Analysis of Metabolic Evolution in Bacteria Using Whole-Genome Metabolic Models

Ali A. Faruqi, William A. Bryant, and John W. Pinney*

Imperial College, South Kensington Campus
London, SW7 2AZ
j.pinney@imperial.ac.uk
http://www.theosysbio.bio.ic.ac.uk/bacterial-metabolism/

Abstract. Recent advances in the automation of metabolic model reconstruction have led to the availability of draft-quality metabolic models (predicted reaction complements) for multiple bacterial species. These reaction complements can be considered as trait representations and can be used for ancestral state reconstruction, to infer the most likely metabolic complements of common ancestors of all bacteria with generated metabolic models. We present here an ancestral state reconstruction for 141 extant bacteria and analyse the reaction gains and losses for these bacteria with respect to their lifestyles and pathogenic nature. A simulated annealing approach is used to look at coordinated metabolic gains and losses in two bacteria. The main losses of *Onion yellows phytoplasma* OY-M, an obligate intracellular pathogen, are shown (as expected) to be in cell wall biosynthesis. The metabolic gains made by *Clostridium difficile* CD196 in adapting to its current habitat in the human colon is also analysed. Our analysis shows that the capability to utilize N-Acetylneuraminic acid as a carbon source has been gained, rather than having been present in the *Clostridium* ancestor, as has the capability to synthesise phthiocerol dimycocerosate which could potentially aid the evasion of the host immune response. We have shown that the availability of large numbers of metabolic models, along with conventional approaches, has enabled a systematic method to analyse metabolic evolution in the bacterial domain.

Keywords: Metabolic Evolution, Ancestral State Reconstruction, Metabolic Models, Hierarchical Clustering, Simulated Annealing, Pathogenicity.

1 Introduction

One of the aims of systems biology has been to integrate information regarding metabolism in order to construct metabolic models and thus to analyse the effects of genetic perturbations on metabolism at the system level. In recent years, a number of attempts have been made to study the evolution of metabolic networks

* Corresponding author.

M. Deng et al. (Eds.): RECOMB 2013, LNBI 7821, pp. 45–57, 2013.

and these have provided insights into the mechanisms of evolution of various extant bacteria [1–3]. Understanding the evolution of bacterial metabolism is of great importance for a number of reasons. In particular, it has the potential to provide insights into the evolution of pathogenicity and its relationship with metabolism.

Bacteria not only evolve through vertical inheritance, but also through horizontal gene transfer (HGT). Often HGT can provide metabolic genes [4, 5], and potentially antibiotic resistance and toxin encoding genes [6] to bacteria. On the other hand evolution through gene loss can occur in some environments [7]. These processes directly involve gene losses and gains, but it is not the genes themselves that are of most interest, but their function and how they interrelate with the functions of all other genes in the system.

Evolution is often studied through Ancestral State Reconstruction (ASR) for various biological traits [8, 9]. ASR relies on biological trait information from extant organisms to infer trait occurrence in the common ancestors of those organisms. This information can be provided in the form of a character matrix for the characteristics under investigation. Depending on the context, a parsimony or maximum likelihood approach can be used on a phylogenetic tree to obtain the probabilities of different ancestral nodes possessing the considered traits. This approach has been taken in looking at gene families in the metabolic context [10], and metabolic reaction occurrences have been compared according to inferred metabolic models for a small set of 16 *E. coli* strains to investigate the evolution of these strains [11].

Previously genomic comparisons have been done using information from the WIT database, examining differences between the metabolic pathway complements of various extant organisms [12]. Additionally phylogenetic profiles have been inferred based on enzyme evolutionary predictions [13] to establish the ancestral relationships between a large number of prokaryotes and eukaryotes.

With the advent of automatic methods for bacterial metabolic model reconstruction – such as the Model SEED pipeline [14] – it is possible for the first time to establish direct reaction complements for any bacterium for which there is a complete genome sequence. Data from these draft-quality automatically generated metabolic models can be used as the input to ASR, since these models make direct assertions about which reactions are present and absent in each bacterium. Consequently, it is possible to infer ancestral metabolic complements directly and to investigate the precise metabolic changes accumulated by bacteria in the evolution towards their current lifestyles and ecological niches at the system level. This improves on previous approaches by being reaction-specific, rather than at a pathway level. Also, information about specific reactions can be made based on enhanced inferences (achieved through the Model SEED pipeline) about reaction presence and absence, not just based on direct observation of annotated enzymes.

