Molecular BioSystems

RSCPublishing

Journals Molecular BioSystems Advance Articles DOI: 10.1039/b915913b

Mol. BioSyst., 2009 | DOI: 10.1039/b915913b | Paper

An integrative approach towards completing genome-scale metabolic networks† ‡

Nils Christian [a], Patrick May [a], Stefan Kempa [b], Thomas Handorf [c] and Oliver Ebenhöh [ade]

[a]*Max-Planck-Institute for Molecular Plant Physiology, Potsdam-Golm, Germany. E-mail: ebenhoeh@abdn.ac.uk*

[b]*Max-Delbrück-Centrum für Molekulare Medizin, Berlin-Buch, Germany*

[c]*Institute for Biology, Humboldt University Berlin, Berlin, Germany*

[d]*Institute for Biochemistry and Biology, University of Potsdam, Potsdam, Germany*

[e]*Institute for Complex Systems and Mathematical Biology, University of Aberdeen, United Kingdom*

Received 10th July 2009 , Accepted 13th August 2009

First published on the web 10th September 2009

Genome-scale metabolic networks which have been automatically derived through sequence comparison techniques are necessarily incomplete. We propose a strategy that incorporates genomic sequence data and metabolite profiles into modeling approaches to arrive at improved gene annotations and more complete genome-scale metabolic networks. The core of our strategy is an algorithm that computes minimal sets of reactions by which a draft network has to be extended in order to be consistent with experimental observations. A particular strength of our approach is that alternative possibilities are suggested and thus experimentally testable hypotheses are produced. We carefully evaluate our strategy on the well-studied metabolic network of *Escherichia coli*, demonstrating how the predictions can be improved by incorporating sequence data. Subsequently, we apply our method to the recently sequenced green alga *Chlamydomonas reinhardtii*. We suggest specific genes in the genome of Chlamydomonas which are the strongest candidates for coding the responsible enzymes.

Introduction

The rapid development of high-throughput techniques has enabled biological researchers to acquire immense amount of data and has triggered the advent of a multitude of 'omics' disciplines. The improvement of sequencing technologies resulted in the availability of the full genomic sequences for several hundred organisms.[1] Monitoring the activity of the genes has almost become routine through the advances in microarray technologies.[2,3] Many of the proteins, the gene products, may today be quantified by measuring peptides with modern mass spectrometry approaches and matching the identified peptides to databases.[4,5] Finally, various chromatographical methods combined with mass spectrometry allow to simultaneously measure the level of hundreds of metabolites.[6] A major focus of systems biology research is now to integrate this flood of data to arrive at a comprehensive, systems-wide view of living organisms.[7]

A particular challenge is imposed by the fact that every measured dataset is necessarily incomplete. Even though obtaining a genome sequence is now relatively easy, by far not all gene models are defined and annotated and despite the rapid technological improvement in mass spectrometry, still only a small fraction of metabolites can unambiguously be identified. This demonstrates the necessity to develop theories and methods which can cope with incomplete data and nevertheless provide a systemic description of cellular processes. The traditional bottom-up view on metabolism, in which reactions form a pathway and interacting pathways define the metabolic system is clearly insufficient if one has to deal with unavoidable gaps in the knowledge of single reactions.

The difficulty imposed by incomplete data becomes evident when considering recent approaches to analyze genome-scale metabolic networks. For example, Ibarra and co-workers[8] have successfully applied the mathematical framework of flux balance analysis (FBA, see for example refs. 9 and 10) to predict flux distributions resembling optimal growth rates in *Escherichia coli* for different nutrient conditions. This analysis required a complete, or at least consistent, genome-scale metabolic network model. The development of such models involves a time consuming manual verification of every single reaction,[11] which is in stark contrast to the modern high-throughput technologies which yield a tremendous amount of data in a very short time.

In this work, we present a systems biology strategy aiming at integrating available data from the various modern 'omics' technologies. Our top-down approach is specifically designed to accept the incomplete nature of experimental data. By embedding genomic and metabolomic data into bioinformatics and structural modeling approaches, our strategy is suited to extend incomplete metabolic network models to make them consistent with

experimental observations. The simple rationale behind our approach is that if the precise chemical composition of the growth medium is known and metabolic products have been observed or are inferred from biological reasoning, the underlying metabolic network must provide routes to produce these metabolites from the nutrients. Theoretically, it has been shown that such a metabolic reconstruction approach is an *NP*-hard problem.[12] Therefore, we designed an algorithm that calculates a large variety of theoretically possible minimal network extensions which all make the incomplete draft network compliant with the observed functions. Bioinformatics prediction methods allow to rank different extensions with respect to their biological plausibility. Thus, our proposed methodology is qualified to deduce experimentally testable hypotheses from unfiltered and incomplete information of heterogeneous origin.

Several approaches have been reported that aim at completing draft metabolic networks which have been derived from genome sequences by homology matching of DNA or protein sequences with known genes of metabolic enzymes. Such approaches, often referred to as *genome context analysis*, usually involve the identification of the missing parts, the identification and ranking of candidate genes and their experimental verification.[13] In most studies (see for example refs. 14–17) the local context of certain reactions within predefined pathways is investigated to identify gaps in metabolic maps. These bottom-up approaches bear the danger of missing reaction routes that deviate from classical pathways. A systemic top-down strategy was proposed by ref. 18, aiming at extending draft networks to fulfill the condition that a defined set of metabolites, the biomass, can be produced at steady state from selected nutrients. The approach employs FBA, which is mathematically described as an optimization problem, resulting in a single solution or a small number thereof.

Our approach overcomes this restriction by providing a large variety of solutions without the need to reiterate the calculations with additional constraints. Genomic sequence information is exploited to obtain hints which solutions are more plausible to be correct. Our strategy is widely applicable and not restricted to particular organisms. Moreover, our approach is error tolerant and yields plausible predictions for networks that are directly retrieved from databases without the need of prior manual curation. As a proof of principle, we first apply our methodology to the well-characterized metabolic network of *E. coli*. Secondly, we use our approach to extend the draft metabolic network of the recently sequenced green alga *Chlamydomonas reinhardtii*[19] to achieve an improved genome annotation and a more complete metabolic network. Experimentalists can benefit from our theory by obtaining testable hypotheses on gene functions and metabolic pathways. The generation of genome-scale network models which are consistent with available experimental observations are in turn of great value to theoreticians concerned with the analysis of genome-scale metabolism.

Results

General strategy

Our strategy is based on the simple biological observation that all metabolites which have been experimentally detected within a cell culture or an organism must have been produced by the organism's metabolism from the available nutrients.

The general approach to infer completions of draft metabolic networks is illustrated in Fig. 1. Genomic information is used to initially draft a metabolic network model. Experimental data, in particular measured metabolites, are exploited to define functions that the network necessarily must possess. Whether a network provides the synthesis routes necessary to fulfill these functions is tested employing the method of network expansion.[20–22] The draft network is then embedded in a much larger reference network from biochemical databases such as KEGG[23] or MetaCyc.[24]

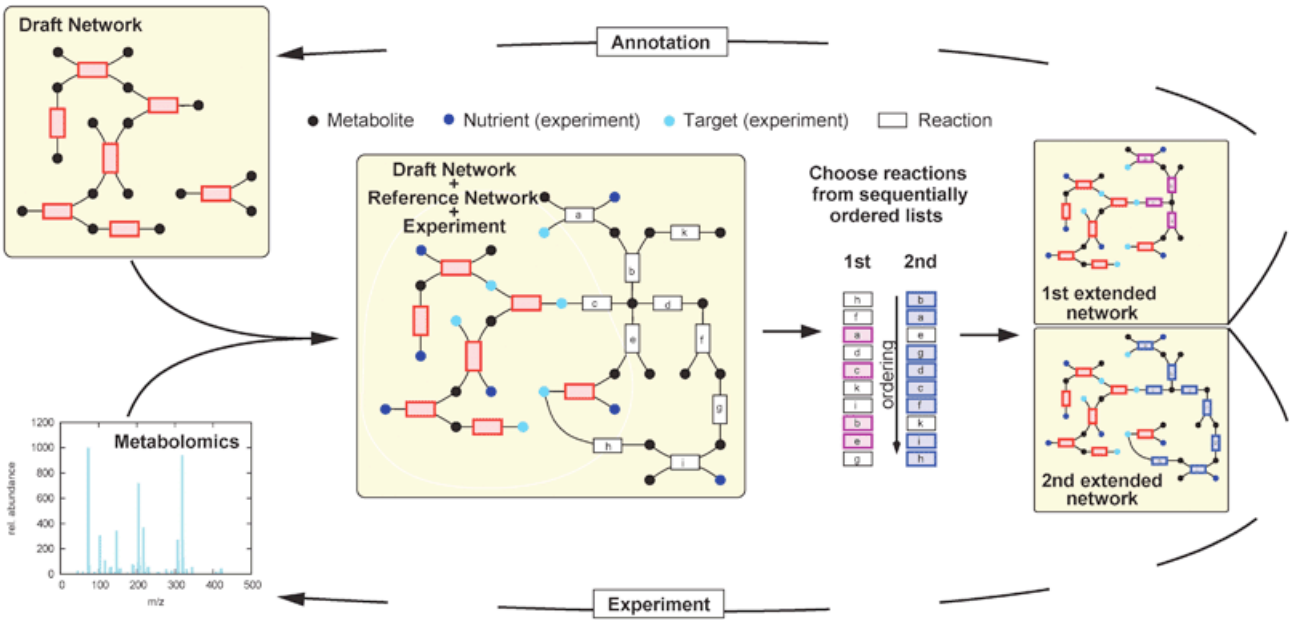

Fig. 1 *Integrative approach using omics techniques and mathematical modeling to improve the metabolic network.* The initial draft network is derived from genomic sequence data. In general, the architecture of a draft network is not sufficient to explain the presence of all metabolites observed in metabolomics measurements. The draft network is embedded in a reference network consisting of reactions collected in databases such as MetaCyc or KEGG. A greedy algorithm calculates minimal sets of reactions, so called *extensions*, that have to be added to the draft network to make it compliant with all experimental data. A network is in agreement with observations if it is able to carry fluxes producing the measured metabolites from the applied nutrient medium. The calculation of a large number of extensions is achieved by initializing the algorithm with many differently ordered lists of reactions (see text). In this process, genomic sequence information is incorporated to ensure that as a tendency those reactions are preferentially included in an extension for which there exists high significance that a gene coding for a catalyzing enzyme is present. The solutions are compared and used to derive hypotheses about the existence of biochemical reactions and genes encoding the respective enzymes. These hypotheses can be tested experimentally or with bioinformatics methods. With this strategy, modeling, bioinformatics and experiment are combined in an iterative process to improve gene annotations and arrive at more complete genome-scale metabolic networks.

