An approach to pathway reconstruction using whole genome metabolic models and sensitive sequence searching

Mansoor Saqi [1], Richard JB Dobson [1], Preben Kraben [2], David A. Hodgson [3] , David L. Wild [4] *

[1]Barts and The London School of Medicine, Queen Mary, University of London

[2]Department of Biochemical Engineering, University College London

[3]Biological Sciences, University of Warwick

[4]Centre for Systems Biology, University of Warwick

Summary

Metabolic models have the potential to impact on genome annotation and on the interpretation of gene expression and other high throughput genome data. The genome of *Streptomyces coelicolor* genome has been sequenced and some 30% of the open reading frames (ORFs) lack any functional annotation. A recently constructed metabolic network model for *S. coelicolor* highlights biochemical functions which should exist to make the metabolic model complete and consistent. These include 205 reactions for which no ORF is associated. Here we combine protein functional predictions for the unannotated open reading frames in the genome with 'missing but expected' functions inferred from the metabolic model. The approach allows function predictions to be evaluated in the context of the biochemical pathway reconstruction, and feed back iteratively into the metabolic model. We describe the approach and discuss a few illustrative examples.

1 Introduction

Currently available metabolic pathway databases contain hundreds or even thousands of separate metabolic maps describing different parts of general metabolism [10], [13]. However, it is still difficult to get an overall picture of metabolic processes and to answer automatically specific 'network' questions, such as: what are viable external substrates? what are the possible pathways for converting a particular small molecule to its final product? what will be the consequences of one or multiple gene knockouts? can we identify new gene function to extend genome annotation and metabolic and signaling pathway reconstruction? Metabolic models of cellular metabolism help to address these questions.

Metabolic models (or stoichiometric models) are constructed by combining information from numerous biological resources to collect together reactions and corresponding enzymes involved in cellular function in the organism. The stoichiometric models are based on mass balances for each of the compounds in the system, which includes both intracellular and extracellular components. The stoichometric matrix **S** contains the stoichiometric coefficients in

*To whom correspondence should be addressed. E-mail: D.L.Wild@warwick.ac.uk

the system. Each compound corresponds to a row in the matrix and each reaction corresponds to a column in the matrix. The unknown fluxes of each reaction are described by the vector **v**. All intracellular compounds can be assumed to be in pseudo-steady state, because the flux through their pool is much bigger than their pool sizes. The vector **b** contains the rates of production/consumption of each compound in the system, i.e. zero for intracellular compounds, negative values for substrate, and positive values for products, which includes biomass. The stoichiometric matrix, fluxes and rates of production/consumption are related by the equation:

$$\mathbf{S.v=b}$$

Where possible gene names are assigned to enzyme functions, and, using the stiochiometric matrix, the models are balanced (e.g. if a compound is produced but not consumed it must be excreted from the cell, or consumed in another reaction not identified in the organism.)

To make the model consistent it may be necessary to introduce additional functions with no currently identified genes. This opens up the possibility of using the models as an annotation tool: as sequence databases grow and with the development of powerful methods for detecting similarities below the so-called twilight zone of sequence similarity, it may be possible to suggest function for a number of the as-yet unannotated genes. If the predicted function matches what are believed to be essential functions for which no gene has been identified, this may lend greater weight to the prediction, and these genes can become candidates for further experimental investigations. Additionally, if any of the predicted functions have not already been included in the model, then there is scope to refine the model.

Such metabolic models have been constructed for a number of organisms, including *S. coelicolor* A3(2) [4]. *S. coelicolor* is a soil bacterium that excretes biologically active compounds including antibiotics as well as other medically important compounds. Of the 25-30,000 antibiotics known in 2002 about 63% were produced by microbes and about 35% were produced by higher plants. 70% of the microbial antibiotics were produced by bacteria with the vast majority of the rest coming from fungi. Of the 11600 bacterial antibiotics, 8700 where produced by actinomycetes and of these 6550 were produced by a member of the genus *Streptomyces*, i.e. 24% of all known antibiotics are produced by a single bacterial genus. Between 100 and 120 of the streptomycete antibiotics are used commercially, which represents about 68% of the commercial antibiotics that are natural products [3]. *S. coelicolor* A3(2) has a large genome compared to other sequenced bacteria, with 7825 predicted open reading frames, of which about 5492 have some functional annotation [2]. In the model of Borodina et al. [4], 711 proteins (out of 926 that have an Enzyme Commission number assigned) were included. Borodina et al. identified 79 enzyme reactions with no open reading frame (ORF) and then went on to suggest putative candidate genes that may fill roles in the processes of phospholipid biosynthesis and polyprenoid biosynthesis.