Here we present an ancestral state reconstruction of the metabolic reactions inferred to be present in 141 bacteria by the Model SEED server. A hierarchical clustering was used to establish the metabolic similarity of these 141 bacteria,

and this was compared to the 23S rRNA phylogenetic tree inferred for these same bacteria. Further we related this clustering to the lifestyles of the bacteria according to three categorisations: habitat, respiratory mode and pathogenic mode and showed that each of these categorisations encapsulates information about how the evolution of these bacteria has proceeded.

Results for the gain and loss of reactions for each of the extant bacteria have been produced using the metabolic networks inferred for the common ancestors of these bacteria and for two cases these gains and losses have been investigated at the system level to look for coordinated sets of reactions (those reactions adjacent in the metabolic networks inferred from their respective metabolic models) that have been lost (in the case of an obligate intracellular pathogen) and gained (in the case of a free-living pathogen). This has been achieved by using an approach based on active modules [15] called `ambient` which finds connected subnetworks in the bipartite network of reactions and metabolites associated with strong evidence for reaction gain or loss for both these bacteria [Bryant et al. - in submission]. `ambient` has picked out several reaction pathways in *C. difficile* CD196 that would not be seen by gene-based analysis (since several of the reactions have no gene association) but are clearly found by taking advantage of the generated metabolic model used here.

2 Methods

2.1 23S rRNA Phylogeny Construction

23S rRNA sequences for all 141 organisms in the current analysis along with an out-group organism (*Thermoplasma acidophilum*) were obtained from the NCBI Nucleotide Database. Multiple sequence alignment of the 23S rRNA sequences of these organisms was obtained using `MAFFT` [16]. A threshold score of $E = 8.4e{-}11$ was used (the default threshold value used by MAFFT).

Based on the results of multiple sequence alignment, a phylogeny was constructed using `PhyML 2.4.4` [17]. Bootstrapping was performed 100 times on the tree to obtain the most likely phylogeny. After rooting the tree, the out-group was removed. For visualisation of the phylogeny obtained and for the creation of phylogeny images `Dendroscope` was used [18]. The phylogeny can be seen in Supplementary Fig. 1 available at our website[1].

2.2 Comparison of Reaction Numbers and Lifestyle

Three lifestyle classifications were used to assess how they related to the evolutionary histories of the bacteria in this study. The classifications are named i) habitat, indicating the usual environment the bacteria experience, ii) respiratory mode, indicating their ability to tolerate oxygen and iii) pathogenic mode, each bacterium falling into one of four categories: free-living, facultative host-associated, obligate intracellular mutualists and obligate intracellular pathogens.

[1] `http://www.theosysbio.bio.ic.ac.uk/bacterial-metabolism/`

These classifications were taken from work by Zientz et al. [19] and Merhej et al. [10]. It should be noted that although the last classification is termed 'pathogenic mode' this is just an alternative classification of habitat, based on the types of environment experienced by bacteria in their eukaryotic hosts.

A Mann Whitney U test was conducted between each category for each classification to establish correlations between reaction numbers and lifestyles. The Benjamini-Hochberg multiple testing correction was used to control for false positives and the corrected p-values were used to establish significance.

2.3 Ancestral State Reconstruction

For Ancestral State Reconstruction (ASR), `Mesquite` was used [20]. The Ancestral State Reconstruction algorithm, as implemented in Mesquite, looks for ancestral states which maximize the probability of the observed characteristics in extant organisms.

Maximum likelihood reconstruction methods look for ancestor states that maximise the probability of producing the current state, having evolved under a defined model of evolution [8,21]. It is equivalent to the marginal reconstruction method as implemented in PAUP [22]. Every reaction was classified as present or absent according to the Model SEED metabolic model creation server [14]. The Asymmetrical Markov k-state 2 parameter model (AsymmMK model) in `Mesquite` was chosen as it allows different rates for reaction gains and losses. In the ASR, the out-group organism was removed from the phylogeny and the reaction traits for the out-group were not specified in the character matrix.

A boolean character matrix was created for all the 2526 metabolic reactions that were present in at least one of the bacteria under investigation. Maximum Likelihood ASR was performed for this categorical, discrete dataset of reactions. Values for the probability (P_q^j) of the presence of a particular reaction (j) in a particular ancestral organism (q) were calculated by Mesquite based on the AsymmMK model.