The core of our strategy is an algorithm that determines minimal sets of reactions, so called *extensions*, which have to be added to the network draft in order to make it compliant with all experimental observations. In a first step of our greedy algorithm, all reactions from the reference network are added to the draft network. From these additional reactions, every single one is temporarily removed and it is verified whether the network is still fully functional. If this is the case, the considered reaction was not strictly necessary and is permanently removed. Otherwise, the reaction is kept in the extension, because its presence is apparently required to obtain agreement with experimental data.

Clearly, the obtained extensions strongly depend on the order in which the reactions are temporarily removed. We explicitly exploit this fact and systematically determine large numbers of extensions. This is achieved by creating many randomized lists of reactions. With available gene and protein sequences, we build reaction-specific but species independent profile hidden Markov models[25] (see Fig. 2 and Methods). These models allow to assess which enzymes are most likely to be encoded in the genome. We incorporate this information into the randomization procedure in such a way that preferentially those reactions are included in an extension for which a high probability is observed that a catalyzing enzyme is encoded in the genome. This leads to a considerably improved prediction of missing reactions.

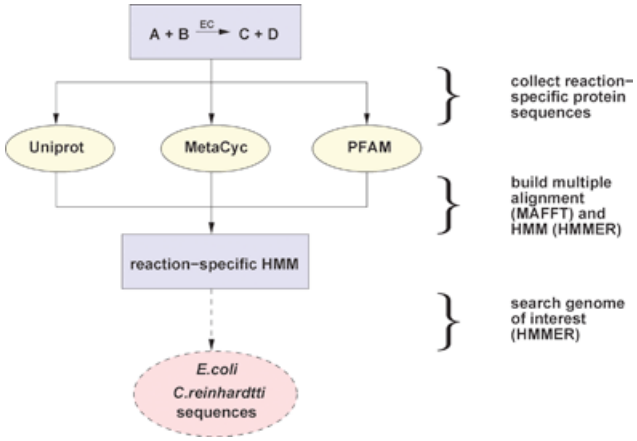

Fig. 2 *Reaction-specific HMMs and E-values.* For a given reaction, enzyme sequences are collected from databases and are used to build reaction-specific and species-independent profile hidden Markov models (HMMs). From these HMMs *E*-values are calculated which reflect the probability that an enzyme catalyzing a reaction is encoded in the genome of the studied

organism. In this paper HMMs were applied to the genomes of *E. coli* and *Chlamydomonas reinhardtii*.

We systematically evaluate our computational predictions by applying our strategy to the well-investigated model organism *E. coli*. We then apply our methodology to the recently sequenced green alga *Chlamydomonas reinhardtii* to put forth hypotheses regarding the extension of its draft metabolic network. We contribute to improved gene annotations by proposing candidate genes coding for enzymes catalyzing missing reactions.

Proof of principle

To demonstrate that our method yields biologically relevant results, we verify that our predictions meet the expectations for a well-studied example. As one of the most thoroughly investigated organisms, we selected *E. coli* and retrieved the metabolic reactions from the EcoCyc[26] database. While it can be expected that this collection will further expand in the future, we assume that the metabolic network of *E. coli* in its present state is, among all available organism-specific networks, closest to being complete. From the information contained in EcoCyc, we have assembled a reaction network containing more than 1500 reactions, including enzymatic and spontaneous reactions, connecting about 1500 chemical compounds (see Methods). We have decided to perform our case study on a network which has been directly retrieved from a metabolic database instead of a published curated model, as for example presented in refs. 27 and 28, for two reasons. First, because networks retrieved directly from databases tend to contain some stoichiometric inaccuracies, in particular with respect to protons and water moieties on both sides of the reactions, we can show that our method is more tolerant against these types of errors than those that use FBA or related concepts. Secondly, the technical task to embed a network into a larger network of a different format is tedious and it is hard to rule out the possibility that the predicted extensions contain artifacts resulting from different naming conventions of metabolites or reactions.

E. coli, as a generalist, may grow on a huge variety of growth media. For example, it displays rapid growth when it is grown on glucose as the only carbon source. This implies that *E. coli*'s metabolism is capable of producing all necessary precursors for higher level processes, such as macromolecule assembly, by consuming exclusively glucose and other, non-carbon containing substrates. This includes in particular the formation of all twenty amino acids, the nucleotide phosphates ATP, CTP, GTP and UTP, as well as the deoxy forms dATP, dCTP, dGTP and dTTP required for protein, RNA, and DNA synthesis. For our case study, we focus on the minimal metabolic function that a metabolic network describing the biochemistry of *E. coli* must be able to synthesize all these 28 target metabolites from glucose and inorganic material. As expected, the retrieved metabolic network can perform this essential function. The selected 28 target metabolites make up by far the largest fraction of metabolites commonly considered as substrates necessary for biomass production in other genome-scale metabolic models of *E. coli*.[27,28] Thus, the chosen condition corresponds approximately to the condition that genome-scale models would show growth.

In the following, we construct a large ensemble of incomplete networks through the removal of reactions. We first investigate how the functionality is impaired upon this deletion. Secondly, we repair the networks using a set of more than 4500 reactions extracted from MetaCyc and compare our predicted extensions to the reactions that were originally removed, allowing to define and assess the quality of the predicted network extensions.

We mimic draft networks of different levels of incompleteness by randomly removing 20, 50, 100 and 200 reactions from the full *E. coli* network, respectively. For each case, 100 draft networks have been constructed. Each of the resulting 400 reduced networks imitates a draft network for an organism whose genome has only been partially annotated and that therefore shows an incomplete architecture. In contrast to the full network, most of the constructed draft networks do not display the full capability to produce all 28 target metabolites. Clearly, how many and which particular precursors cannot be produced depends on the specific reactions that have been removed. In Fig. 3A, a histogram over the number of targets which cannot be produced by the reduced networks is shown. The bars are separated to indicate how strong the original network has been reduced (20, 50, 100 and 200 reactions removed). As expected, the tendency can be observed that a larger number of removed reactions leads to more targets which can no longer be produced. In Fig. 3B, the effect of reaction removal on the particular target production routes is depicted. The production of simpler, non-aromatic, amino acids seems more robust than that of more complex amino acids and nucleotides. This indicates that for the latter metabolites the synthesis routes show a lower degree of redundancy and therefore the removal of reactions is more likely to result in the loss of their producibility.

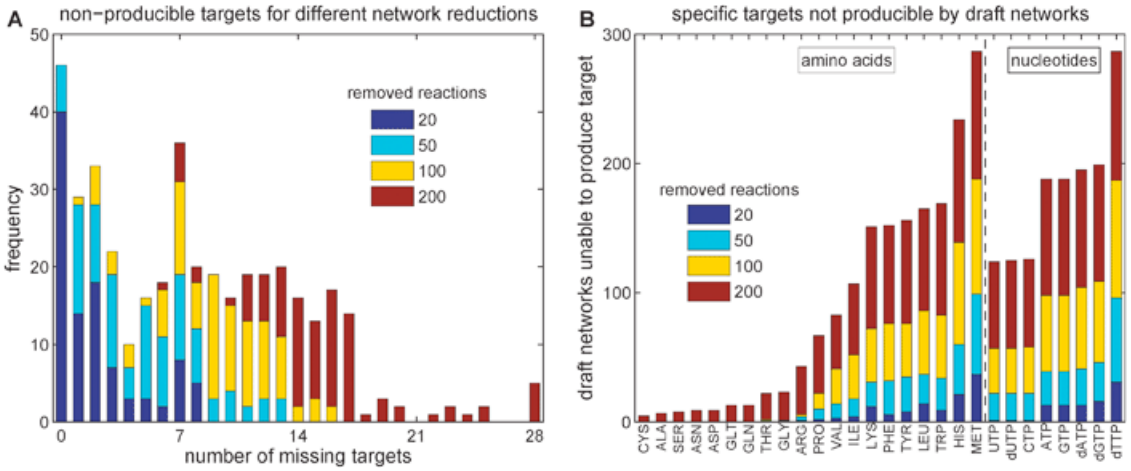

Fig. 3 *Effect of network reductions on the producibility of target metabolites*. (A) Shown is a histogram of the numbers of target metabolites which can no longer be produced after removal of reactions. (B) Shown are the numbers of draft networks which are unable to synthesize particular target metabolites. In both figures, the bars are separated to indicate networks of different degrees of incompleteness (20, 50, 100 and 200 reactions have been removed to generate 4 × 100 = 400 reduced networks).

While the investigation of the effect of random network perturbations is extremely interesting and useful to elucidate the robustness of network architectures with respect to essential functions, the main focus of the present paper lies elsewhere, namely in the identification of missing links in incomplete networks. These gaps are filled by minimal collections of reactions, so-called extensions, recovering the functionality of the original network. Since we identify minimal sets of reactions, it cannot be expected that all removed reactions are recovered, because many reactions are not involved in the production of the target metabolites. Hence, it is impossible to predict such reactions by a strategy based on recovering the particular network function given by the producibility of the targets. We will, however, demonstrate that it is possible to identify those missing reactions which are strictly required to perform essential metabolic functions.

In the artificially produced draft networks we know exactly which reactions were removed. Therefore, they provide an ideal background to assess the predicted extensions. A good prediction should propose a high number of previously removed reactions and a low number of other reactions which are not found in the original metabolic network of *E. coli*. To quantify the correctness of an extension, we introduce the quality measure $q(E)$ of an extension E

$$q(E) = \frac{T(E)}{N(E)}, \tag{1}$$

where $T(E)$ denotes the number of correctly predicted reactions within an extension E and $N(E)$ denotes the extension size, *i.e.* the total number of reactions within extension E. In this way, a value of $q = 1$ describes a perfect prediction containing exclusively reactions which were previously removed and a value of $q = 0$ characterizes the worst possible prediction consisting only of reactions not found in the original *E. coli* network.

Since the calculated extensions strongly depend on the order in which the reactions are traversed, we have generated for each of the 400 draft networks 100 completely randomized reaction lists. Every resulting extension ensures that the capability to produce all 28 target metabolites is regained. A histogram of the corresponding prediction quality measures (1) is shown in Fig. 4A. Interestingly, those extensions containing exclusively reactions that have previously been removed ($q = 1$) and those containing none of these reactions ($q = 0$) show the highest relative abundance. However, this does not hold true for strongly reduced networks. For those cases in which 200 reactions have been removed, almost no extension falls into one of these categories. One reason for this is that less incomplete network can in general be fixed by simpler extensions and these are more likely to assume extreme values.