Here we discuss how the integration of information from metabolic models and sequence analysis can narrow down the functional hypothesis space for a number of as-yet unannotated ORFs and can hence lead to enhanced genome annotation and facilitate the discovery of new pathways that can be included in the model. We envisage an iterative approach to genome scale metabolic modelling leading to updated models as new information from prediction and experiment becomes available.

2 Computational Method

Sequences from *S. coelicolor* annotated as hypothetical proteins were matched against Pfam [6] families using the hidden Markov profile-profile matching program HHSEARCH [15]. No predicted secondary structure information was used for the query profile. Other sequence database searching programs could have been used. The PFAM2GO resource was used to attach functional annotation from the Gene Ontology (GO) Database [8] to matches to Pfam families. Then the EC2GO resource was used to identify possible enzyme families that corresponded to the respective GO identifiers. Out of a total of 3796 total Pfam identifiers, 2630 had no E.C. mappings. Of the remaining 1166 Pfam identifiers, which had at least one E.C. mapping, 1144 identifiers had 1 E.C. mapping, 12 had mappings to 2 E.C. numbers and 10 mapped to 4 E.C. numbers. A list of 'missing but expected enzymes' was obtained from the supplementary data in the paper by Borodina et al. [4]. All data was stored in a MySQL database for retrieval.

3 Results

All microbial cells have to be able to form biomass in order to multiply. Biomass is a complex mixture of mainly different macromolecules, *e.g.* protein, DNA, RNA, carbohydrate, lipids, and a stoichiometric model must contain the reactions to form these macromolecules from precursor metabolites. Precursor metabolites such as amino acids for protein biosynthesis must either be formed inside the cell or taken up from the external medium. *S. coelicolor* A3(2) has the ability to grow on a chemically defined medium with glucose, ammonium, sulphate, and diverse trace elements, and *S. coelicolor* A3(2) must therefore be able to form all the precursors for the macromolecules and thus be able to divide. Reactions that are necessary for precursor biosynthesis are therefore termed *essential*.

3.1 Diversity of Missing Enzymes

Why do we expect that sensitive sequence searching might suggest candidates for hitherto unidentified functions? Many enzymes are part of large and diverse sequence families. Standard sequence searching methods may fail to detect remote relatives to target sequences and such weak relationships may become apparent using more powerful search tools such as sequence profile matching (e.g. PSIBLAST [1])or profile-profile matching (e.g. HHSEARCH [15]). In order to estimate the utility of such an approach we must first find out how many of the missing functions are orphan enzymes and estimate the diversity associated with the non-orphan enzyme families (Table 1).

An orphan enzyme is an enzyme for which no gene (in any organism) is associated. It has been estimated that some 36% of enzymes activities represented by E.C. classification numbers have no associated gene sequences [11] and, for most of these, the absence of sequence information is real [14]. ORENZA [12], a web resource for orphan enzymes, identifies 3 enzymes in *S. coelicolor* as orphans, namely E.C. 1.5.3.2 N-methyl-L-amino-acid oxidase, E.C. 2.1.1.142 cycloartenol 24-C-methyltransferase and E.C. 2.7.1.136 macrolide 2'-kinase. However none of the missing enzymes from Borodina et. al. [4] correspond to orphan enzymes.

