carré W, Ball SG, Chaparro C, Tonon T, et al. Genome structure and
854 metabolic features in the red seaweed *Chondrus crispus* shed light on evolution of the
855 Archaeplastida. Proc Natl Acad Sci USA. 2013; 110: 5247 – 5252.

856 26. Young EG, Smith DG. Amino acids, peptides, and proteins of Irish moss, *Chondrus crispus*. J
857 Biol Chem. 1958; 233: 406 – 410.

858 27. Laycock MV, Craigie JS. The occurrence and seasonal variation of gigartinine and L-
859 citrullinyl-L-arginine in *Chondrus crispus* Stackh. Can J Biochem. 1977; 55: 27 – 30.

860 28. Matsuhiro B, Urzua C. Heterogeneity of carrageenans from *Chondrus crispus*
861 Phytochemistry. 1992; 31: 531 – 534.

862 29. Tasende M. Fatty acid and sterol composition of gametophytes and sporophytes of *Chondrus*
863 *crispus* (Gigartinaceae, Rhodophyta). Sci Mar. 2000; 64: 421 – 426.

864 30. Kräbs G, Watanabe M, Wiencke C. A monochromatic action spectrum for the photoinduction
865 of the UV-absorbing mycosporine-like amino acid shinorine in the red alga *Chondrus crispus*.
866 Photochem Photobiol. 2004; 79: 515 – 519.

867 31. Gaquerel E, Hervé C, Labrière C, Boyen C, Potin P, Salaün JP. Evidence for oxylipin
868 synthesis and induction of a new polyunsaturated fatty acid hydroxylase activity in *Chondrus*
869 *crispus* in response to methyljasmonate. Biochim Biophys Acta, Mol Cell Biol Lipids. 2007;
870 1771: 565 – 575.

871 32. Banskota AH, Stefanova R, Sperker S, Lall S, Craigie JS, Hafting JT. Lipids isolated from the
872 cultivated red alga *Chondrus crispus* inhibit nitric oxide production. J Appl Phycol. 2014; 26:
873 1565 – 1571.

874 33. Pina A, Costa A, Lage-Yusty M, López-Hernández J. An evaluation of edible red seaweed
875 (*Chondrus crispus*) components and their modification during the cooking process. LWT -
876 Food Sci Technol. 2014; 56: 175 – 180.

877 34. Melo T, Alves E, Azevedo V, Martins AS, Neves B, Domingues P, et al. Lipidomics as a new
878 approach for the bioprospecting of marine macroalgae – Unraveling the polar lipid and fatty
879 acid composition of *Chondrus crispus*. 2015; Algal Res 8: 181 – 191.

880 35. Alcaide A, Devys M, Barbier M. Remarques sur les stérols des algues rouges. Phytochemistry.
881 1968; 7: 329 – 330.

882 36. Kremer BP, Kirst GO. Biosynthesis of photosynthates and taxonomy of algae
883 Z Naturforsch. 1982; 37c: 761 – 771.

884 37. Pettit T, Jones A, Harwood J. Lipid metabolism in the red marine algae *Chondrus crispus* and
885 *Polysiphonia lanosa* as modified by temperature. Phytochemistry. 1989; 28: 2053 – 2089.

886 38. van Ginneken VJ, Helsper JP, de Visser W, van Keulen H, Brandenburg WA. Polyunsaturated
887 fatty acids in various macroalgal species from north Atlantic and tropical seas. Lipids Health
888 Dis. 2011. 10: 104.

889 39. Santos SA, Vilela C, Freire CS, Abreu MH, Rocha SM, Silvestre AJ. Chlorophyta and
890 Rhodophyta macroalgae: A source of health promoting phytochemicals. Food Chem. 2015;
891 183: 122 – 128.