2.4 Correlation between Dendrograms and Lifestyle Classifications

Two dendrograms were obtained from the 23S rRNA alignment and the metabolic traits comparison. The `cutree` package in `R` was used to examine every possible clustering of each dendrogram and the maximum Adjusted Rand Index [23] from all possible clusterings was obtained for each dendrogram against the three classifications in this analysis: habitat, respiratory mode and pathogenic mode. Adjusted Rand Index measures the similarity of different partitions of a set; in this case the partitions are the three classifications and the set is all bacteria under consideration.

2.5 Inference of Gains and Losses in Extant Bacteria

Each branch in the phylogeny connects two nodes. One node is the parent (ancestor) node and other node is the child (descendant) node. In order to assess

the gain and loss of reactions from an ancestor to its descendant, δP^j values were calculated according to the following formula:

$$\delta P^j = P^j_{child} - P^j_{parent} \quad (1)$$

where P^j_{parent} and P^j_{child} are the probabilities of the presence of reaction j in the parent (ancestral organism)and child (descendant organism) nodes of a particular branch respectively. Therefore, a δP^j value close to 1 indicates a high likelihood of gain of reaction j in a branch and a δP^j value close to -1 indicates a loss of reaction.

There are a total of 140 internal (parent) nodes in the phylogeny. Each node gives rise to 2 branches giving a total of 280 branches. δP^j values were calculated for all the reactions on all of the branches. Thus, there are a total of 707,280 δP^j values (280 x 2526) for the entire phylogeny.

A δP cutoff of ± 0.9 was used to define those reactions gained or lost. Using this threshold δP value, ancestral state reconstruction predicted a total of 10,396 gain and loss events. $\delta P \leq$ -0.9 (loss) had 5001 events and $\delta P \geq +0.9$ (gain) had 5395 events.

2.6 Metabolic Traits Hierarchical Clustering

The construction of a metabolic trait-based hierarchical clustering was done using the `Pars` programme in the `PHYLIP` package [24]. Each reaction present in at least one, but less than 141 of the bacteria under investigation, was used as a metabolic trait, as for the ASR. The `Pars` programme produced a total of 12 trees, from which a consensus tree was obtained using the `CONSENSE` program in the `PHYLIP` package.

2.7 Analysis of Coordinated Metabolic Changes

`ambient` [Bryant et al. - in submission] was used to run simulated annealing on the bipartite network of reactions and metabolites to find the 100 most significant coordinated metabolic changes in two bacteria representing the obligate intracellular (*Onion yellows phytoplasma* OY-M) and free-living (*Clostridium difficile* CD196) pathogenic lifestyles adopted by many of the bacteria investigated here.

The metabolic network used for both bacteria was the complete 'meta-' metabolic network consisting of the union of all 141 networks used in this paper. This allowed both gains and losses to be seen for each bacterium. `ambient` uses scores for each reaction and metabolite in its attempt to find connected network components encompassing many highly changed reactions. In this case the scores for reactions were taken from δP values for the relevant bacteria. Metabolites were scored in the using the default `ambient` scoring method - with a penalty equal to their connectivity in the metabolic network, to select against currency metabolites.

`ambient` was run to look for coordinated areas of loss of reactions in *Onion yellows phytoplasma* and gain of reactions in *C. difficile*. `ambient` was run with the following non-default parameters: maximum number of steps (-N) was set to 2,500,000, temperature gradient (-U) to 0.95, initial temperature factor (-T) to 3, reaction score offset (-Y) to -0.15 and number of steps between equilibrium tests (-i) to 6000.

3 Results and Discussion

3.1 Distribution of Organism Lifestyles and Reactions

Information about the number and types of organisms and reactions [14] was integrated with data about the lifestyles of those organisms [10,19]. Fig. 1 shows the distribution of the number of reactions in each organism with respect to their lifestyles: habitats, respiratory modes and pathogenic mode. The median number of reactions in the organisms is 1014. The reactions common to all 141 organisms account for about 1% of the total number of reactions.

As can be seen from Fig. 1, most of the organisms that have fewer than 700 reactions are host-associated; indeed from the distribution of pathogenic modes these bacteria represent the vast majority of obligate intracellular symbionts and pathogens. A Mann Whitney U test was conducted to establish whether there was any statistically significant relationship between lifestyle and number of reactions present in each bacterium. Results for each individual test and their p-values corrected for multiple testing can be seen in Supplementary Table 1.