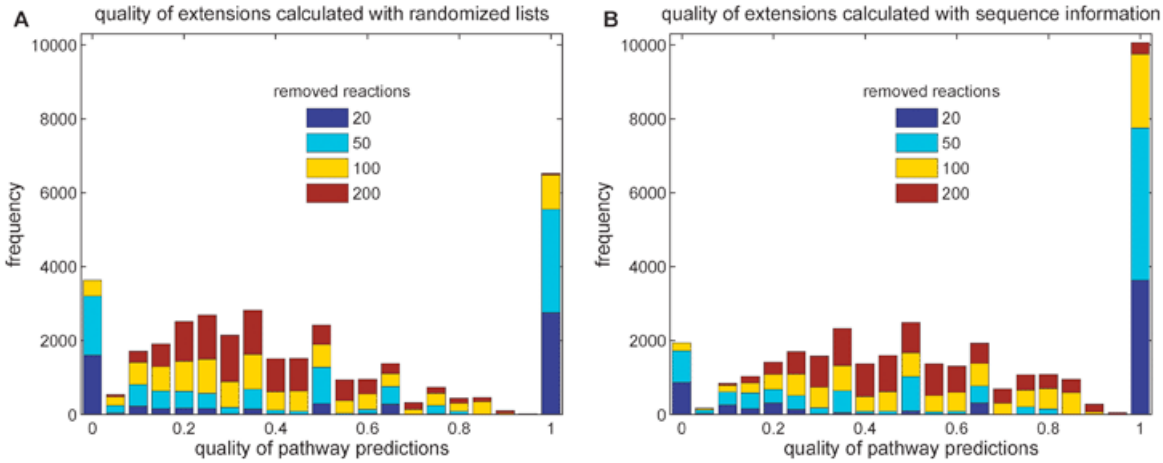

Fig. 4 *Histogram of the quality of predicted extensions for the artificially reduced metabolic network of E. coli.* (A) Extensions have been calculated without including sequence information. (B) Extensions were determined such that reactions with a high probability that coding enzymes are present in the genome of *E. coli* are preferentially included. The shuffling parameter was set to $\beta = 0.04$. In both figures the quality is measured by the fraction of correctly predicted reactions in a calculated extension. The bars are stacked indicating the contributions of pathway predictions obtained for different numbers of removed reactions.

Randomizing the sequential order of the reaction lists means that all reactions are treated equally. If, for example, two reactions could alternatively be included in a functional extension, both reactions will be found with equal probability. This procedure is adequate if no further information about the putative reactions is available. However, the genome sequence may provide valuable hints on which enzymes may be encoded. To consider such information, we collect for each reaction in MetaCyc all available protein sequences to build reaction-specific profile hidden Markov models (HMMs)[25] from the resulting multiple sequence alignment. These multiple-species models allow to search protein sequences for functionally related sequences (see Fig. 2 and Methods). For the best match of every reaction-specific HMM to a protein encoded in the genome of *E. coli*, an *E*-value was determined, where a low value close to zero indicates a high chance that a catalyzing enzyme is present.

To ensure that reactions with a low *E*-value are preferentially included in the extensions, we create reaction lists which are not completely randomized but retain the tendency that reactions with lower *E*-values are positioned towards the end of the lists. For this, a scalable parameter β is introduced which is related to the probability that a reaction with a higher *E*-value may be placed behind a reaction with a lower *E*-value (see Methods). An infinite value β results in a strictly ordered list whereas $\beta = 0$ characterizes the completely randomized case described above. We have systematically investigated the effect of the ordering parameter β (see Supporting Text S1, ESI‡) and chosen an intermediate value of $\beta = 0.04$ for the calculations presented here. The corresponding histogram of quality measures (1) is depicted in Fig. 4B. Comparison of Fig. 4A and B reveals that considering sequence information indeed improves the prediction quality in general and the fraction of perfect predictions in particular (10008 or 28% against 6463 or 18% perfect predictions).

To ensure that these findings hold true not only for the special case of the well studied organism *E. coli*, we have repeated the presented analysis for *B. subtilis*, for which the tier 3 network has been retrieved from the BioCyc database collection. For *B. subtilis*, a simple chemically defined growth medium exists,[29] from which all necessary precursors must be produced. In fact, the retrieved network is already capable to provide these essential precursors. In analogy to our studies for *E. coli*, we have again generated 400 reduced networks and for each calculated 100 extensions. The figure corresponding to Fig. 4 is given in the Supporting Text S2, ESI.‡ It is striking that inclusion of sequence information yields a much stronger improvement of the prediction quality than for *E. coli*. The explanation for this lies in the fact that the retrieved network for *B. subtilis* has been obtained exclusively by sequence homologies without further curation. As a consequence, our algorithm will identify the original reactions with a higher probability if sequence information in the form of *E*-values is included.

We expect that including genomic information will also increase the fraction of correct predictions when applying our algorithm to real draft networks for which a quality measure cannot be determined. Since the applied algorithm detects a large variety of theoretically possible extensions to regain functionality, in the real case a level of uncertainty will necessarily remain. Bioinformatics methods are useful to obtain hints on the likelihood of alternative predictions, but ultimately the candidate extensions have to be verified experimentally.

Extending the network of *Chlamydomonas reinhardtii*

We have applied our proposed method to the metabolic network of the green alga *Chlamydomonas reinhardtii*. The sequenced genome of this organism has recently been published[19] and based on this sequence, a draft metabolic network has been assembled and compiled in the ChlamyCyc database[30] (see Methods). At the current state, the network contains about 1500 metabolites and 1200 reactions. Chlamydomonas cultures grown photoautotrophically have been subjected to extensive metabolomics measurements and 159 metabolites could uniquely be identified using gas chromatography–time of flight–mass spectrometry (GC–TOF–MS).[31,32] Of these, 138 metabolites could unambiguously be mapped to compounds present in the MetaCyc database. The experimental evidence for the presence of these metabolites entails that the metabolic network of Chlamydomonas must be capable of producing these from the supplied nutrient medium. We have tested whether the draft network is already able to carry fluxes necessary for their production and found that 87 from the 138 metabolites may in fact be produced. Among the producible compounds are also the 28 essential precursor metabolites considered in the previous section for *E. coli*. Interestingly, 20 of the remaining 51 metabolites cannot even be produced by the reference network comprising all reactions found in the MetaCyc database (a complete list is given in the Supporting Text S6, ESI‡). This finding indicates that our knowledge on metabolism, even when combining information from hundreds of organisms, is still far from complete.

To identify possible extensions for the draft network of Chlamydomonas we have applied our algorithm by embedding the draft network into the complete MetaCyc network with the task to compute sets of reactions that, when added to the draft network, enable it to carry fluxes for the production of the remaining 31 metabolites. For this, we have first calculated *E*-values for each reaction present in MetaCyc but not in the draft network, based on sequence similarities to the coding region in the Chlamydomonas genome using the reaction-specific hidden Markov models described in the Methods section. Based on the *E*-values, we have generated 10 000 randomized lists of reactions such that as a tendency those reactions with a high probability that coding enzymes are found in the genome are placed near the end, resulting in their preferential incorporation into network extensions (see Methods; similarly to *E. coli* we set $\beta = 0.04$). Based on these lists, 10 000 possible extensions have been calculated.

The distribution of the extension sizes for the calculated network extensions are depicted in Fig. 5A. The smallest set of reactions providing the Chlamydomonas network with maximal functionality (capability to produce the 31 target metabolites) contains 52 reactions, the largest set 95 reactions. These values give an impression at how incomplete the existing draft network, built exclusively on genomic sequence information, still is. In total, the 10 000 extensions contained 598 distinct reactions. In Fig. 5B the relative occurrence of all these reactions within all 10 000 calculated extensions is displayed where the reactions have been ordered with decreasing frequency. Interestingly, 15 reactions are found in every extension, while 466 reactions occur in less than 10% of all extensions.

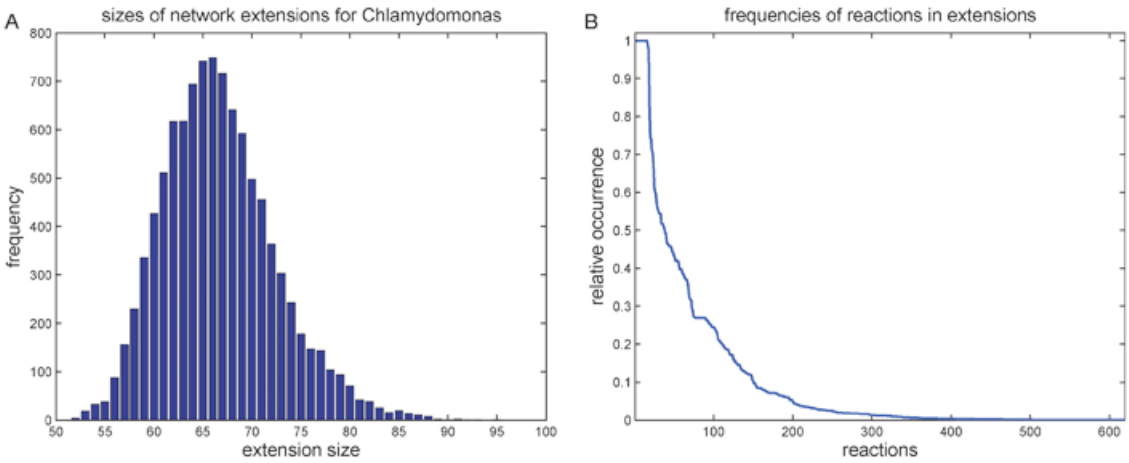

Fig. 5 *Extension sizes and frequency of reactions within extensions for the draft network of Chlamydomonas reinhardtii*. (A) Shown is a distribution of the numbers of reactions within the calculated extensions. (B) The relative occurrence of single reactions within the 10 000 calculated extensions are displayed. The reactions have been ordered with decreasing frequency.

Various reasons may be responsible for the fact that metabolites have been observed but the draft network does not include reactions required for their production. Many reactions have been characterized biochemically several decades ago, but no protein sequences for catalyzing enzymes are available. In such a case, annotation of the coding genes is impossible. Gene annotations may also have failed due to low sequence similarities. Further, it cannot be excluded that the observed metabolites have been produced chemically post extraction and are not truly part of the metabolism. Finally, the possibility has to be considered that reactions producing a particular metabolite are simply not known.

In the following paragraph we will discuss in detail examples of incomplete annotation.

Ergosterol. The Chlamydomonas genome encodes for several proteins involved in sterol biosynthesis. However, the annotated enzymes do not suffice to allow for full biosynthetic pathways. For example, ciliary membranes of Chlamydomonas containing multiple sensory proteins have been identified as enriched with ergosterol,[33] a sterol which may not be produced by the draft network. We obtained several predictions within the ergosterol biosynthesis pathway, depicted in Fig. 6, which we will discuss in the order of the reactions. A key step in sterol biosynthesis is squalene monooxygenase (EC 1.14.99.7). This enzyme has not yet been annotated in Chlamydomonas, but the corresponding reaction has been found in every calculated extension. This step is of major importance also for the production of other sterols. By sequence comparison we could identify a good but not yet annotated candidate gene model with a gene product homolog to several proteins in a variety of species. Fig. 7 displays the phylogeny of the protein squalene monooxygenase. Clearly, the candidate protein belongs to the group of monooxygenases. It cannot, however, be unambiguously assigned to groups formed by plant, animal or fungal proteins. It rather seems to represent a distinct clade together with other algal species, depicted in cyan in Fig. 7. These findings indicate that this essential protein has diverged early during evolution from the orthologs found in other eukaryotes. Moreover, we got additional evidence from proteomics studies that this gene model encodes a protein present in Chlamydomonas. We found at least one peptide matching the proposed protein sequence with high confidence (unpublished data).