The missing enzymes of Borodina et. al. fall into three categories: missing essential, missing non-essential and other missing enzymes. We find that four of the missing essential enzymes from Borodina et. al [4] do have an *S. coelicolor* gene associated with them in KEGG, namely E.C. 2.3.1.157 (glucosamine-1-phosphate N-acetyltransferase) corresponding to SCO3122, E.C. 4.1.1.17(benzoylformate decarboxylase) corresponding to SCO6035, E.C. 4.1.1.36 (phosphopantothenoylcysteine decarboxylase) corresponding to SCO1477 and E.C. 6.3.2.5 (phosphopantothenate-cysteine ligase) corresponding to SCO1477. This perhaps reflects subsequent annotation of the genome. We have used KEGG to identify how many genes are associated with the missing *S. coelicolor* enzymes in other organisms. We see, for example, that E.C. 2.5.1.33 (farnesyl pyrophosphate), a missing essential enzyme in *S. coelicolor*, has an associated gene in only one other organism, namely the archaeon *Picrophilus torridus*, whereas another missing essential enzyme, E.C. 2.7.4.9 (dTMP kinase) maps to 647 genes from other organisms. In order to get an idea of the sequence diversity of these enzyme families, we collected together the protein sequences of all the genes assigned by KEGG to the missing enzymes, and clustered them at the level of 50% sequence identity using the program BLASTCLUST from NCBI. Although the numbers will change as the sequence databases grow, the resulting number of clusters give a rough indication of sequence diversity (see Table 1). We observe that, for example, the sequence family associated with E.C. 2.7.4.9 (dTMP kinase) is very diverse (diversity measure 125) whereas that associated with E.C. 4.1.3.36 (naphthoate synthase) which is also quite a large family, has a much lower diversity (with diversity measure of 9)

3.2 Examples of holes filled

3.2.1 Thymidylate kinase

One essential missing enzyme predicted by the model of Borodina et al. [4] is thymidylate kinase (E.C. number 2.7.4.9). In order to incorporate thymidylate into the DNA it must first be activated by two sequential phosphorylation steps. The first phosphorylation step is carried out by thymidylate kinase and without phosphorylation thymidylate cannot be incorporated into new DNA and cell multiplication cannot occur.

Sequence searching among the set of unannotated proteins suggests six candidates to fulfill the functional role of thymidalate kinase: SCO0163, SCO2993, SCO1996, SCO1975, SCO1952, SCO0723. These proteins match the Pfam family PF02223 (namely thymidylate kinase). An alignment of SCO0163 (the sequence with the best score), with a representative of the probable thymidylate kinase related cluster UniRef entry Q8PXV5, KTHY_METMA, is shown below.

As thymidylate kinase describes what is an essential function, the corresponding reactions have already been included in the model. However the hypothesis space for identifying the correct gene(s) has been reduced from many fold. The six candidates above can become the focus for further experimental study, e.g. purification and biochemical characterization or gene disruption studies.

Enzyme	Enzyme name	SCO gene	N	D	Putative candidates
1.2.1.21	glycolaldehyde dehydrogenase		27	5	
1.5.1.3	dihydrofolate reductase		461	106	SCO7627, SCO2813, SCO5252, SCO7023
2.3.1.157	glucosamine-1-phosphate N-acetyltransferase	SCO3122	310	24	
2.5.1.11	trans-octaprenyltranstransferase		15	4	
2.5.1.33	trans-pentaprenyltranstransferase		1	1	
2.7.4.9	dTMP kinase		647	125	SCO0723, SCO1952, SCO1975, SCO1996, SCO2993, SCO0163
2.7.2.23	not in KEGG		0	0	
2.7.7.39	glycerol-3-phosphate cytidylyl-transferase		93	17	
2.7.8.12	CDP-glycerol glycerophospho-transferase		12	7	
3.1.3.15	histidinol-phosphatase		227	68	
3.1.3.27	phosphatidylglycerophosphatase		287	72	
3.1.3.7	3'(2'),5'-bisphosphate nucleoti-dase		82	41	
4.1.1.17	benzoylformate decarboxylase	SCO6035	222	47	
4.1.1.36	phosphopantothenoylcysteine decarboxylase	SCO1477	478	53	
4.1.3.36	naphthoate synthase		207	9	
4.1.3.38	aminodeoxychorismate lyase		260	93	
5.4.4.2	isochorismate synthase		241	62	
6.2.1.26	o-succinylbenzoate—CoA ligase		240	105	
6.3.2.5	phosphopantothenate—cysteine ligase	SCO1477	444	39	

Table 1: Missing essential enzymes (with four figure E.C. numbers) from Borodina et al. [4], N is the number of genes associated with the E.C. number from KEGG; D gives a measure of diversity as described in the text; column 6 suggests putative candidates from searching the as-yet unannotated *S. coelicolor* sequences.