The results show that differing habitats do not necessarily have a large impact on numbers of reactions that the bacteria maintain, except when comparing the free-living bacteria with those which are host-associated. There is also some impact of respiratory mode on number of reactions, but this could be due to a dependence of respiratory mode on bacterial habitat.

The most significant results come from the comparison of the different pathogenic lifestyles of these organisms, as classified by Merhej et al. [10]. Supplementary Table 1C clearly shows, as expected from observations of symbiotic and parasitic bacteria, that the number of reactions available for each bacterium is strongly dependent on their relationship with their eukaryotic host. This is not just true for obligate intracellular bacteria, but also to an extent for host-associated pathogenic bacteria. Unsurprisingly, obligate intracellular mutualists and parasites do not differ significantly in the size of their metabolic network, since their lifestyles, restricted to within a eukaryotic host, mean they experience the same nutrient availability and limitations.

3.2 Ancestral State Reconstruction

Ancestral state reconstruction for each reaction (trait) was performed on the phylogenetic tree inferred from the 23S rRNA alignment. A total of 30 metabolic reactions were present in all the 141 bacteria and these were excluded from the analysis so 2526 reactions were considered.

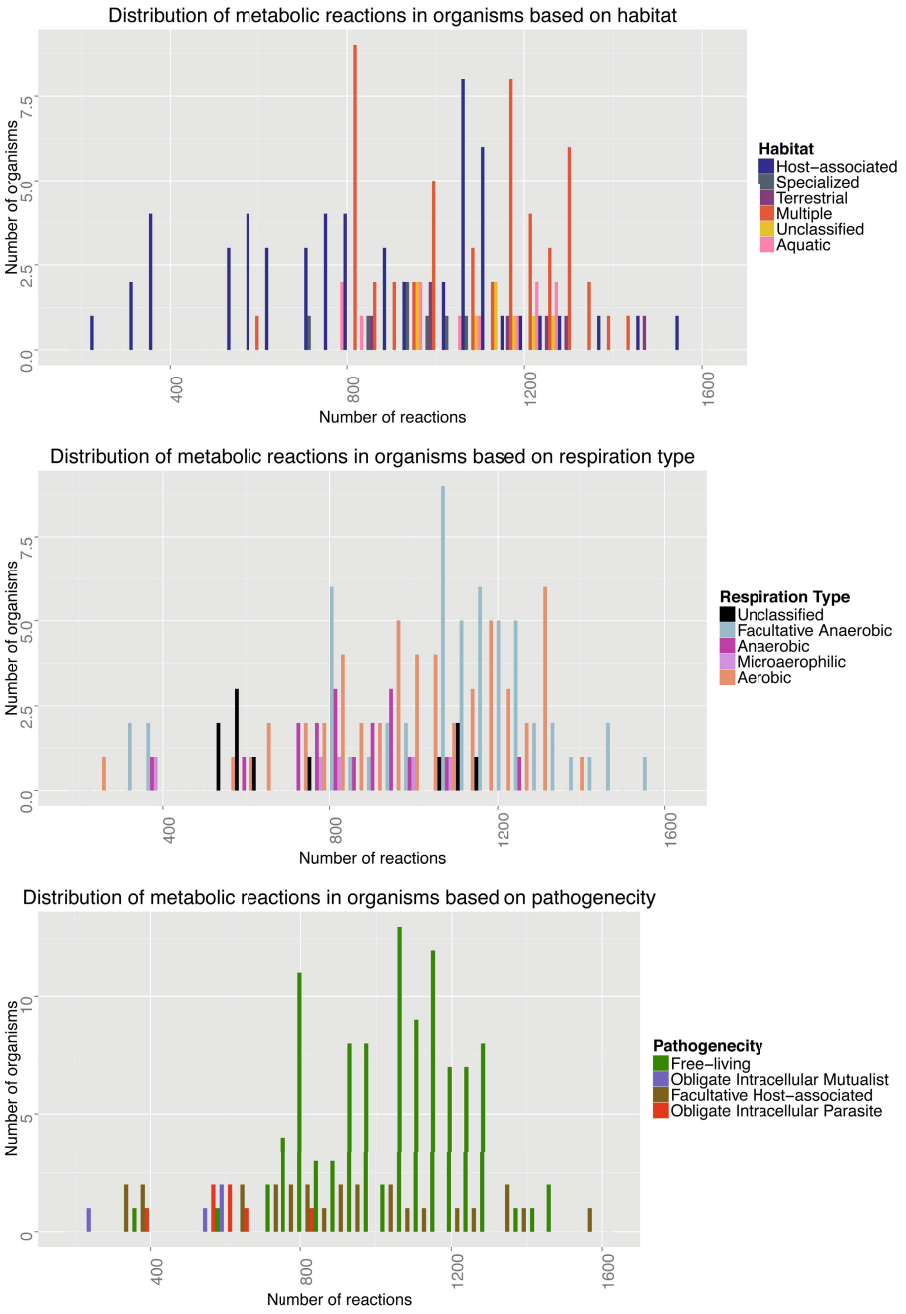

Fig. 1. Histograms showing the relationship between reaction numbers and bacterial lifestyles. Each diagram shows the distribution of total number of reactions in organisms based on habitat type (top), respiratory type (middle) and pathogenicity (bottom) according to Model SEED reconstructions of 141 bacteria.

ASR predicted the presence and absence of every reaction at every ancestral node on the phylogeny. Inferences about the gains and losses of reactions through evolutionary history could be established by using the results of the ASR. For each branch the change in likelihood of the presence of a particular reaction between the parent node and the child node was calculated, called δP (see Methods). A cutoff value for δP of ± 0.9 was used to infer which reactions were most likely gained and lost along each branch of the phylogeny, thus establishing where these metabolic changes occurred in the history of each extant bacterium.

The results obtained from ASR appear to be consistent with our biological knowledge about the different habitats and lifestyles of the bacteria under investigation. Using the aforementioned threshold score, the top five branches that showed the greatest number of gain and loss events were terminal branches leading to various extant bacteria.

The greatest gain was observed in the terminal branch leading to the bacterium *Clostridium difficile* CD196. These metabolic changes could be related to *C. difficile* pathogenicity, and are thus of interest. The gains made by this *C. difficile* strain were analysed by `ambient`, see below, to establish whether these gains occurred in a modular way (adjacent reactions in the metabolism of the bacterium).

