2 Development of StyCyc and StmCyc and validation of known metabolic lesions

2.1 Introduction

Highly curated metabolic pathway databases exist either as collections across multiple species or as more specific resources for particular organisms. Pathway Tools (SRI International, California, USA) is a freely available piece of software for the latter, which predicts a repertoire of metabolic pathways given a fully annotated genome sequence. The Pathway Tools group is responsible for the BioCyc collection of metabolic databases (http://www.biocyc.org), with a particular focus on EcoCyc, the “encyclopaedia of *Escherichia coli* K-12 genes and metabolism” (http://www.ecocyc.org) (Karp et al. 1997; Keseler et al. 2009; Ouzounis and Karp 2000).

A Pathway/Genome Database (PGDB) built using Pathway Tools is designed to represent data on genes, proteins and enzymatic reactions, through to the metabolic pathways they make up. This integrated approach has led to the development of numerous bioinformatic tools to extract the maximum amount of information from a well-annotated genome sequence (Karp et al. 2002). The hallmark of Pathway Tools is to make use of high quality annotation rather than rely entirely upon computational methods. Thus, the software often uses name-matching algorithms to deduce enzymatic function, for example the Transport Identification Parser is used to predict cellular transport reactions, identifying the substrates that an organism can successfully import or export from the cytoplasm (Lee et al. 2008). All reactions, whether in pathways or transporters, are then visualised in an overview diagram which acts as a framework for any gene-, protein- or metabolite-based data to be ‘painted’ on (Paley and Karp 2006). The outstanding feature of Pathway Tools however, is the Pathway Hole Filler algorithm (Green and Karp 2004), which facilitates the identification of novel enzymatic functions within the genome sequence being analysed and hence provides updated annotation.

EcoCyc provides an information retrieval system for *E. coli* K12 that is publically available via the internet. This has allowed studies performed on this strain to be analysed in the context of metabolic pathways, from large scale systems biology (Hyduke et al. 2007) to gene expression (e.g.(Konig and Eils 2004; Schramm et al. 2007)) to metabolic engineering (Chassagnole et al. 2002). The level of curation that has been achieved from over 10 years of evidence-based literature searching makes EcoCyc one of the most comprehensive resources available on a single organism.

Salmonella and *E. coli* are closely related members of the Enterobacteriaceae, and as such, EcoCyc represents a wealth of information that serves as a useful resource for the creation of a *Salmonella* Pathway/Genome database (PGDB).

For *Salmonella*, only one pathway database had been published at the beginning of this work. However, this map of Typhimurium LT2 metabolism was edited from the *E. coli* version depicted in EcoCyc, rather than being built *de novo*. While this was a useful shorthand method that was used to describe the metabolism of four *Salmonella* serovars during typhoid or typhoid-like disease (Becker et al. 2006), it has not been made publically available. Also, using a single database to model different serovars makes it more difficult to ask serovar-specific questions. Thus, at the outset of this project, no curated pathway database existed for Typhi.

The level of genome degradation in Typhi and Paratyphi A, coupled with evidence that these serovars have more restricted metabolic profiles than other serovars, prompts the hypothesis that there is a causal link between host restriction and metabolic ability (Uzzau et al. 2000). However, there is currently very little evidence to support this in the literature. Whilst various complete *Salmonella* genome sequences are now available, this lack of evidence is likely due in part to the paucity of information describing gene function and an inability to assign individual pseudogenes to metabolic pathways. The aim therefore was to build one database to represent the metabolism of Typhi in order to visualise the metabolic pathways interrupted by pseudogenes and one to represent Typhimurium as a comparator.

2.2 Methods

2.2.1 Pathway Tools

Pathway Tools version 11.0 (SRI International, California, USA), was installed on a Debian Linux system under the academic license available at http://www.biocyc.org. The software contains the PathoLogic suite, which allows the user to build metabolic databases from scratch.

2.2.1.1 Input file formatting

The Sanger-sequenced Typhi CT18 genome sequence and annotation was used as the basis of the Pathway/Genome Database (PGDB), named StyCyc. A custom Perl script (Appendix 8.1.1 CD_001) was used to parse the Typhi GenBank annotation file to ensure that information contained in qualifiers would be assigned correctly within the PGDB framework. All STY unique gene IDs (/systematic_id) were given the qualifier '/label' and all '/note's were renamed '/product_comment'. A first-pass automated PGDB build indicated that genes annotated with the '/pseudo' qualifier would be removed by the software before pathway construction. In order to understand where potential enzymes encoded by these genes would be placed with metabolic pathways, this qualifier was removed. Instead, all pseudogenes had the string "(pseudogene)" added to the gene product name to enable identification in the database.

2.2.1.2 Pseudogene re-annotation

It was possible that the enzymes encoded by pseudogenes would be candidates for pathway reactions. To allow these enzymes to be considered, the annotated translation needed to ‘correct’ any stop codons or frameshifts present in the sequence. Typhi CT18 was one of the first bacterial sequences completed at the Sanger Institute, before standardised methods were employed for the annotation of pseudogenes. Thus, frameshifted pseudogenes were annotated only in the frame in which the start codon was encoded. Each of these was reannotated using the Artemis Comparison Tool (ACT (Carver et al. 2005)) to allow automated translation of the whole reading frame. Typhimurium strains LT2 and SL1344 were used as comparators to reconstruct the intact sequence. Pseudogenes caused by stop codons were not reannotated as the automated translation was unaffected.