Figure 1: An alignment of SCO0163 (the sequence with the best score), with a representative of the probable thymidylate kinase related cluster UniRef entry Q8PXV5

Although we detect hits to PF02223, the mapping of sequences to functional families may be one to many. In this case, we note also that SCO0163 also matches proteins annotated as shikimate kinase (E.C. 2.7.1.71), gluconokinase (E.C. 2.7.1.12) and adenylate kinase (E.C. 2.7.4.3). All three of these enzymes have nucleotide triphosphate binding domains, as they are metabolite kinases. We note the conservation of what may be a possible ATP/GTP-binding site motif at the beginning of the alignment shown. Shikimate kinase has a similar topology to adenylate kinase. Could this sequence fulfill the functional role of thmidylate kinase? The function E.C. 2.7.1.12 is already fulfilled by the gene SCO1679 and the function of E.C. 2.7.1.71 by SCO1495 (*aroK*), and E.C. 2.7.4.3 by SCO4723 (*adk*). 5-dehydroquinate synthase (E.C. 4.2.3.4 – AroB – SCO1494), shikimate 5-dehydrogenase (E.C. 1.1.1.25 – AroD– SCO1498), shikimate kinase (E.C. 2.7.1.71 – AroI(K)– SCO1495) and chorismate synthase (E.C. 4.2.3.5 – AroF – SCO1496) are four enzymes in the seven-enzyme common aromatic amino acid pathway that results in the synthesis of chorismate - the common precursor of phenylalanine, tyrosine and tryptophan. It is therefore unlikely that SCO1495 is involved in thymidylate phosphorylation, as it is in the middle of a five-gene cluster that contains four genes annotated as involved in chorismate biosynthesis, i.e. SCO1494-SCO1498 *aroBI(K)F – aroD*.

3.2.2 Dihydrofolate reductase

Another missing essential enzyme is dihydrofolate reductase (DHFR), E.C. 1.5.1.3. Dihydrofolate is an intermediate in the linear pathway towards folate. Folate is a essential enzyme cofactor in all living organisms, where it plays a central role in the transfer of carbon atoms between different metabolites. The enzyme dihydrofolate reductase catalyses the last reaction in the folate biosynthesis. We find four candidates for this functional role, SCO7627, SCO2813, SCO5252 and SCO7032.

3.3 New Pathways

3.3.1 Cobalamin

Examples of genes currently labelled as hypothetical proteins for which database searching reveals similarity to proteins of known function are SCO1858 and SCO1116. SCO01858 has recently appeared in KEGG but not, as yet, SCO01116. The similarity of SCO1858 (and SCO1116) to CbiX is clear and it appears that this was unannotated in the original annotation of *S. coelicolor*, perhaps due to very conservative thresholds being set. These genes match to CbiX, cobalt chelatase (E.C. 4.99.1.3), that acts in the anaerobic cobalamin biosynthesis pathway (KEGG MAP Porphyrin and chlorophyll metabolism). Cobalamin or vitamin B12 is a complex small molecule that is essential for some organism, e.g. humans. However, it is not essential for *S. coelicolor* because the organism contains pathways and/or parallel reactions that can carry out the same reactions/pathways as the cobalamin-dependent reactions. *S. coelicolor* A3(2) is known to contain several enzymes that use cobalamin (vitamin B12) as a cofactor. There is a cobalamin-dependent homocysteine methyltransferase (E.C. 2.1.1.13 MetH SCO1657) and a cobalamin-independent enzyme (E.C. 2.1.1.14 MetE SCO0985) [9]. A similar situation arises for ribonucleotide reductase with a cobalamin-dependent, oxygen-independent type II form (E.C. 1.17.4.2 NrdJ SCO5805) and a cobalamin independent, oxygen-dependent type Ia form (E.C. 1.17.4.1 NrdAB (NrdLM) SCO5225-6) [5]. In addition there are a number of cobalamin-dependent mutases, e.g. methylmalonyl-CoA mutase (E.C. 5.4.99.2 MutA SCO6832, MutA2 SCO4869) isobutyryl-CoA mutase (E.C. 5.4.99.13 IcmA SCO5415) and a unknown mutase (MeaA SCO6472). Borovok et al. [5] identified nine gene transcripts with cobalamin dependent riboswitches in the 5' untranslated leaders.

Studies on the CbiX proteins [16] suggest the importance of conservation of the two histidine residues (the catalytic residues) as well as a glycine, proline and aspartic acid. A multiple sequence alignment of SCO01858, SCO1116 and several other CbiX proteins suggests that SCO1116 does not have the required patterns of conservation, although a global alignment of these two SCO proteins does indeed reveal global similarity. It may be the case that SCO1116 is a non-functional paralog.