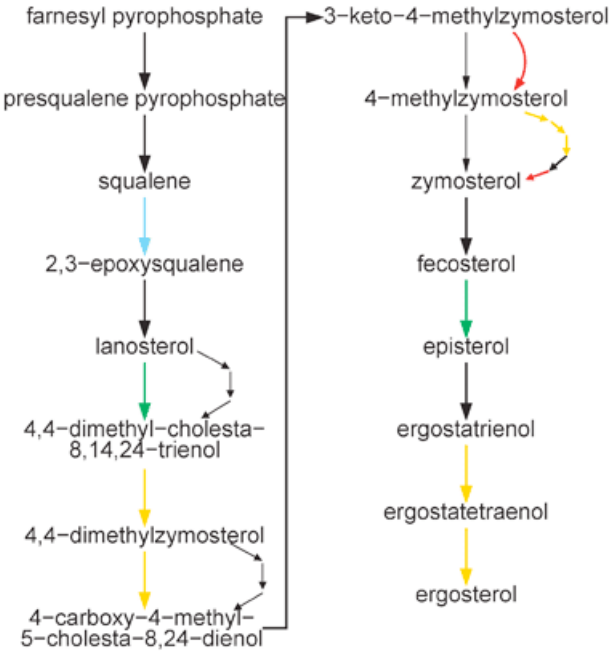

Fig. 6 *The ergosterol biosynthesis pathway*. Shown is the pathway as annotated in yeast (Main pathway). The alternative routes correspond to reactions annotated in the human cholesterol pathways. These reactions represent analog chemical conversions but may differ in the degree of detailedness and used cofactors. Both possibilities have been detected by our algorithm. Thick black arrows indicate reactions present in the draft network. The remaining reactions are color coded to indicate the species where the closest homologs were found (green—plants, yellow—fungi, red—animals). The blue arrow for squalene monooxygenase indicates that this enzyme forms a distinct subgroup in algae. Reactions for which no clear sequence similarities could be identified are marked by thin black arrows. A detailed graphical representation of the pathway is given in the Supporting Text S7, ESI.‡

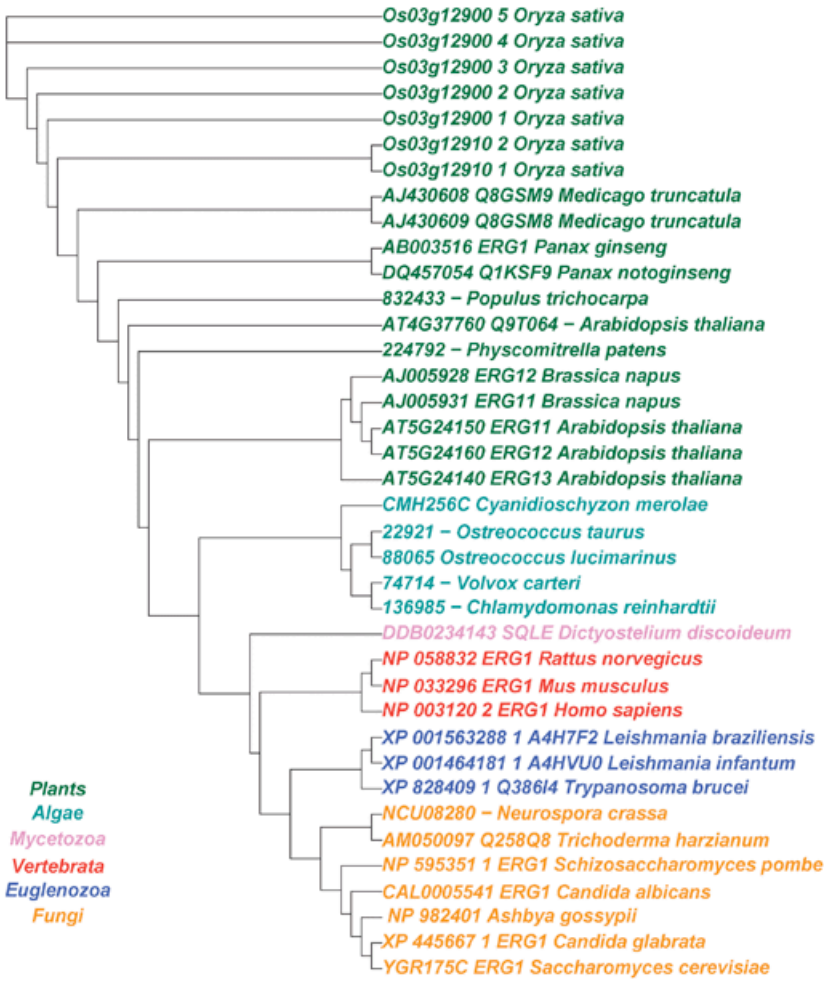

Fig. 7 Phylogenetic relationship of the putative squalene monooxygenase found in Chlamydomonas to orthologs from other organisms.

For some of the intermediate steps of ergosterol biosynthesis, our algorithm seemingly found alternative routes. In four cases, two alternative reaction sequences were identified from which at least one is required to explain the existence of ergosterol. A detailed investigation of the alternative reaction sequences reveals that they represent similar or even identical reaction steps. While they appear in different pathways, as they are defined in MetaCyc, they do in fact simply represent a different degree of detail, thus explaining the different number of steps. The fact that these alternative descriptions were found demonstrates the principle ability of our approach to identify alternative solutions to explain the presence of observed metabolites. Despite the close resemblance of the alternatives, it is still interesting to observe that in all four cases one alternative corresponds to a part of the ergosterol biosynthesis pathway annotated in yeast, while the other alternative corresponds to a part of the cholesterol pathway as annotated in human (for a detailed representation of the alternative routes, see Supporting Text S7, ESI‡). Apart from the different detailedness regarding the number of reaction steps, the human and yeast pathways differ in the cofactors that are used. While these differences are not critical for the extension of the draft network to consistency, it is nevertheless interesting to investigate whether in Chlamydomonas this pathway is closer related to the kingdom of fungi, animals, or plants. We performed phylogenetic analyses for various of the involved proteins (see Supporting Text S8, ESI‡) and found that there is no clear tendency towards a similarity with one particular kingdom. Rather, some proteins seem to bear a high similarity to plant proteins while others are more similar to fungal or human proteins.

The final steps in the biosynthesis of ergosterol involve five consecutive reactions in which zymosterol is converted into ergosterol. Two of the five catalyzing enzymes have already been assigned to coding genes in the genome of Chlamydomonas. The remaining three are found in every calculated extension. For the C-8 sterol isomerase (isomerizing fecosterol into episterol) we identified a gene with homologs in *Arabidopsis thaliana* and mouse, for the C-22 sterol desaturase (desaturating an ergostatetraenol at the 22nd position to an ergostatetraenol), we found a gene with a clear ortholog in yeast. The last step from ergostatetraenol to ergosterol, catalyzed by the C-24 sterol reductase (EC 1.3.1.71), is biochemically very similar to the C-14 sterol reductase (EC 1.3.1.70) for which we found a clear ortholog against several organisms. Indeed, this protein also displays a high similarity to the C-24 sterol reductase (ERG4) gene in yeast. We therefore speculate that the identified protein is able to catalyze both reactions.

In Table 1 we list the putative enzymes of the Ergosterol synthesis pathway. In the first column reactions are specified by their EC number. In cases where no EC number is assigned to a reaction, an alternative name is given. In the second column, we state evidence obtained by sequence homology.

Table 1 Evidence for predicted enzymes of the Ergosterol pathway in *Chlamydomonas reinhardtii*

Reaction/EC number	Evidence
1.14.99.7	Blast hit (136 985) against human (ERG1)
1.1.1.270	Blast hit (191 061) against human (DHB7)
1.3.1.70	Orthologs (196 516, 126 431) to yeast (ERG24)
1.3.1.71	Blast hit (196 516) against yeast (ERG4)
1.14.13.70	Ortholog (196 411) to Arabidopsis (AT1G11680)
1.14.13.72	Orthologs (142 288,186 886) to human (NP_006736.1)
C-8 sterol isomerase	Blast hit (160 258) against Arabidopsis (AT1G20050) but more likely C-8,7 sterol isomerase (5.3.3.5)
5.3.3.5	Ortholog (160 258) to Arabidopsis (AT1G20050)
C-22 sterol desaturase	Ortholog (196 874) to yeast (ERG5)

Rhamnose. The deoxyhexose L-rhamnose was uniquely identified but the draft metabolic network does not provide a synthesis route. Interestingly, though, the chemically more complicated compound UDP-rhamnose may be produced by the draft network. The three reactions

$$\text{L-rhamnose} \rightleftharpoons \text{L-rhamnulose} \tag{2}$$

$$\text{L-rhamnulose} + \text{ATP} \rightarrow \text{L-rhamnulose-1-phosphate} + \text{ADP} \tag{3}$$

$$\text{L-rhamnulose-1-phosphate} \rightleftharpoons \text{L-lactaldehyde} + \text{dihydroxyacetonephosphate}, \tag{4}$$

catalyzed by the enzymes L-rhamnose isomerase (EC 5.3.1.14), L-rhamnulose kinase (EC 2.7.1.5) and rhamnulose-1-phosphate aldolase (EC 4.1.2.19) respectively, all appear in every predicted extension. This reaction chain constitutes the rhamnose degradation pathway characterized in *E. coli*. However, the algorithm predicts that all these reactions are operating in reverse direction. This seems unrealistic since rhamnulose kinase is likely to be irreversible under physiological conditions considering the change in free energy resulting from hydrolysis of the γ-phosphate of ATP.

To understand why our algorithm yields this somewhat counterintuitive result, the usual synthesis pathway of rhamnose has to be considered. Rhamnose is, for example, an important component of plant cell wall pectins.[34] The incorporation into pectins occurs through the activated intermediate UDP-L-rhamnose, a compound which may already be produced by the draft network. In Chlamydomonas, the rhamnosylated macromolecules have not yet been identified and the presence of pectin-like structures was not observed. However, fucose, another deoxyhexose, was observed in Chlamydomonas as a constituent of the extracellular matrix.[35] A later degradation of pectin releases free rhamnose, providing a simple explanation why L-rhamnose was experimentally observed. However, this plausible chain of events is not represented in metabolic databases. Such databases focus on the description of biochemical reactions involving relatively small molecules. Macromolecules, if described at all, are represented as generalized compound classes, such as ‘a pectin’ or ‘a protein’. Since our algorithm depends on detailed reaction stoichiometries involving well defined reactants, it is not able to detect pathways involving such compound classes, explaining why instead the degradation pathway in reverse direction was predicted.