892 40. Robertson RC, Guihéneuf F, Bahar B, Schmid M, Stengel DB, Fitzgerald GF, et al. The anti-
893 inflammatory effect of algae-derived lipid extracts on lipopolysaccharide (LPS)-stimulated
894 human THP-1 macrophages. Mar Drugs. 2015; 13: 5402 – 5424.

895 41. Desmond E, Gribaldo, S. Phylogenomics of sterol synthesis: insights into the origin,
896 evolution, and diversity of a key eukaryotic feature. Genome Biology and Evolution. 2009; 1:
897 364 – 381.

898 42. Moss GP. IUPAC-IUB Joint Commission on Biochemical Nomenclature. Nomenclature of
899 steroids (Recommendations 1989). Eur J Biochem. 1989; 61: 1783 – 1822.

900 43. Sumner LW, Amberg A, Barrett D, Beale MH, Beger R, Daykin CA, et al. Proposed
901 minimum reporting standards for chemical analysis Chemical Analysis Working Group
902 (CAWG) Metabolomics Standards Initiative (MSI). Metabolomics. 2007; 3: 211 – 221.

903 44. Saito A, Idler DR. Sterols in Irish Moss (*Chondrus crispus*). Can J Biochem. 1966; 44: 1195 –
904 1199.

905 45. Goldberg A, Hubby C, Cobb D, Millard P, Ferrara N, Galdi G, et al. Sterol distribution in red
906 algae from the waters of eastern Long Island. Botanica Marina. 1982; 25: 351 – 355.

907 46. Sonawane PD, Pollier J, Panda S, Szymanski J, Massalha H, Yona M, et al. Plant cholesterol
908 biosynthetic pathway overlaps with phytosterol metabolism Nature Plants. 2016; 3, 16205.

909 47. Alcaide A, Barbier M, Potier P, Magueur AM, Teste J. Nouveaux résultats sur les stérols des
910 algues rouges. Phytochemistry. 1969; 8: 2301 – 2303.

911 48. Horgen FD, Sakamoto B, Scheuer PJ. New triterpenoid sulfates from the red alga
912 *Tricleocarpa fragilis*. J Nat Prod. 2000; 63: 210 – 216.

913 49. Athukorala Y, Trang S, Kwok C, Yuan YV. Antiproliferative and antioxidant activities and
914 mycosporine-like amino acid profiles of wild-harvested and cultivated edible Canadian marine
915 red macroalgae. Molecules. 2016; 21: E119.

916 50. Guihéneuf F, Gietl A, Stengel DB. Temporal and spatial variability of mycosporine-like
917 amino acids and pigments in three edible red seaweeds from western Ireland. J Appl Phycol.
918 2018; 30: 2573 – 2586.

919 51. Karsten U, Franklin LA, Lüning K, Wiencke C. Natural ultraviolet radiation and
920 photosynthetically active radiation induce formation of mycosporine-like amino acids in the
921 marine macroalga *Chondrus crispus* (Rhodophyta). Planta, 1998; 205: 257 – 262.

922 52. Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TCG. More than just orphans: are
923 taxonomically-restricted genes important in evolution? Trends Genet. 2009; 25: 404 – 413.

924 53. Fabris M, Matthijs, M, Carbonelle, S, Moses T, Pollier J, Dasseville R, et al. Tracking the
925 sterol biosynthesis pathway of the diatom *Phaeodactylum tricornutum*. New Phytol. 2014;
926 204: 521 – 535.

927 54. Brawley SH, Blouin NA, Ficko-Blean E, Wheeler GL, Lohr M, Goodson HV, et al. Insights
928 into the red algae and eukaryotic evolution from the genome of *Porphyra umbilicalis*
929 (Bangiophyceae, Rhodophyta). Proc Natl Acad Sci USA. 2017; 114: E6361 – E6370.

930 55. Aite M, Chevallier M, Frioux C, Trottier C, Got J, et al. Traceability, reproducibility and wiki-
931 exploration for "à-la-carte" reconstructions of genome-scale metabolic models, Plos Comput
932 Biol. 2018; 14:e1006146.