The E.C. number of CbiX is 4.99.1.3 and the associated reaction is: Sirohydrochlorin + Co(2+) <=> cobalt-sirohydrochlorin + 2 H(+) This reaction could now be fed back into the model now that the gene for CbiX has been identified.

Additionally we can now look for other genes in *S. coelicolor* which may participate in the pathway. One example is SCO0993, which reveals a plausible match to Precorrin-6x reductase CbiJ/CobK (E.C. 1.3.1.54), which catalyses the reaction: precorrin-6B + NADP+ <=> precorrin-6A + NADPH + H+. The KEGG pathway database reveals that this is the 7th step in the pathway which coverts sirohydrochlorin to Cob(II)yrinate a,c diamide with CbiX (E.C. 4.99.1.3) catalyzing the first step in this pathway. A search of the annotated *S. coelicolor* genome reveals that most of the enzymes catalyzing the intermediate steps in this pathway have been annotated (see Figure 2). CbiL (E.C. 2.1.1.151, a cobalt-factor II C20-methyltransferase), which catalyses the second step, has no *S. coelicolor* gene is associated with its function but we have detected several putative methyltransferases from our sequence searching among the hypothetical proteins and proteins of unknown function.

The identification of the enzyme E.C. 4.99.1.3, catalyzing the first step, suggests that this path-

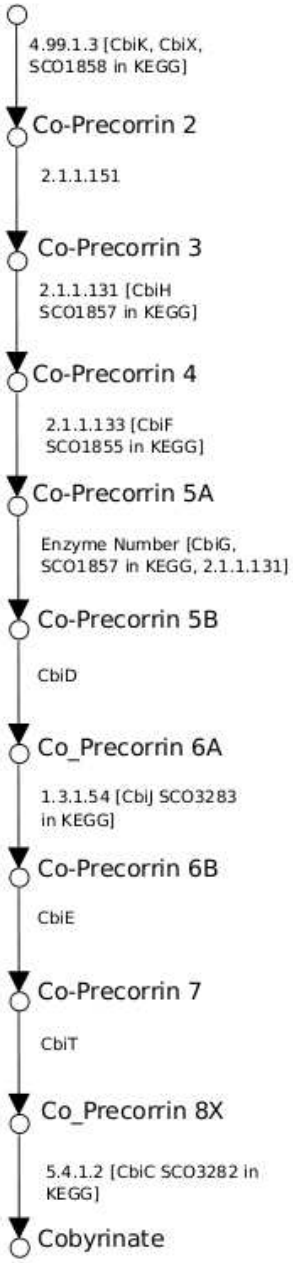

Figure 2: Pathway for sirohydrochlorin to Cob(II)yrinate a,c diamide. Some steps have no associated EC number in the KEGG maps. As only biochemically characterised enzymes are assigned EC numbers, it may be the case that some of the enzymes are theoretical at this stage.

way can now be included in the metabolic model.

3.3.2 Rhamnose

Another example of an enzyme function not currently included in the metabolic model is E.C. 1.1.1.133. A number of proteins in the *S. coelicolor* genome show plausible similarity to this PFAM family (RmlD substrate binding domain). These include SCO7108, which also matches the 3-beta hydroxysteroid dehydrogenase/isomerase family and the NAD-dependent epimerase/dehydratase family, as well as dTDP-4-dehydrorhamnose reductase.

The enzyme function described by E.C. 1.1.1.133 appears in three KEGG maps, map00520 (Nucleotide sugars metabolism), map00521 (Streptomycin biosynthesis),[1] map00523 (Polyketide sugar unit biosynthesis). We suggest below a possible role for E.C. 1.1.1.133 in *S. coelicolor*.

L-rhamnose is a saccharide required for the cell wall components of some bacteria. Its precursor, dTDP-L-rhamnose, is synthesised by four different enzymes, the final one of which is dTDP-4-dehydrorhamnose reductase (RmlD) (E.C. 1.1.1.133). The RmlD substrate binding domain is responsible for binding a sugar nucleotide. The enzyme catalyzes the reaction: dTDP-6-deoxy-L-mannose + NADP+ <=> dTDP-4-dehydro-6-deoxy-L-mannose + NADPH + H+, which the KEGG pathway database reveals as the fifth step in the activation of rhamnose.