The prediction of degradation pathways is nevertheless informative. The existence of free rhamnose strongly suggests that it can be reincorporated into other metabolic processes. Any other assumption is implausible considering the energy required for sugar production. Moreover, a continuous accumulation of rhamnose has not been observed. We did not identify any clear similarity to proteins involved in the predicted degradation pathway (see Table 2), indicating that homology to the bacterial pathway is unlikely. However, the authors are not aware of any studies concerned with the metabolic recycling of free rhamnose in eukaryotic organisms. It remains therefore unclear by which mechanism rhamnose is degraded or otherwise recycled. With the present knowledge of metabolism, the only known route explaining rhamnose degradation is a pathway analogous to the one characterized in *E. coli*.

Table 2 Evidence for predicted reactions outside the ergosterol pathway. Listed are reactions for which we see strong evidence that they must be included in the draft network of *Chlamydomonas reinhardtii*. The presented

candidate reactions have been predicted by our algorithm and their presence is supported by biological reasoning (see text). In the first column, the target metabolite, for which a particular reaction is responsible, is given. Reactions (second column) are specified by their EC number. In cases where no EC number is assigned to a reaction, an alternative name is given. In the third column, we state additional evidence obtained by sequence homology studies. Where applicable, we explicitly state the best candidate genes encoding for corresponding enzymes

Target	Reaction/EC number	Evidence
L-rhamnose	5.3.1.14	No hit
	2.7.1.5	No hit
	4.1.2.19	No hit
	2.7.7.64	Ortholog (32 796) to Arabidopsis (AT5G52560)
	3.1.3.23	Blast hit (196 269) against *E. coli* (SUPH)
Hydroxyproline	Hydroxyproline oxidase	Ortholog (146 649) to Arabidopsis (AT3G30775)
	2.6.1.23 (= 2.6.1.1)	Ortholog (186 959) to Arabidopsis (AT4G31990)
	4.1.3.16	No sequences available
Phenylacetaldehyde	4.1.1.43	Ortholog (135 197) to yeast PDC5
	4.1.1.53	Blast hit (40 158) against *Solanum lycopersicum* AADC1A
Lumichrome	3.5.99.1	No hit
N-acetyl-L-phenylalanine	2.3.1.53	No sequences available

Hydroxyproline. Even though the amino acid derivative hydroxyproline is chemically and functionally unrelated to the sugar rhamnose, some common principles can be understood when studying the predicted production routes. Similar to the case of rhamnose, also for hydroxyproline the degradation pathway was predicted to operate in reverse direction. Hydroxyprolines are important structural components of collagen in animals[36] and of cell walls in plants[37] and algae, including Chlamydomonas.[38] Moreover, in plants hydroxyproline-rich glycoproteins play an important role for providing structural integrity.[39] The usual biosynthesis pathway involves the enzyme prolyl hydroxylase (EC 1.14.11.2), hydroxylating proline to hydroxyproline. It accepts only peptidyl proline as a substrate.[40] Again, the automatic network extension algorithm is not able to relate this reaction to the synthesis of hydroxyproline, explaining why also in this case only degradation pathways are predicted.

In Chlamydomonas under sulfur limiting conditions high concentrations of hydroxyproline have been observed[32] and increased expression levels of mRNAs from hydroxyproline-rich polypeptides have been reported.[41] Upon resupply of sulfur, these mRNAs displayed a rapid decline, suggesting high turnover rates also on protein levels. The resulting free hydroxyprolines have to be reincorporated into metabolism by some degradation pathway. In contrast to rhamnose, hydroxyproline degradation is described and annotated in *A. thaliana*. It involves four consecutive enzymatic reactions, resulting in the degradation products pyruvate and glyoxylate. The first step is mediated by hydroxyproline oxidase, for which we could identify clear orthologs in the genome of Chlamydomonas to enzymes annotated in *A. thaliana*. The second enzyme, Δ^1-pyrroline-3-hydroxy-5-carboxylate dehydrogenase (EC 1.5.1.12), has been annotated previously and was therefore already incorporated into the draft network. For the remaining two enzymes, 4-hydroxyglutamate transaminase (EC 2.6.1.23) and 4-hydroxy-2-ketoglutarate aldolase (EC 4.1.3.16), no protein sequences are available such that more detailed searches could not be performed. Considering the strong evidence that the first two reactions of the degradation pathway are present in the metabolic network of Chlamydomonas, we strongly assume that also enzymes for the latter two reactions are encoded in its genome, since they provide the only plausible route allowing for reutilization of free hydroxyproline (see Table 2).

Phenylacetaldehyde. The aromatic compound phenylacetaldehyde is a volatile compound and plays a role in many floral scents.[42] Its precise role in Chlamydomonas is unclear, however, its presence has been experimentally proven. To explain its production, our algorithm proposed several alternative solutions involving a single reaction. Since a direct precursor, the aromatic amino acid L-phenylalanine, is producible by the draft network, one possibility is to extend the network by the reaction

$$H_2O + O_2 + \text{L-phenylalanine} \rightleftharpoons NH_3 + H_2O_2 + CO_2 + \text{phenylacetaldehyde}, \quad (5)$$

catalyzed by phenylacetaldehyde synthase (no EC number assigned). An alternative is the extension of the network by phenylalanine decarboxylase (EC 4.1.1.53), catalyzing the reaction

$$\text{L-phenylalanine} \rightleftharpoons CO_2 + \text{phenylethylamine}, \quad (6)$$

since phenylethylamine oxidase (EC 1.4.3.6), converting phenylethylamine into phenylacetaldehyde, is already annotated and included in the draft network. The third possibility uses phenylpyruvate as precursor which is also producible by the draft network. The corresponding reaction that needs to be included is

$$\text{phenylpyruvate} \rightleftharpoons CO_2 + \text{phenylacetaldehyde}, \quad (7)$$

catalyzed by phenylpyruvate decarboxylase (EC 4.1.1.43).

Sequence comparison reveals a high probability that the latter two alternatives are indeed present in the genome-scale network of Chlamydomonas. For phenylalanine decarboxylase we obtained a high similarity against a protein in the tomato *Solanum lycopersicum*, for phenylpyruvate decarboxylase we found a clear ortholog in yeast (see Table 2).

Lumichrome. Also lumichrome is among the unambiguously identified metabolites that cannot be produced by the draft network. Lumichrome has also been identified in Chlamydomonas with an independent technique (reverse phase HPLC and UV detection).[43] There, the authors have demonstrated that lumichrome secreted by Chlamydomonas is capable of activating quorum sensing receptors in bacteria, thus disrupting their quorum sensing regulation. This hints at a role of lumichrome in a defense mechanism against bacterial pathogens.

The reaction

$$H_2O + \text{riboflavin} \rightleftharpoons \text{lumichrome} + \text{ribitol}, \quad (8)$$

catalyzed by the enzyme riboflavinase (EC 3.5.99.1) occurred in all 10 000 calculated extensions. Both, riboflavin and ribitol are present in Chlamydomonas' draft network and may be produced from the applied medium. In fact, this is the only reaction found in MetaCyc with lumichrome as reactant. The only protein sequence recorded in the enzyme database BRENDA[44] for an enzyme with this catalytic activity has been obtained from the bacterium *Enterobacter* sp. (strain 638). To this particular sequence, we could not detect any homolog within the genome of Chlamydomonas (see Table 2). This, however, does not exclude the possibility that a gene coding for riboflavinase is present. We consider the existence of a riboflavinase highly plausible since all other explanations for the presence of lumichrome would include hitherto unknown mechanisms.

Acetyl phenylalanine. Also *N*-acetyl-L-phenylalanine, observed but not producible by the draft network, is involved in only one reaction in the MetaCyc database. The enzyme phenylalanine *N*-acetyltransferase (EC 2.3.1.53) catalyzes the reaction

$$\text{L-phenylalanine} + \text{acetyl-CoA} \rightleftharpoons \text{\textit{N}-acetyl-L-phenylalanine} + \text{coenzyme A}. \quad (9)$$

Enzymatic activity was demonstrated in ref. 45 in *E. coli*, but no sequence is available (see Table 2). Therefore it remains uncertain whether the production of *N*-acetyl-L-phenylalanine in Chlamydomonas proceeds by a homologous enzyme or by a different, unknown, mechanism.

Glycerol-2-phosphate. Similarly, glycerol-2-phosphate, detected but not producible by the draft network, participates in exactly one reaction found in MetaCyc, namely

$$H_2O + \text{glycerol-2-phosphate} \rightleftharpoons \text{phosphate} + \text{glycerol}, \quad (10)$$

catalyzed by the enzyme glycerol-2-phosphatase (EC 3.1.3.19). For this enzyme, no protein sequence is available and therefore sequence homology studies could not be performed. As shown in ref. 46, glycerol-2-phosphate can result from the breakdown of *sn*-glycerol-3-phosphodiester derivatives such as phospholipids. We cannot exclude that this reaction occurred during sample preparation and therefore do not consider this reaction as a good candidate to be included in the network of Chlamydomonas.

Discussion

Following the spirit of systems biology research, we propose a novel strategy integrating data resulting from modern high-throughput technologies to improve gene annotations and derive more complete genome-scale metabolic networks. The ability to handle incomplete and heterogeneous information and nevertheless derive biologically plausible and experimentally testable hypotheses makes our methodology widely applicable and

flexible. We envisage further enhancements of our method in several respects. Besides genomic and metabolomic data, existing enzymes can be inferred from proteomics measurements. This information may be incorporated in the definition of the draft metabolic networks or used for validation of the computational predictions.

We have demonstrated that our method produces meaningful results by extending artificially produced draft networks, which we generated by removing reactions from the genome-scale network of *E. coli*. We considered a restricted set of minimal functions that the metabolic network must fulfill, and thus focused our proof of principle analysis to the central metabolism involving the synthesis of amino acids and nucleotides. The functionality of the reduced networks was regained and the predicted extensions were compared with the truly removed reactions. Including sequence information in the form of multiple-species profile hidden Markov models, *E*-values were calculated, reflecting the likelihood that proteins are encoded in the genome. Reactions catalyzed by enzymes with a low *E*-value, indicating a high probability, were preferentially included in the extensions. The importance of the *E*-value is controlled by a parameter β, whose effect is studied in detail in the Supporting Text S1 (ESI‡) to this article. We could show that including sequence information indeed leads to a higher quality of the predicted extensions.