In terms of loss, the greatest loss is observed in *Onion yellows phytoplasma* OY-M which is an obligate intracellular plant pathogen and contains an even smaller complement of metabolic genes than *Mycoplasma genitalium* [25]. These losses were analysed by `ambient` to discover whether they are linked together in the metabolic network.

3.3 Metabolic Traits Hierarchical Clustering (MHC) Compared with 23S rRNA Phylogeny

Based on 23S rRNA sequence similarity, many organisms appear closely related to each other on the phylogenetic tree. However, the metabolic data presented here indicate the divergence of these organisms at a metabolic level far greater than that implied by their vertical evolution (genetic inheritance from ancestors) alone. This reflects the knowledge that bacteria evolve metabolically by horizontal gene transfer in addition to vertical evolution.

A hierarchical clustering based on metabolic reaction traits was created to show metabolic relatedness of various extant bacteria. Indeed using clusters of orthologous groups of genes as traits to construct a hierarchical clustering has been shown to cluster bacteria along metabolic lines [10]. This clustering is based on metabolic similarity, so should reflect both vertical evolution (where the bulk of metabolic capabilities are inherited from) and horizontal gene transfer, depending on the importance of each of these mechanisms in the evolution of each organism.

The clustering was constructed using the character matrix of metabolic traits, to gain a better understanding of the evolutionary relationships as revealed through the ASR results presented above. Supplementary Fig. 2 shows the

consensus tree obtained based on the metabolic traits of the organisms. The results obtained here clearly show that even though two organisms may be distantly related based on 23S rRNA sequence similarity, they can be very closely related in terms of their metabolic capabilities, i.e. that they have been subject to convergent evolution. A clear example here is between *Mycoplasma pulmonis* UAB CTIP and *Onion yellows phytoplasma* OY-M. They appear evolutionarily distant on the 23S rRNA phylogeny but are very closely related according to their metabolic trait profiles. Supplementary Figures 1 and 2 show the phylogeny and the metabolic hierarchical clustering respectively.

Dendrograms were produced from the trait-based tree and the RNA-based phylogeny and these were analysed to find whether the clusterings in the dendrograms corresponded to the three lifestyle classifications considered here. Maximum Adjusted Rand Indices (ARIs) were produced for each dendrogram / classification pair to quantify their relatedness. For the RNA-based tree none of the maximum ARIs were greater than 0.1, indicating little or no correlation between vertical evolution and current lifestyle. However, when the metabolic traits (i.e. metabolic reaction complement) and the lifestyles were compared a value of 0.15 was obtained for respiratory mode and a value of 0.37 for pathogenic lifestyle (free-living, host-associated, obligate intracellular mutualists and obligate intracellular pathogens). This indicates that the pathogenic mode adopted by a bacterium has a clear influence on its metabolic network.