The exact function of rhamnose in *Streptomyces* is not totally known, but in related microorganisms it serves as a linker molecule between the structural part of the bacterial cell wall and molecules that are exposed to the surrounding environment. The nature and structure of the cellular envelope of *Streptomyces coelicolor* is presently not known to any great extent. Furthermore several antibiotics consist of rhamnose derived moieties. Although streptomycin is produced by many different streptomyces strains, *S. coelicolor* is not one of them. However, we cannot rule out the possibility that *S. coelicolor* might produce a rhamnose derived compound.

Figure 3 shows a schematic diagram of the steps in the pathway and the candidate genes to fulfill the required functionality. The other enzymes in the pathway (apart from 2.4.2.27) have all been annotated in the *S. coelicolor* genome: E.C 5.4.2.2, phosphoglucomutase (SCO7443, SCO3028), E.C. 2.7.7.24, glucose-1-phosphate thymidylyltransferase (SCO5208, annotated as a putative monophosphatase which also shows sequence similarity to glucose-1-phosphate thymidylyltransferase), E.C. 4.2.1.46, dTDP-glucose 4,6-dehydratase (SCO0395, SCO0749), E.C. 5.1.3.13, dTDP-4-dehydrorhamnose 3,5-epimerase (SCO0400). E.C. 2.4.2.27, dTDP-dihydrostreptose streptidine-6-phosphate dihydrostreptosyltransferase, is not, as yet, annotated in the *S. coelicolor* genome.

The identification of the missing enzyme E.C. 1.1.1.133 in the pathway therefore means that it could now be included in the metabolic model.

[1] Note that in KEGG map 00521 this is wrongly annotated as E.C. 1.1.1.13

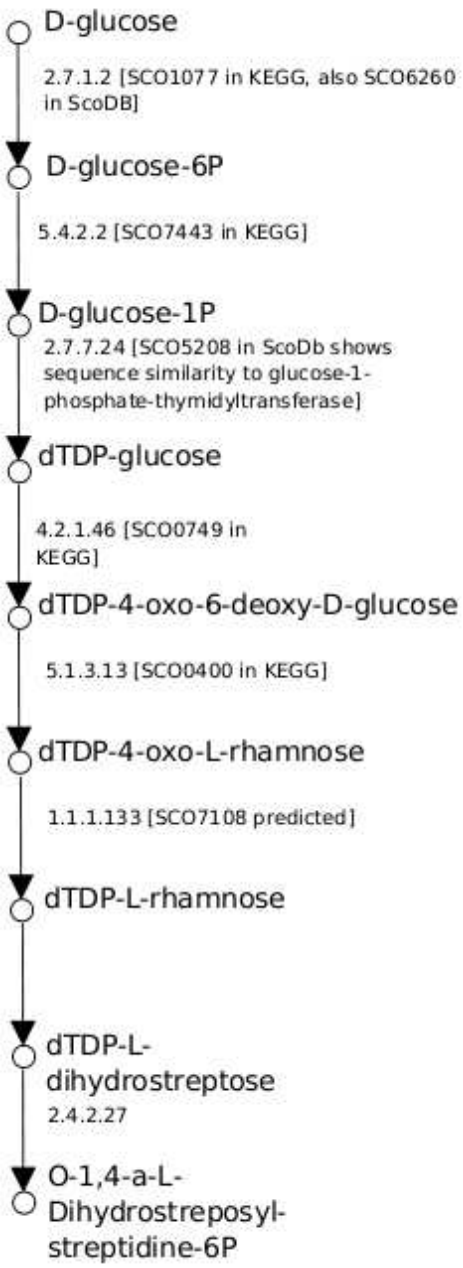

Figure 3: Pathway for rhamnose

3.4 Comparison with the Pathway Hole Filler algorithm

Holes in pathways can also be identified using the Hole Filler algorithm [7] and are these are available from the SCODB resource (see `http://scocyc.jic.bbsrc.ac.uk:1555/SCO/`. We compare the list of missing enzymes (generated using Hole Filler algorithm) with the missing essential enzymes from Borodina et al. and find the only overlap is E.C. 1.5.1.3 (dihydrofolate reductase). We see that the choice of approach used can affect the predictions of missing enzymes. Table 3.4 gives examples of predictions obtained from sensitive sequence searching for putative proteins for the missing enzymes identified by the Hole Filler algorithm.