The chosen target compounds compose the largest fraction of biomass precursors as defined in the published genome-scale models for *E. coli*.[27,28] We have decided to select a minimal version of biomass producibility since this condition is sufficient to demonstrate that our method is able to identify realistic completions of networks to consistency. It is in principle straight forward to employ more restricted conditions. Similarly, alternative carbon sources that *E. coli* can reportedly grow on or other combinations of nutrients that experimentally ensure growth can in principle be used to assess the completeness and identify missing reactions. In fact, a greater variety of minimal growth media and a larger set of verified target metabolites will yield more testable hypotheses. An automated algorithm to infer target metabolites is provided by the GapFind,[16] which uses the network topology to find metabolites without a producing reaction and their dependent products. We envisage a possible improvement of the predictions by combining the functional search for missing targets of our approach with the structural prediction of targets provided by GapFind. In the subsequent analysis of the draft metabolic network of *Chlamydomonas reinhardtii* we used a significantly enlarged set of target metabolites which were directly motivated by experimental observations.

In contrast to *E. coli*, Chlamydomonas is a eukaryotic organism. For these, metabolic pathways are not as well characterized as for the simpler prokaryotes. This is in particular true for the localization of the enzymes and thus for the compartments that single pathways operate in. This is reflected in the limited information on enzyme localization contained in metabolic databases. For our calculations, we therefore neglected compartmentation. As a consequence, our predictions cannot include putative transport processes or the localization of enzymes. However, this simplified approach is nevertheless suitable to detect inconsistencies because it is impossible that functions missing in a non-compartmented model can be performed in a more detailed model including subcellular structures. There is no principle limitation of our approach to non-compartmented models and we expect that with the increasing knowledge on enzyme localization and intracellular transport processes, the predictive power of our method will further increase.

For a 'real' draft network, such as that of Chlamydomonas, it is not *a priori* known whether the reference network does provide the correct solution or, in fact, contains any solution at all. We have presented a detailed analysis for the example of the ergosterol biosynthesis pathway. The computed extensions illustrate how the ability to propose several alternatives triggers new research activities. In this particular case, one of two predicted alternatives could always be associated with the known pathway in yeast while the other represented the human pathway. This finding inspired a phylogenetic analysis of the involved enzymes. We obtained hints that squalene monooxygenase, a key enzyme required for the synthesis of many sterols, has diverged early from homolog enzymes found in higher eukaryotes. As closest homologs we could identify genes in other algal species which also have not yet been annotated. This demonstrates how our computational approach in combination with phylogenetic studies may lead to improved gene annotations even for organisms which were not originally the main subject of investigation. Some of the other enzymes in the ergosterol pathway bear a close resemblance to yeast proteins while others are most similar to human or plant proteins. This intricate patchwork structure raises new questions about the evolutionary origin of the sterol biosynthesis pathways in green algae. We speculate that key enzymes have diverged early to form a distinct subgroup while others present in more specialized pathways have diverged later during evolution. More detailed phylogenetic studies will be necessary to further elucidate the evolutionary history of sterol synthesis.

The detailed investigations of the predictions of our algorithm for the target metabolites L-rhamnose and hydroxyproline revealed a fundamental difficulty in defining consistent large-scale metabolic network models. Both metabolites belong to a class of substances which are not directly produced, their free forms rather result from macromolecule assembly and subsequent breakdown. This indirect mode of production raises the question where the limits of metabolism should be defined. A detailed description of the actual synthesis pathways requires the consideration of polysaccharides and proteins. However, macromolecules are usually not considered as metabolites but as higher level structures composed of a limited number of different building blocks which are provided by metabolism, as depicted in Fig. 8. To resolve this conflict, we propose a macromolecule-free description by considering the overall reaction. For the incorporation of UDP-rhamnose into polysaccharides and the subsequent release of rhamnose, for example, the net conversion can be summarized as

$$\text{UDP-L-rhamnose} \rightarrow \text{UDP} + \text{L-rhamnose}. \quad (11)$$

Such a simplified description of the production of free rhamnose has the advantage that inclusion of macromolecules as metabolites is avoided, but meaningful calculations are unproblematic. The flux through this reaction, for example, is simply interpreted as the net turnover of rhamnose residues.

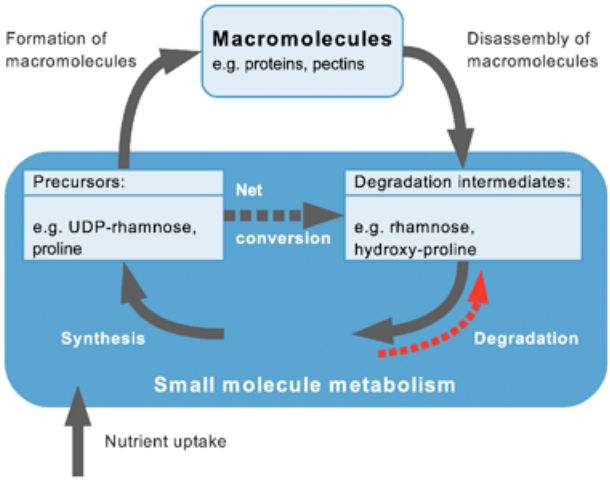

Fig. 8 *Limits of metabolism*. This schematic view illustrates that some metabolites may only be produced in conjunction with macromolecules; their free forms result through macromolecule degradation. Since this route is not included in the description of small molecule metabolism, for such metabolites the reverse degradation pathway is predicted (dashed red arrow). For a consistent description of a metabolic network including the production of such metabolites, we suggest incorporation of overall reactions describing the net conversion (dashed gray arrow).

The calculations presented in this paper have been performed on networks which have been directly retrieved from databases. By this choice we could demonstrate that our method is robust against stoichiometric inaccuracies which are found in many biochemical databases.[47,48] This robustness results from the fact that the underlying test whether a network may perform a given function was carried out using the method of network expansion (see Supporting Text S4, ESI‡). This forward approach to identify producible metabolites is stricter than the condition that there exists a flux distribution for its production and it has been shown[49] that the identified metabolites are even producible if a constant increase of cellular volume is considered, which is not necessarily the case if a producing flux distribution exists, especially if intermediates from a conserved pool of metabolites are involved. While it is in principle straight-forward to replace this analytic method with the mathematically more stringent theory of flux balance analysis (FBA), the latter is very sensitive to stoichiometric imbalances. To demonstrate this, we have systematically compared our predictions for the artificially created draft networks of *E. coli* with predictions obtained from solving constrained optimization problems. For this, we implemented an algorithm that closely resembles those used for GapFill[16] and in ref. 18 (for details, see Supporting Text S4, ESI‡). We found that in general, the FBA based approach obtained smaller solutions which were subsets of extensions found by our approach. This is expected, since the former approach is designed to find a solution minimal with respect to the number of contained reactions. However, often the constrained optimization failed to identify reactions that are necessary, because due to inaccuracies in the stoichiometries in some reactions, some metabolites are wrongly considered as producible. While the FBA based approach tends to overestimate the set of producible metabolites, our approach, which relies on the method of network expansion, tends to underestimate this set and therefore imposes stricter conditions on the network functions. As a consequence, our approach is more robust against stoichiometric imbalances (see Supporting Text S4, ESI‡). Motivated by the systematic comparison of the two methods performed in ref. 49, we expect for the ideal case, in which all imbalances of reactions contained in metabolic databases have been corrected, that the results obtained by an FBA based approach would be almost identical. In such a case, the main advantage of using the method of network expansion is simply a considerably smaller computational effort. An expected advantage of an FBA based approach is that besides the simple existence of possible production routes also the distribution of fluxes may be assessed. This further expands the applicability of our integrative approach by the incorporation of flux data for validation of the computational predictions.

Graph-based approaches and local pathway searches can only provide a rough approximation of our results. The KEGG tool PathComp[50] only considers binary relationships between substrate-product pairs to calculate connecting pathways. Thus, the resulting pathways do not necessarily provide possible production routes since other substrates may be required. The pathway hole filler within in the PathoLogic program provided by the

Pathway Tools software suite[15] requires that for each draft network a new Cyc-database is created and is therefore impractical for a systematic comparison. Moreover, it is restricted to the identification of reactions annotated in predefined pathways. We found that approximately 25% of the reactions identified by our method are not associated with any MetaCyc pathway and thus would not have been found by PathoLogic (see Supporting Text S3, ESI‡).

While we could demonstrate that our method yields plausible and testable predictions regarding missing reactions, it cannot identify reactions that are wrongly included in the network models and that should be excluded. This, however, is a more difficult problem. Even for enzymes for which encoding genes have been identified with high confidentiality, it is possible that they are only expressed under rare conditions and it is therefore difficult to assess whether they should be retained or excluded. Theoretically, unnecessary reactions could be identified by cleaving all those reactions from the draft network for which either no clear sequence homologies are found or which cannot carry a flux under any known external growth condition. Subsequently, our algorithm can be applied to find minimal extensions. The resulting networks will then only contain those reactions which are absolutely required to explain the presence of the observed metabolites. This approach is suitable if one aims at finding a consistent network containing only reactions with a high confidentiality.

The predicted extensions draw their reactions from reported enzymatic reactions stored in databases. This limits the predictions to previously characterized reactions and pathways. While this limitation is unproblematic as long as one aims at completing pathways included in primary metabolism, for which our understanding is far advanced, it will become necessary to include completely novel reactions when aiming at predicting synthesis routes for secondary metabolites, for which experimental knowledge is still sparse. A challenge for future research will be to expand our strategy to construct reactions that are chemically feasible but not yet described. A possible approach is based on reaction patterns defined by the EC nomenclature system or the generalized reaction patterns as defined in ref. 51. The difficulties to be expected include the fact that the number of possible reactions from which the extensions are calculated will be tremendous. We estimate that such an approach can still be successful when focussing the search to a restricted aspect of metabolism, for example a well defined group of structurally similar secondary metabolites.

Our work has direct implications for experimental as well as theoretical researchers. For theoreticians, the proposal of consistent network models is of considerable value. Even though predicted extensions may be incorrect, the resulting networks are at least consistent with experimental observations and hence they are far more suited for subsequent quantitative analyses than the draft networks lacking observed metabolic functions. For experimentalists, the proposal of missing pathway elements and putative genes coding for the missing enzymes can facilitate the design of new experiments aiming at proving the existence of certain metabolic pathways.

Methods

Data preparation

To apply the extension algorithm, we have constructed two organism-specific metabolic networks for *E. coli* and *Chlamydomonas reinhardtii*, and one reference network containing enzymatic reactions from a large number of organisms. These networks were obtained by parsing the flat files of the EcoCyc,[26] ChlamyCyc[30] and MetaCyc[24] databases. EcoCyc and MetaCyc were both version 12 as released on April 1st, 2008. The ChlamyCyc pathway database for Chlamydomonas version 1.0 (June 2008) featured 272 pathways with their annotated genes, enzymes, and compounds. ChlamyCyc was assembled based on the recently published genome sequence[19] and MapMan[52] annotations of Chlamydomonas genes using the Pathway Tools software[53] within the BioCyc family of databases. The predicted pathways and reactions were verified by using orthology information from fifteen other species as well as manual curation.