3.4 Active Module Analysis

While overall gains and losses of reactions in bacteria are informative in establishing some of the principles of metabolic evolution, the specific changes and how coordinated these changes are might shed more light on the dependence of metabolic evolution on bacterial lifestyles and pathogenicity. Most metabolic processes rely on multiple distinct reactions, therefore on multiple genes encoding those enzymatic functions, so gains and losses of adjacent metabolic functions (pathways) might be expected to occur simultaneously. Here we used `ambient` [Bryant et al. - in submission] to look for reaction gains and losses that form connected components of the metabolic networks of the bacteria under consideration. Two bacteria were analysed, representing two different lifestyles: the obligate intracellular (*Onion yellows phytoplasma* OY-M) and the free-living (*Clostridium difficile* CD196).

The analysis of *C. difficile* produced 14 metabolic modules significant at the $q = 0.001$ level, which can be seen in Supplementary Fig. 3. Table 1 shows a summary of the functions of the modules found. Several modules are involved in monosaccharide utilisation and some in cell wall biosynthesis. Of particular interest is the apparent gain of phthiocerol dimycocerosate biosynthesis capabilities; this lipid has been shown to protect *Mycobacterium tuberculosis* when growing in a mammalian host [26], so could potentially perform the same function for *C. difficile*.

It has been established previously that *C. difficile* CD196 utilises as carbon sources N-Acetyl-glucosamine and N-Acetyl-neuraminic acid, which are both

Table 1. A summary of the metabolic functions gained by *C. difficile* since branching from the rest of the bacteria of the genus *Clostridium* represented in this analysis. Each line is an individual module (connected metabolic component) that has significantly higher scores for gains than would be expected in the whole metabolic network (at the corrected $p = 0.001$ level). The 'Metabolic Function' column represents a summary of the enzymatic functions present in the module.

AMBIENT Module ID	Number of Reactions	Metabolic Function	Corrected p-value
1	12	Methylamine metabolism	$< 1e-5$
2	15	Polyamine metabolism	$< 1e-5$
3	12	Phthiocerol dimycocerosate biosynthesis	$< 1e-5$
4	15	Salicin metabolism	$< 1e-5$
5	8	Niacin, Cob(I)alamin metabolism	$< 1e-5$
6	6	Fatty acid biosynthesis	$< 1e-5$
7	6	4-Hydroxybuanoate metabolism	$< 1e-5$
8	5	Monosaccharide metabolism	$< 1e-5$
9	5	Lipid metabolism	$3.6e-4$
10	5	Amino acid metabolism	$3.6e-4$
11	5	Monosaccharide utilisation	$2.2e-4$
12	5	D-Lactate metabolism	$3.6e-4$
13	4	D-Proline metabolism	$2.2e-4$
14	6	N-Acetyl-D-neuraminic acid utilisation	$8.6e-4$

represented in the metabolic network used here. It appears that the reactions around N-Acetyl-glucosamine are shared with the other *Clostridium* strains in this study. One of the significant modules found by `ambient` shown in Fig. 2, shows that *C. difficile* gained the ability to utilise N-Acetyl-neuraminic acid since its divergence from the other *Clostridia* in the study. The assimilation of N-Acetyl-neuraminic acid proceeds by conversion through several intermediates to Fructose-6-Phosphate, which is part of central carbon metabolism.

The reactions responsible for this interconversion, allowing *C. difficile* to utilise this carbon source, have been inferred by Model SEED to be present in this *C. difficile* strain. Some of the reactions in the model were predicted to be present without having a gene associated with them. In the case of this module two genes, nanA and CD196_2092, were associated with two of the reactions, ATPN-acyl-D-mannosamine 6-phosphotransferase and N-Acetylneuraminate pyruvate-lyase, in the module. These genes are transcribed in the same direction and have just three closely spaced same-sense genes between them, each of unknown function. This establishes the tantalising possibility that these three intervening genes could encode proteins with other functions within this coordinately gained metabolic module.