Enzyme	**SCOCyc Pathway**	**KEGG pathway**	**Candidate SCO genes**
5.1.3.14	teichoic acid (poly-glycerol) biosynthesis	Aminosugars metabolism (MAP00530)	SCO1779 SCO0383
1.7.2.1	denitrification pathway	Nitrogen metabolism (MAP00910)	SCO0798 SCO0563 SCO4955 SCO5953 SCO0305 SCO6621
1.13.11.20	L-cysteine degradation I	Cysteine metabolism (MAP00272) Taurine and hypertaurine metabolism (MAP0430)	SCO5572 SCO3035
2.9.1.1	selenocystein biosynthesis	Selinoamino acid metabolism (MAP00450)	SCO1868
1.1.1.44	formaldehyde oxidation I	Pentose phosphate pathway (MAP00030)	SCO0965 SCO2482 SCO3384 SCO5465
2.7.7.22	colanic acid building blosks biosynthesis, GDP-mannose metabolism	Fructose and mannose metabolism (MAP00051)	SCO1208 SCO0820 SCO1737 SCO3035 SCO3803 SCO3923 SCO4483 SCO4789 SCO6435 SCO7126 SCO7127
1.5.1.3	formylTHF biosynthesis I, formylTHF biosynthesis, tetrahydrofolate biosynthesis, tetrahydrofoliate biosynthesis I	One carbon pool by folate (MAP00670), Folate biosynthesis (MAP00790)	SCO7627 SCO2813 SCO5252 SCO7032

Table 2: Holes identified by HoleFiller and possible candidates identified from sequence searching

4 Discussion

As sensitive methods for probing into the so-called twilight zone of sequence similarity continue to develop, and more genome sequences are added to the sequence databases, it is possible that putative annotation may be assigned to some of the as-yet unannotated genes. This annotation can be evaluated in the context of metabolic networks. The approach described in this paper has the potential to reduce the potential functional space of the genome: firstly by suggesting candidates for missing functions, and secondly by suggesting new functions that can be examined in light of known pathways and the extent to which steps in these can be mapped to genes in the organism. It is probable, using this approach, that a number of candidate genes can be identified which could encode enzymes to fill primary metabolic pathway gaps. This is due the presence of redundant primary metabolic pathways and multiple isozymes present in streptomycetes, possibly a consequence of its oligotrophic lifestyle. However, the presence of large numbers of genes encoding enzymes involved in secondary metabolism complicate the issue. These genes are derived from genes encoding primary metabolic enzymes that have been recruited to secondary metabolism. This is a particular problem when considering fatty acid synthesis, which is confused with polyketide secondary metabolite synthesis, and amino acid catabolic enzyme genes which are often recruited to secondary metabolism, for example, lysine catabolism and β-lactam biosynthesis [9].

A further complication is that streptomycetes often use metabolic pathways that are different to more well-studied bacteria such as enteric bacteria and *Bacillus* species. For any pathway the order in which particular reactions are carried out in a pathway is a product of the evolutionary history of the bacterium. For example, actinomycetes use arogenate as an intermediate in tyrosine synthesis instead of hydroxyphenylpyruvate, as used by enteric bacteria. Therefore, a pathway reconstruction approach that is based on enzyme activity rather than gene homology would be more likely to yield useful information. Not all bacterial metabolism is the same. Often there are extra pathways and shunt pathways that do not exist in more well-studied bacteria e.g. the trans-sulphuration pathway of *Streptomyces* is more characteristic of filamentous fungi than bacteria. This reflects the ecological niche of streptomycetes which are facultative oligotrophs, i.e. capable of living in low- nutrient environments, unlike enteric bacteria and bacilli. Soil, the natural habitat of streptomycetes, is a low-nutrient environment that is carbon-rich and nitrogen- and phosphate-poor. This is because it is derived from carbon-rich and nitrogen- and phosphate-poor plants. This explains why streptomycetes have so many genes encoding carbon catabolic pathways, again when compared to more well studied bacteria. The metabolism of filamentous fungi, which inhabit the same ecological niche, are often a better model for streptomycetes than bacteria [9].

A webserver implementation of the approach is available at
`http://cluster.wsbc.warwick.ac.uk/cgi-bin/strep/metagaps.pl`

5 Acknowledgements

The authors acknowledge support from grants BBSRC/EPSRC GR/S29256/01, BBSRC BB/F003498/1 (SysMo initiative) and EU Marie Curie IRG 46444

References

[1] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–402, 1997.