For the organism-specific networks, thermodynamic constraints were considered if a reaction was given as irreversible in the corresponding database. In such a case, only the physiological direction was included. For the reference network, from which reactions are recruited to calculate the extensions, we have explicitly included every reaction in both directions. This means that extensions may include reactions operating in a non-physiological direction. We deliberately chose not to exclude this possibility, since in principle all reactions may be reversed and the information found in databases about the directionality of reactions is often incorrect. Moreover, much about the network architecture can be learned from the cases in which only extensions are found including reactions operating in their ‘wrong’ directions. This became clear when discussing rhamnose and hydroxyproline metabolism.

Many reactions involve compound classes, such as ‘a hexose’, describing a large set of metabolites. Since it is difficult to automatically determine substrate and product pairs from their corresponding classes, we have conducted this only for the particularly simple classes NAD(P)H and $NAD(P)^+$, which both contain only two metabolites. All other classes were taken into our network as if they were ordinary compounds.

In the databases, some metabolites contain chemical formulas with unspecified residues (denoted by a chemical ‘element’ *R*). Reactions involving such reactants were ignored. Similarly, reactions for which the sum formulas on the left and the right side clearly disagree, have been disregarded.

While embedding the EcoCyc and ChlamyCyc networks into the reference network, we observed the following rules. Reactions with the same ID in the organism-specific and the reference network were considered only once. In the rare cases where these reactions described different stoichiometries, the reaction from the

organism-specific database was used. Also if two or more reactions described exactly the same chemical conversions, only one of them has been considered.

The resulting metabolic networks consist of more than 16 000 reactions and 5000 compounds.

Species-independent reaction models and *E*-values

To determine similarity scores for the annotated protein sequences from *E. coli* and Chlamydomonas we utilized the Pfam[54] and Uniprot[55] databases by using protein domain information, EC numbers, multiple sequence alignments and profile hidden Markov models (HMMs) (Fig. 2). First, we extracted all reactions from MetaCyc with at least one annotated protein sequence. If for such a reaction the Pfam domain family was annotated we used the multiple seed alignment and the according profile HMM as given by Pfam for further computations. If no Pfam family was annotated but a 4-digit EC number for the reaction was given, we downloaded all protein sequences from Uniprot annotated with the according EC number. Otherwise, we used standard BLAST[56] with a score cut-off of 50 and an *E*-value threshold of $1e^{-10}$ to collect significant sequence hits. The *E*-value is the expectation value representing the number of different alignments with the same or better total score, that could be expected to occur within the database purely by chance. The lower the *E*-value, the more significant the score. Afterwards, the reaction-specific protein sequences were aligned using the multiple alignment program MAFFT.[57] The resulting multiple sequence alignment was then converted into a reaction-specific profile HMM using the HMMER software.[25] HMMER turns a multiple-sequence alignment into a probability based position-specific scoring system. Using such calibrated reaction-specific profile HMMs we searched the protein sequences from *E coli* and Chlamydomonas with an HMMER *E*-value cut-off of 1.

If several database entries corresponded to one reaction stoichiometry, for each an *E*-value was obtained and the reaction was associated with the minimum of these values. For reactions with an explicit direction stated in the database, the *E*-value was associated only with the physiologically observed direction. In this way, reactions operating in the physiological direction are included in the extensions with a higher probability.

Extensions of the metabolic network

The extension algorithm (see Protocol 1) is designed in a similar fashion as the algorithm used in ref. 58, where we identified minimal sets of required nutrients. Here we modify the algorithm to determine a minimal extension E to a metabolic draft network R_D, enabling it to produce the target compounds T from the nutrient compounds S (also called seed). Candidate reactions for the extension are taken from an ordered set of reactions R_O. As only a fraction of the reactions R_O are used to extend the network, the sequential order of R_O is crucial. Thus, different orderings are used to produce a large number of different extensions. R_O ideally consists of many known reactions found in other metabolic networks.

Protocol 1: minimalExtension

Input: R_D: set of reactions r belonging to the draft network
R_O: ordered set of reactions r with $R_D \cap R_O = \emptyset$
S: set of compounds belonging to the seed
T: set of target compounds
Result: E: set of reactions r representing the minimal extension
Data: R_E: set of reactions r, the extended network
$R_E \leftarrow R_D \cup R_O$;
$E \leftarrow \emptyset$;
for $i \leftarrow 1 \ldots \|R_O\|$ **do**
  $r \leftarrow R_O[i]$;
  $R_E \leftarrow R_E \setminus r$;
  if not targetsProducible(R_E, S, T) **then**
    $R_E \leftarrow R_E \cup r$;
    $E \leftarrow E \cup r$;
  end
end
return E

First, we start with the fully extended network ($R_D \cup R_O$) that is a union of the draft network and the network derived from the database, ensuring the producibility of all target compounds T. Then the reactions originating from the database (set R_O) are removed one by one in a strict order. This is why the order of the list is important and different orderings may result in different extensions. It is then tested whether still all target metabolites T are producible. If this is the case, the reaction is permanently removed and will not be an element of the minimal extension E. If not all targets T are producible after removal of this reaction, it was apparently essential for some production route. It is therefore returned to the network and belongs to the minimal extension E. The resulting set of reactions is minimal in the sense that the removal of any reaction would make at least one target metabolite unreachable by the remaining network.

The test of producibility, denoted as *targetsProducible*, was implemented applying the method of network expansion.[21] This method depends on heuristics regarding the choice of cofactors, an issue which is discussed in detail in ref. 22. A summary of this discussion and the complete list of used cofactors is given in the Supporting

Text S5, ESI.‡

Two methods are applied to create different sequential orderings R_O and hence different minimal extensions. In the first approach we randomly shuffle the list. Doing this, all reactions have the same probability to end up at a certain position in the list. The second approach of randomization favors reactions according to a similarity score associated with each reaction. This score is defined as the logarithm of the *E*-value, which has been determined as described above. Since reactions situated at the end of the sequentially ordered set R_O have a higher probability to be included in the extension, we sort the list such that reactions with a low score are preferentially positioned near the end (see Protocol 2). For that, each element of the ordered set is exchanged with another randomly selected element with the probability $e^{-\beta\Delta S}$, where ΔS is the difference between the scores of the corresponding reactions and β an adjustable parameter quantifying the randomness of the mixing. A value of $\beta = \infty$ means complete ordering (no randomness at all), while a value of $\beta = 0$ corresponds to the random shuffling of the first approach. The functions *randomFloat* and *randomInt* return uniformly distributed random numbers (floating point and integer respectively) in the given interval. The function *exchangeElements* swaps the position of the two elements in the given ordered set.

Protocol 2: shuffleListByScore

Input: R_O: ordered set of reactions r
β: parameter that determines the amount of randomness
Result: R_{NO}: new ordered set of reactions r
Data: i: index
i_r: random index
dS: delta score
$R_{NO} \leftarrow R_O$;
for $i \leftarrow 1 \ldots \|R_{NO}\|$ **do**
  $i_r \leftarrow$ `randomInt`$(1, \|R_{NO}\|)$;
  $dS \leftarrow$ `score`$(R_{NO}[i])$ − `score`$(R_{NO}[i_r])$;
  if $i > i_r$ **then**
    $dS \leftarrow -dS$;
  end
  if $dS < 0$ **then**
    `exchangeElements`(R_{NO}, i, i_r);
  else if `randomFloat`$(0, 1) < e^{-\beta\ dS}$ **then**
    `exchangeElements`(R_{NO}, i, i_r);
  end
end
return R_{NO};

To achieve a proper mixing according to the chosen β, the function *shuffleListByScore* is called 100 times before calculating the first extension (this is the so-called *thermalization*).

Comparative sequence analysis

Orthologs have been identified by using the Inparanoid[59] software tool. The phylogenies were built as follows: we downloaded all available protein sequences for a given reaction from Uniprot.[55] Chlamydomonas genes were then aligned together with the reaction-specific sequences using the MAFFT[57] multiple alignment webserver. Phylogenetic trees were then constructed using the Neighbor-Joining algorithm.

Acknowledgements

The authors thank Alexander Skupin for critically reading the manuscript and for his valuable suggestions. This work was funded by the German Research Foundation, in particular the International Research Training Group "Genomics and Systems Biology of Molecular Networks" (N. Christian) and the Collaborative Research Center "Theoretical Biology: Robustness, Modularity and Evolutionary Design of Living Systems" (T. Handorf), as well as by the German Federal Ministry of Education and Research, Systems Biology Research Initiative "*Go*FORSYS" (P. May, O. Ebenhöh, S. Kempa).

References

1 K. Liolios, K. Mavromatis, N. Tavernarakis and N. C. Kyrpides, *Nucleic Acids Res.*, 2007, **36**, D475–D479 External Links.

2 G. Ramsay, *Nat. Biotechnol.*, 1998, **16**, 40–44 External Links.

3 H. Goda, E. Sasaki, K. Akiyama, A. Maruyama-Nakashita, K. Nakabayashi, W. Li, M. Ogawa, Y. Yamauchi, J. Preston, K. Aoki, T. Kiba, S. Takatsuto, S. Fujioka, T. Asami, T. Nakano, H. Kato, T. Mizuno, H. Sakakibara, S. Yamaguchi, E. Nambara, Y. Kamiya, H. Takahashi, M. Y. Hirai, T. Sakurai, K. Shinozaki, K. Saito, S. Yoshida and Y. Shimada, *Plant J.*, 2008, **55**, 526–542 External Links.

4 P. Jones, R. G. Ct, L. Martens, A. F. Quinn, C. F. Taylor, W. Derache, H. Hermjakob and R. Apweiler, *Nucleic Acids Res.*, 2006, **34**, D659–D663 External Links.

5 J. Hummel, M. Niemann, S. Wienkoop, W. Schulze, D. Steinhauser, J. Selbig, D. Walther and W. Weckwerth, *BMC Bioinformatics*, 2007, **8**, 216 External Links.

6 K. Hollywood, D. R. Brison and R. Goodacre, *Proteomics*, 2006, **6**, 4716–4723 External Links.

7 H. Kitano, *Science*, 2002, **295**, 1662–1664 External Links.

8 R. U. Ibarra, J. S. Edwards and B. O. Palsson, *Nature*, 2002, **420**, 186–189 External Links.

9 J. S. Edwards and B. O. Palsson, *BMC Bioinformatics*, 2000, **1**, 1 External Links.

10 K. J. Kauffman, P. Prakash and J. S. Edwards, *Curr. Opin. Biotechnol.*, 2003, **14**, 491–496 External Links.

11 M. J. Herrgård, N. Swainston, P. Dobson, W. B. Dunn, K. Y. Arga, M. Arvas, N. Blthgen, S. Borger, R. Costenoble, M. Heinemann, M. Hucka, N. L. Novre, P. Li, W. Liebermeister, M. L. Mo, A. P. Oliveira, D. Petranovic, S. Pettifer, E. Simeonidis, K. Smallbone, I. Spasi, D. Weichart, R. Brent, D. S. Broomhead, H. V. Westerhoff, B. Kirdar, M. Penttil, E. Klipp, B. O. Palsson, U. Sauer, S. G. Oliver, P. Mendes, J. Nielsen and D. B. Kell, *Nat. Biotechnol.*, 2008, **26**, 1155–1160 External Links.