As expected from an obligate intracellular pathogen, `ambient` finds extensive coordinated losses in the *Onion yellows phytoplasma* OY-M metabolic network, with over 350 reactions lost in connected metabolic modules (as shown in Supplementary Fig. 4). The closest relatives of *Onion yellows* in this study share only the same Phylum (Firmicutes), so this represents a long period of evolutionary

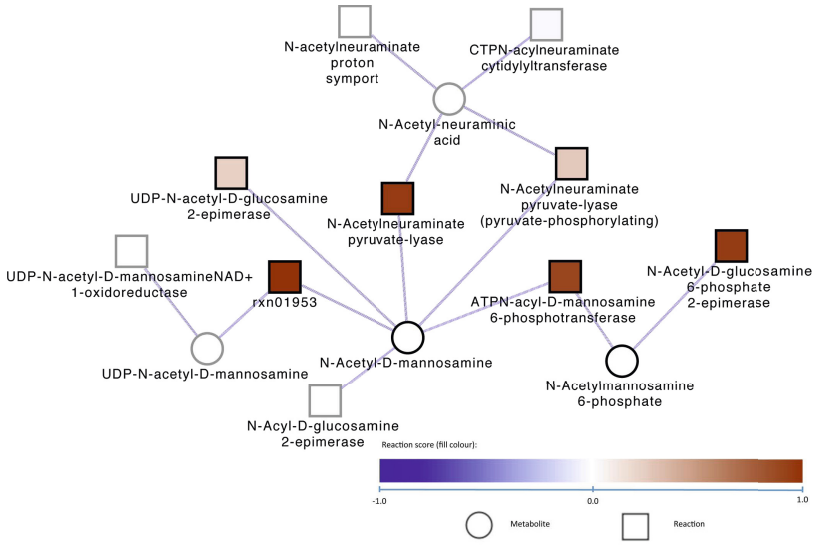

Fig. 2. Metabolic module 14 gained in *C. difficile* CD196 and its metabolic neighbourhood, according to `ambient` analysis of the reaction gains and losses from its closest ancestor on the 13S rRNA phylogenetic tree. Members of module 14 are outlined in black and those not in the module are outlined in grey. The fill colours of the reactions correspond to δP values.

history. Nonetheless *Onion yellows* has only gained (and retained) 91 reactions in the same period, indicating a very strong bias towards metabolic function loss, as expected from the bacterium's lifestyle. By far the largest module shows the complete loss of lipid biosynthesis, as expected since *Phytoplasmas* lack a cell wall.

4 Conclusion

The ancestral state reconstruction results and metabolic traits phylogeny have been able to unpick and clarify the significant gains and losses of metabolic capabilities in various organisms during their evolutionary history. The findings have correlated well with previous biological knowledge of the lifestyles of these organisms. The hierarchical clustering of these bacteria using metabolic traits has shown that as expected metabolic evolution is far more intimately linked with current lifestyle than is bacterial ancestry.

The adaptation of bacteria to different conditions has led to a considerable gain and/or loss of reactions over time. Considerable gain has been observed in *Clostridium difficile*, which is consistent with the expectations for a non-intracellular opportunistic pathogen. Considerable losses, including those of lipid biosynthesis, have been observed in *Onion yellows phytoplasma*, which is a known obligate intracellular plant pathogen which does not produce a cell wall.

The metabolic traits based hierarchical clustering has provided insight into examples of convergent evolution with respect to bacterial metabolism.

The `ambient` analysis presented here has clearly picked out some relevant and biologically meaningful metabolic modules that have been gained or lost in a coordinated fashion. This approach, combined with the multiple metabolic models produced by Model SEED, which can infer reaction presence even in the absence of known enzymes, is a powerful tool that goes beyond previous approaches to investigating metabolic evolution.

Acknowledgments. Thanks to David Hughes, Lesley Hoyles, Pakorn Aiewsakun and Ghazal A Milani for kindly collating the reaction presence/absence tables for the 141 bacteria analysed in this paper from the Model SEED website. AAF was supported by the Mohamedali Habib Welfare Trust, Karachi. WAB was supported by the BBSRC, grant BB/G020434/1. JWP is supported by a University Research Fellowship from the Royal Society.