[2] S.D. Bentley, K.F. Chater, A.M. Cerdeno-Tarraga, G.L. Challis, N.R. Thomson, K.D. James, D.E. Harris, M.A. Quail, H. Kieser, D. Harper, A. Bateman, S. Brown, G. Chandra, C.W. Chen, M. Collins, A. Cronin, A. Fraser, A. Goble, J. Hidalgo, T. Hornsby, S. Howarth, C.H. Huang, T. Kieser, L. Larke, L. Murphy, K. Oliver, S. O'Neil, E. Rabbinowitsch, M.A. Rajandream, K. Rutherford, S. Rutter, K. Seeger, D. Saunders, S. Sharp, R. Squares, S. Squares, K. Taylor, T. Warren, A. Wietzorrek, J. Woodward, B.G. Barrell, J. Parkhill, and D.A. Hopwood. Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2). *Nature*, 417(6885):141–7, 2002.

[3] J. Berdy. Bioactive microbial metabolites - a personal view. *Journal of Antibiotics*, 58:1–26, 2005.

[4] I. Borodina, P. Krabben, and J. Nielsen. Genome-scale analysis of Streptomyces coelicolor A3(2) metabolism. *Genome Res*, 15(6):820–9, 2005.

[5] I. Borovok, B. Gorovitz, R. Schreiber, Y. Aharonowitz, and G. Cohen. Coenzyme b12 controls transcription of the streptomyces class ia ribonucleotide reductase nrdabs operon via a riboswitch mechanism. *Journal of Bacteriology*, 188:2512–2520., 2006.

[6] R.D. Finn, J. Mistry, B. Schuster-Bockler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S.R. Eddy, E.L. Sonnhammer, and A. Bateman. Pfam: clans, web tools and services. *Nucleic Acids Res*, 34(Database issue):D247–51, 2006.

[7] M.L. Green and P.D. Karp. A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics*, 5:76, 2004.

[8] M.A. Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, J. Richter, G.M. Rubin, J.A. Blake, C. Bult, M. Dolan, H. Drabkin, J.T. Eppig, D.P. Hill, L. Ni, M. Ringwald, R. Balakrishnan, J.M. Cherry, K.R. Christie, M.C. Costanzo, S.S. Dwight, S. Engel, D.G. Fisk, J.E. Hirschman, E.L. Hong, R.S. Nash, A. Sethuraman, C.L. Theesfeld, D. Botstein, K. Dolinski, B. Feierbach, T. Berardini, S. Mundodi, S.Y. Rhee, R. Apweiler, D. Barrell, E. Camon, E. Dimmer, V. Lee, R. Chisholm, P. Gaudet, W. Kibbe, R. Kishore, E.M. Schwarz, P. Sternberg, M. Gwinn, L. Hannick, J. Wortman, M. Berriman, V. Wood, N. de la Cruz, P. Tonellato, P. Jaiswal, T. Seigfried, and R. White. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*, 32(Database issue):D258–61, 2004.

[9] D. A. Hodgson. Primary metabolism and its control in streptomycetes: a most unusual group of bacteria. *Advances in Microbial Physiology*, 42:47–238, 2000.

[10] M. Kanehisa and S. Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27–30, 2000.

[11] P.D. Karp. Call for an enzyme genomics initiative. *Genome Biol*, 5(8):401, 2004.

[12] O. Lespinet and B. Labedan. ORENZA: a web resource for studying ORphan ENZyme activities. *BMC Bioinformatics*, 7:436, 2006.

[13] N. Maltsev, E. Glass, D. Sulakhe, A. Rodriguez, M.H. Syed, T. Bompada, Y. Zhang, and M. D'Souza. PUMA2–grid-based high-throughput analysis of genomes and metabolic pathways. *Nucleic Acids Res*, 34(Database issue):D369–72, 2006.

[14] Y. Pouliot and P.D. Karp. A survey of orphan enzyme activities. *BMC Bioinformatics*, 8:244, 2007.

[15] J. Soeding. Protein homology detection by HMM-HMM comparison. *Bioinformatics*, 21(7):951–60, 2005.

[16] J. Yin, L.X. Xu, M.M. Cherney, E. Raux-Deery, A.A. Bindley, A. Savchenko, J.R. Walker, M.E. Cuff, M.J. Warren, and M.N. James. Crystal structure of the vitamin B12 biosynthetic cobaltochelatase, CbiXS, from Archaeoglobus fulgidus. *J Struct Funct Genomics*, 7(1):37–50, 2006.