12 Z. Nikoloski, S. Grimbs, P. May and J. Selbig, *J. Theor. Biol.*, 2009.

13 A. Osterman and R. Overbeek, *Curr. Opin. Chem. Biol.*, 2003, **7**, 238–251 External Links.

14 S. M. Paley and P. D. Karp, *Bioinformatics*, 2002, **18**, 715–724 External Links.

15 M. L. Green and P. D. Karp, *BMC Bioinformatics*, 2004, **5**, 76 External Links.

16 V. Satish Kumar, M. Dasika and C. Maranas, *BMC Bioinformatics*, 2007, **8**, 212 External Links.

17 M. DeJongh, K. Formsma, P. Boillot, J. Gould, M. Rycenga and A. Best, *BMC Bioinformatics*, 2007, **8**, 139 External Links.

18 J. L. Reed, T. R. Patel, K. H. Chen, A. R. Joyce, M. K. Applebee, C. D. Herring, O. T. Bui, E. M. Knight, S. S. Fong and B. O. Palsson, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 17480–17484 External Links.

19 S. S. Merchant, S. E. Prochnik, O. Vallon, E. H. Harris, S. J. Karpowicz, G. B. Witman, A. Terry, A. Salamov, L. K. Fritz-Laylin, L. Marchal-Drouard, W. F. Marshall, L.-H. Qu, D. R. Nelson, A. A. Sanderfoot, M. H. Spalding, V. V. Kapitonov, Q. Ren, P. Ferris, E. Lindquist, H. Shapiro, S. M. Lucas, J. Grimwood, J. Schmutz, P. Cardol, H. Cerutti, G. Chanfreau, C.-L. Chen, V. Cognat, M. T. Croft, R. Dent, S. Dutcher, E. Fernndez, H. Fukuzawa, D. Gonzlez-Ballester, D. Gonzlez-Halphen, A. Hallmann, M. Hanikenne, M. Hippler, W. Inwood, K. Jabbari, M. Kalanon, R. Kuras, P. A. Lefebvre, S. D. Lemaire, A. V. Lobanov, M. Lohr, A. Manuell, I. Meier, L. Mets, M. Mittag, T. Mittelmeier, J. V. Moroney, J. Moseley, C. Napoli, A. M. Nedelcu, K. Niyogi, S. V. Novoselov, I. T. Paulsen, G. Pazour, S. Purton, J.-P. Ral, D. M. Riao-Pachn, W. Riekhof, L. Rymarquis, M. Schroda, D. Stern, J. Umen, R. Willows, N. Wilson, S. L. Zimmer, J. Allmer, J. Balk, K. Bisova, C.-J. Chen, M. Elias, K. Gendler, C. Hauser, M. R. Lamb, H. Ledford, J. C. Long, J. Minagawa, M. D. Page, J. Pan, W. Pootakham, S. Roje, A. Rose, E. Stahlberg, A. M. Terauchi, P. Yang, S. Ball, C. Bowler, C. L. Dieckmann, V. N. Gladyshev, P. Green, R. Jorgensen, S. Mayfield, B. Mueller-Roeber, S. Rajamani, R. T. Sayre, P. Brokstein, I. Dubchak, D. Goodstein, L. Hornick, Y. W. Huang, J. Jhaveri, Y. Luo, D. Martnez, W. C. A. Ngau, B. Otillar, A. Poliakov, A. Porter, L. Szajkowski, G. Werner, K. Zhou, I. V. Grigoriev, D. S. Rokhsar and A. R. Grossman, *Science*, 2007, **318**, 245–250 External Links.

20 O. Ebenhöh, T. Handorf and R. Heinrich, *Genome Informatics*, 2004, **15**, 35–45.

21 T. Handorf, O. Ebenhöh and R. Heinrich, *J. Mol. Evol.*, 2005, **61**, 498–512 External Links.

22 T. Handorf and O. Ebenhöh, *Nucleic Acids Res.*, 2007, **35**, W613–W618 External Links.

23 M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu and Y. Yamanishi, *Nucleic Acids Res.*, 2007, **36**, D480–D484 External Links.

24 R. Caspi, H. Foerster, C. A. Fulcher, P. Kaipa, M. Krummenacker, M. Latendresse, S. Paley, S. Y. Rhee, A. G. Shearer, C. Tissier, T. C. Walk, P. Zhang and P. D. Karp, *Nucleic Acids Res.*, 2007, **36**, D623–D631 External Links.

25 S. R. Eddy, *Bioinformatics*, 1998, **14**, 755–763 External Links.

26 P. D. Karp, I. M. Keseler, A. Shearer, M. Latendresse, M. Krummenacker, S. M. Paley, I. Paulsen, J. Collado-Vides, S. Gama-Castro, M. Peralta-Gil, A. Santos-Zavaleta, M. I. Pealoza-Spnola, C. Bonavides-Martinez and J. Ingraham, *Nucleic Acids Res.*, 2007, **35**, 7577–7590 External Links.

27 J. L. Reed, T. D. Vo, C. H. Schilling and B. O. Palsson, *Genome Biol.*, 2003, **4**, R54 External Links.

28 A. M. Feist, C. S. Henry, J. L. Reed, M. Krummenacker, A. R. Joyce, P. D. Karp, L. J. Broadbelt, V. Hatzimanikatis and B. O. Palsson, *Mol. Syst. Biol.*, 2007, **3**, 121.

29 J. Leitch and P. Collier, *Lett. Appl. Microbiol.*, 1996, **22**, 18–20 External Links.

30 P. May, J.-O. Christian, S. Kempa and D. Walther, *BMC Genomics*, 2009, **10**, 209 External Links.

31 P. May, S. Wienkoop, S. Kempa, B. Usadel, N. Christian, J. Rupprecht, J. Weiss, L. Recuenco-Munoz, O. Ebenhöh, W. Weckwerth and D. Walther, *Genetics*, 2008, **179**, 157–166 External Links.

32 C. Bölling and O. Fiehn, *Plant Physiol.*, 2005, **139**, 1995–2005 External Links.

33 C. Iomini, L. Li, W. Mo, S. K. Dutcher and G. Piperno, *Curr. Biol.*, 2006, **16**, 1147–1153 External Links.

34 E. Zablackis, J. Huang, B. Mller, A. G. Darvill and P. Albersheim, *Plant Physiol.*, 1995, **107**, 1129–1138 External Links.

35 B. G. Moore and R. G. Tischer, *Science*, 1964, **145**, 586–587 External Links.

36 F. W. Kotch, I. A. Guzei and R. T. Raines, *J. Am. Chem. Soc.*, 2008, **130**, 2952–2953 External Links.

37 D. T. Lamport, *J. Biol. Chem.*, 1963, **238**, 1438–1440 External Links.

38 W. S. Adair and K. E. Apt, *Proc. Natl. Acad. Sci. U. S. A.*, 1990, **87**, 7355–7359 External Links.

39 H. Wu, B. de Graaf, C. Mariani and A. Y. Cheung, *Cell. Mol. Life Sci.*, 2001, **58**, 1418–1429 External Links.

40 E. Adams and L. Frank, *Annu. Rev. Biochem.*, 1980, **49**, 1005–1061 External Links.

41 H. Takahashi, C. E. Braby and A. R. Grossman, *Plant Physiol.*, 2001, **127**, 665–673 External Links.

42 P. L. Soto-Yarritu, L. Amigo, G. Taborda, I. Martnez-Castro and J. A. Gmez-Ruiz, *J. Dairy Sci.*, 2007, **90**, 5001–5003 External Links.

43 S. Rajamani, W. D. Bauer, J. B. Robinson, J. M. Farrow, E. C. Pesci, M. Teplitski, M. Gao, R. T. Sayre and D. A. Phillips, *Mol. Plant-Microbe Interact.*, 2008, **21**, 1184–1192 External Links.

44 J. Barthelmes, C. Ebeling, A. Chang, I. Schomburg and D. Schomburg, *Nucleic Acids Res.*, 2007, **35**, D511–D514 External Links.

45 R. V. Krishna, P. R. Krishnaswamy and D. R. Rao, *Biochem. J.*, 1971, **124**, 905–913 External Links.

46 N. Shaw, P. F. Smith and H. M. Verheij, *Biochem. J.*, 1972, **129**, 167–173 External Links.

47 M. G. Poolman, B. K. Bonde, A. Gevorgyan, H. H. Patel and D. A. Fell, *Syst. Biol.*, 2006, **153**, 379–384 External Links.

48 A. Gevorgyan, M. G. Poolman and D. A. Fell, *Bioinformatics*, 2008, **24**, 2245–2251 External Links.

49 K. Kruse and O. Ebenhöh, *Genome Informatics Ser.*, 2008, **20**, 91–101.

50 K. F. Aoki-Kinoshita, *J. Pestic. Sci.*, 2006, **31**, 296–299 External Links.

51 Y. Shimizu, M. Hattori, S. Goto and M. Kanehisa, *Genome Informatics Ser.*, 2008, **20**, 149–158.

52 O. Thimm, O. Bläsing, Y. Gibon, A. Nagel, S. Meyer, P. Krger, J. Selbig, L. A. Mller, S. Y. Rhee and M. Stitt, *Plant J.*, 2004, **37**, 914–939 External Links.

53 P. D. Karp, S. Paley and P. Romero, *Bioinformatics*, 2002, **18**(Suppl 1), S225–S232.

54 E. L. Sonnhammer, S. R. Eddy, E. Birney, A. Bateman and R. Durbin, *Nucleic Acids Res.*, 1998, **26**, 320–322 External Links.

55 UniProt Consortium, *Nucleic Acids Res.*, 2007, **35**, D193–D197 External Links.

56 S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman, *J. Mol. Biol.*, 1990, **215**, 403–410 External Links.

57 K. Katoh, K. Kuma, H. Toh and T. Miyata, *Nucleic Acids Res.*, 2005, **33**, 511–518 External Links.

58 T. Handorf, N. Christian, O. Ebenhöh and D. Kahn, *J. Theor. Biol.*, 2008, **252**, 530–537 External Links.

59 M. Remm, C. E. Storm and E. L. Sonnhammer, *J. Mol. Biol.*, 2001, **314**, 1041–1052 External Links.

Footnotes

† This article is part of a *Molecular BioSystems* themed issue on Computational and Systems Biology.

‡ Electronic supplementary information (ESI) available: S1: The shuffling parameter β and ensemble size. S2: Artificially reduced *B. subtilis* draft networks. S3: Known pathways in extensions. S4: Comparison to FBA based methods. S5: Cofactor metabolites. S6: Nutrients and target metabolites. S7: Predicted ergosterol pathway. S8: Phylogenies for putative Chlamydomonas enzymes of the predicted ergosterol pathway. See DOI: 10.1039/b915913b