References

1. Mithani, A., Preston, G.M., Hein, J.: A Bayesian Approach to the Evolution of Metabolic Networks on a Phylogeny. PLoS Computational Biology 6(8), e1000868 (2010)
2. Mazurie, A., Bonchev, D., Schwikowski, B.: Evolution of metabolic network organization. BMC Systems Biology 4(59) (2010)
3. Pfeiffer, T., Soyer, O.S., Bonhoeffer, S.: The evolution of connectivity in metabolic networks. PLoS Biology 3(7) (2005)
4. Pál, C., Papp, B., Lercher, M.J.: Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nature Genetics 37(12), 1372–1375 (2005)
5. Yagi, J.M., Sims, D., Brettin, T., Bruce, D., Madsen, E.L.: The genome of Polaromonas naphthalenivorans strain CJ2, isolated from coal tar-contaminated sediment, reveals physiological and metabolic versatility and evolution through extensive horizontal gene transfer. Environmental Microbiology 11(9), 2253–2270 (2009)
6. Petridis, M., Bagdasarian, M., Waldor, M.K., Walker, E.: Horizontal transfer of Shiga toxin and antibiotic resistance genes among Escherichia coli strains in house fly (Diptera: Muscidae) gut. Journal of Medical Entomology 43(2), 288–295 (2006)
7. Zomorodipour, A., Andersson, S.G.E.: Obligate intracellular parasites: *Rickettsia prowazekii* and *Chlamydia trachomatis*. FEBS Letters 452(1), 11–15 (1999)
8. Schluter, D., Price, T., Mooers, A.Ø., Ludwig, D.: Likelihood of ancestor states in adaptive radiation. Evolution 51, 1699–1711 (1997)
9. Latysheva, N., Junker, V.L., Palmer, W.J., Codd, G.A., Barker, D.: The evolution of nitrogen fixation in cyanobacteria. Bioinformatics 28(5), 603–606 (2012)
10. Merhej, V., Royer-Carenzi, M., Pontarotti, P., Raoult, D.: Massive comparative genomic analysis reveals convergent evolution of specialized bacteria. Biology Direct 4(13) (2009)
11. Baumler, D.J., Peplinski, R.G., Reed, J.L., Glasner, J.D., Perna, N.T.: The evolution of metabolic networks of E. coli. BMC Systems Biology 5(1), 182 (2011)
12. Liao, L., Kim, S., Francois Tomb, J.: Genome comparisons based on profiles of metabolic pathways. In: Proceedings of the 6th International Conference on Knowledge-Based Intelligent Information and Engineering Systems, KES 2002, pp. 469–476 (2002)

13. Whitaker, J.W., Letunic, I., McConkey, G.A., Westhead, D.R.: metaTIGER: a metabolic evolution resource. Nucleic Acids Research 37(Database issue), D531–D538 (2009)
14. Henry, C.S., DeJongh, M., Best, A.A., Frybarger, P.M., Linsay, B., Stevens, R.L.: High-throughput generation, optimization and analysis of genome-scale metabolic models. Nature Biotechnology 28(9), 969–974 (2010)
15. Ideker, T., Ozier, O., Schwikowski, B., Siegel, A.F.: Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics 18(suppl. 1), S233–S240 (2002)
16. Katoh, K., Asimenos, G., Toh, H.: Multiple alignment of DNA sequences with MAFFT. Methods in Molecular Biology 537, 39–64 (2009)
17. Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O.: New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic Biology 59(3), 307–321 (2010)
18. Huson, D., Richter, D., Rausch, C., Dezulian, T.: Dendroscope: An interactive viewer for large phylogenetic trees. BMC Bioinformatics 8(460) (2007)
19. Zientz, E., Dandekar, T., Gross, R.: Metabolic interdependence of obligate intracellular bacteria and their insect hosts. Microbiology and Molecular Biology Reviews 68(4), 745–770 (2004)
20. Maddison, W.P., Maddison, D.R.: Mesquite: a modular system for evolutionary analysis. Version 2.75 (2011), `http://mesquiteproject.org`
21. Pagel, M.: The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. Systematic Biology 48(3) (1999)
22. Swofford, D.L.: PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts (2003)
23. Hubert, L., Arabie, P.: Comparing partitions. Journal of Classification 2(1), 193–218 (1985)
24. Felsenstein, J.: Phylip, `http://evolution.genetics.washington.edu/phylip.html`
25. Oshima, K., Kakizawa, S., Nishigawa, H., Jung, H.Y., Wei, W.: Reductive evolution suggested from the complete genome sequence of a plant-pathogenic phytoplasma. Nature Genetics 36(1), 27–29 (2003)
26. Cox, J.S., Chen, B., McNeil, M., Jacobs, W.R.: Complex lipid determines tissue-specific replication of Mycobacterium tuberculosis in mice. Nature 402(6757), 79–83 (1999)