933 56. Dittami SM, Corre E, Brillet-Guéguen L, Pontoizeau N, Lipinska AP, Aite M, et al. The
934 genome of *Ectocarpus subulatus* highlights unique mechanisms for stress tolerance in brown
935 algae. 2018. bioRxiv doi.org/10.1101/307165 [PREPRINT]

936 57. de Oliveira Dal'Molin CG, Quek LE, Palfreyman RW, Brumbley SM, Nielsen LK. AraGEM,
937 a genome-scale reconstruction of the primary metabolic network in *Arabidopsis*. Plant
938 Physiol. 2010; 152: 579 – 589.

939 58. Imam S, Schäuble S, Valenzuela J, López García de Lomana A, Carter W, Price ND, et al.
940 Refined genome-scale reconstruction of *Chlamydomonas* metabolism provides a platform for
941 systems-level analyses. Plant J. 2015; 84: 1239 – 1256.

942 59. Benveniste P. Biosynthesis and accumulation of sterols. Annu Rev Plant Biol. 2004; 55: 429 –
943 457.

944 60. Mitsche MA, McDonald JG, Hobbs HH, Cohen JC. Flux analysis of cholesterol biosynthesis
945 in vivo reveals multiple tissue and cell-type specific pathways. eLife. 2015; 4: e07999.

946 61. Shick JM, Dunlap WC. Mycosporine-like amino acids and related gadusols: biosynthesis,
947 accumulation, and UV-protective functions in aquatic organisms. Annu Rev Physiol. 2002;
948 64: 223 – 262.

949 62. Meija J, Coplen TB, Berglund M, Brand WA, Bièvre PD, Gröning M, et al. Atomic weights of
950 the elements 2013 (IUPAC Technical Report). Pure Appl Chem. 2016; 88: 265 – 291.

951 63. Neelakandan AK, Song Z, Wang J, Richards MH, Wu X, Valliyodan B, et al. Cloning,
952 functional expression and phylogenetic analysis of plant sterol 24C-methyltransferases
953 involved in sitosterol biosynthesis. Phytochemistry. 2009; 70: 1982 – 98.

954 64. Haubrich BA, Collins EK, Howard AL, Wang Q, Snell WJ, Miller MB, et al.
955 Characterization, mutagenesis and mechanistic analysis of an ancient algal sterol C24-
956 methyltransferase: Implications for understanding sterol evolution in the green lineage.
957 Phytochemistry. 2015; 113: 64 – 72.

958 65. Carreto JI, Carignan MO. Mycosporine-like amino acids: relevant secondary metabolites.
959 Chemical and ecological aspects. Mar Drugs. 2011; 9: 387 – 446.

960 66. Heikinheimo L, Somerharju P. Translocation of phosphatidylthreonine and -serine to
961 mitochondria diminishes exponentially with increasing molecular hydrophobicity.
962 Traffic. 2002; 3: 367 – 377.

963 67. Arroyo-Olarte RD, Gupta N. Phosphatidylthreonine: An exclusive phospholipid regulating
964 calcium homeostasis and virulence in a parasitic protest. Microb Cell. 2016; 3: 189 – 190.

965 68. Parsot C. Evolution of biosynthetic pathways: a common ancestor for threonine synthase,
966 threonine dehydratase and D-serine dehydratase. EMBO J. 1986; 5: 3013 – 3019.

967 69. Percudani R, Peracchi A. A genomic overview of pyridoxal-phosphate-dependent enzymes.
968 EMBO Rep. 2003; 4: 850 – 854.

969 70. Percudani R, Peracchi A. The B6 database: a tool for the description and classification of
970 vitamin B6-dependent enzymatic activities and of the corresponding protein families. BMC
971 Bioinfo. 2009; 10: 273.

972 71. Preciat-Gonzalez GA, El Assal LR, Noronha A, Thiele I, Haraldsdóttir HS, et al. Comparative
973 evaluation of atom mapping algorithms for balanced metabolic reactions: application to Recon
974 3D. Journal of Cheminformatics. 2017; 9: 39.

975 72. Kim J, Kershner JP, Novikov Y, Shoemaker RK, Copley SD. Three serendipitous pathways in
976 E. coli can bypass a block in pyridoxal-5'-phosphate synthesis. Mol Syst Biol. 2010; 6. 436

977 73. Caspi R, Billington R, Ferrer L, Foerster H, Fulcher CA, Keseler IM, et al. The MetaCyc
978 database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome
979 databases. Nucleic Acids Res. 2016; 44: D471 – D480.

980 74. Loira N, Zhukova A, Sherman DJ. Pantograph: A template-based method for genome-scale
981 metabolic model reconstruction. J Bioinform Comput Biol. 2015; 13, 1550006.

982 75. Schönknecht G, Chen WH, Ternes CM, Barbier GG, Shrestha RP, Stanke M, et al. Gene
983 transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote.
984 Science. 2013; 339: 1207 – 1210.

985 76. Cormier A, Avia K, Sterck L, Derrien T, Wucher V, Andres G, et al. Re-annotation, improved
986 large-scale assembly and establishment of a catalogue of noncoding loci for the genome of the
987 model brown alga *Ectocarpus*. The New Phytologist. 2017; 214: 219-232.

988 77. Prigent S, Frioux C, Dittami SM, Thiele S, Larhlimi A, Collet G, et al. Meneco, a topology-
989 based gap-filling tool applicable to degraded genome-wide metabolic networks. PLoS Comput
990 Biol. 2017; 13: e1005276.

991 78. Prigent S, Collet G, Dittami SM, Delage L, Ethis de Corny F, Dameron O, et al. The genome-
992 scale metabolic network of *Ectocarpus siliculosus* (EctoGEM): a resource to study brown
993 algal physiology and beyond. Plant J. 2014; 80: 367 – 381.

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

996 80. King ZA, Lu J, Dräger A, Miller P, Federowicz S, Lerman JA, et al. BiGG Models: A
997 platform for integrating, standardizing and sharing genome-scale models. Nucleic Acids Res.
998 2016; 44: D515 – D522.

999 81. Moretti S, Martin O, Van Du Tran T, Bridge A, Morgat A, Pagni M. MetaNetX/MNXref –
1000 reconciliation of metabolites and biochemical reactions to bring together genome-scale
1001 metabolic networks. Nucleic Acids Res. 2016; 44: D523 – D526.

1002 82. Lifschitz V. What is answer set programming? In AAAI'08 Proc. of the 23rd national
1003 conference on artificial intelligence, Cohn A (ed.), 2008; 1594 – 1597. Chicago, Illinois:
1004 AAAI Press.

1005 83. Gebser M, Kaminski R, Kaufmann B, Schaub T. Answer set solving in practice. Synth Lect
1006 Artif Intell Mach Learn. 2012; 6: 1 – 238.

1007 84. Sievers F, Higgins DG. Clustal Omega, accurate alignment of very large numbers of
1008 sequences. Methods Mol Biol. 2014; 1079: 105 – 116.

1009 85. Gouy M, Guindon S, Gascuel O. SeaView version 4: A multiplatform graphical user interface
1010 for sequence alignment and phylogenetic tree building. Mol Biol Evol. 2010; 27: 221 – 224.

1011 86. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by
1012 maximum likelihood. Syst Biol. 2003; 52: 696 – 704.

1013 87. Le SQ, Gascuel O. Accounting for solvent accessibility and secondary structure in protein
1014 phylogenetics is clearly beneficial. Syst Biol. 2010; 59: 277 – 287.

1015 88. Anisimova M, Gascuel O. Approximate likelihood-ratio test for branches: A fast, accurate,
1016 and powerful alternative. Syst Biol. 2006; 55: 539 – 552.