2.2.1.3 Assigning enzymes and multimers

Using the Pathway Tools software, the initial round of pathway prediction was based upon a name-matching algorithm, relying on high quality gene annotation to determine the presence of particular enzymes and assign them to relevant reactions. The software generated a report on this process, giving details of potential enzymes which could not be matched. This list was manually processed for typographical errors and probable enzymes were assigned to reactions based upon literature searches and/or sequence comparison with Typhimurium and *E. coli*. Some enzymes had non-specific names e.g. alcohol dehydrogenase and therefore could not be assigned to specific reactions.

For over 150 metabolic reactions, more than one enzyme was predicted to perform the relevant function. These were assessed individually to determine whether the enzymes formed a multimeric complex or whether they represented functional isomers. Complexes were confirmed based upon genome position (i.e. within operons), literature searches and/or comparison with Typhimurium and *E. coli*; otherwise the enzymes were, by default, assumed to be isomers.

2.2.1.4 Pathway prediction

Based on the assignment of enzymes to reactions, the presence of certain metabolic pathways was inferred for the organism of interest from MetaCyc (http://www.metacyc.org), a reference database containing experimentally elucidated metabolic pathways from over 1,000 organisms (Caspi et al.). The software is described as over-predictive, as, for example, a pathway is often inferred when some of its constitutive enzymes were not name-matched in the organism of interest. Therefore, once these stages were complete, a manual check of each predicted pathway was performed to ensure only genuine pathways remained in the database. Each pathway (imported from MetaCyc) was marked with the known range of organisms in which that pathway had been found. This information was used to detect and remove false-positive predicted pathways by assessing whether particular enzymes had been placed in single or multiple pathways. However, once this checking was complete, some pathways still contained reactions without an assigned enzyme. These were known as ‘pathway holes’ and were initially processed by the inbuilt algorithm ‘Pathway Hole Filler’.

2.2.1.5 Transport reactions

Another tool in the software, the Transport Identification Parser, was used to predict transport reactions within the PGDB. Based upon gene annotation, gene products were grouped by substrate and transport mechanism, e.g. ATP-driven, channel-type facilitator or secondary transport. The automated predictions were manually assessed and confirmed if the genome annotation was sufficient to support the assignment and/or experimental evidence was available in the literature and/or comparison with Typhimurium and *E. coli* indicated that amino acid sequence was conserved (> 70% identity) and syntenic. Predictions were rejected when the literature searches were unsuccessful and amino acid sequence was either not conserved (< 70% identity) or not present. Predictions from TransportDB (http://www.membranetransport.org) were also taken into account.

2.2.2 Pathway hole filling

2.2.2.1 Automated hole filling

One of the strongest motivations for using Pathway Tools was the inbuilt Pathway Hole Filler algorithm. This four-step method identified gene candidates from Typhi CT18 to fill enzymatic holes in metabolic pathways. To do this, the function of the missing enzyme was inferred from the complete version of the pathway present in MetaCyc. Stage one in the hole-filling process was the retrieval of enzyme sequences from other organisms which catalysed the appropriate reaction. These sequences were used in a BLAST-p search against the genome sequence of Typhi CT18 and the top hits were recorded. A data consolidation step was then performed, pulling together a summary of the BLAST data to be used as evidence that a particular candidate had the function required to fill the pathway hole. All the evidence per candidate was then evaluated using a Bayes classifier. The evaluation included whether the candidate was part of an operon and/or adjacent to a gene coding for the enzyme catalysing an adjacent reaction in the pathway. Finally, the classifier generated a probability that the candidate had the function required to catalyse the missing reaction.

2.2.2.2 Manual evaluation of pathway hole candidates

Every pathway with at least one hole-filling candidate was manually assessed. Pathways which would become fully intact with the assignment of a hole-filling candidate were considered a priority for assessment. Where possible, the pathway was compared with the equivalent in EcoCyc and/or MetaCyc. ACT comparisons (Carver et al. 2005) were used to determine if the Typhi candidate was conserved (> 70% amino acid identity) and syntenic with the enzyme from *E. coli*. Possible amino acid sequence structure homology was detected using FUGUE (Shi et al. 2001). Literature searches were also performed to obtain experimental evidence for assignments. Pathways which had a taxonomic range outside of the Enterobacteriaceae and had multiple holes with no candidates were pruned from the database. Pathways were also deleted when no evidence for the pathway existed, either from unsuccessful literature searches or from lack of amino acid conservation (< 70% identity) of predicted enzymes with potential orthologues in *E. coli*. In addition, if any enzymes previously assigned (automatically) were not unique to the pathway, and no further evidence was found, the pathway was also deleted.

2.2.2.3 PGDB build process for Typhimurium

Subsequent to StyCyc, a new PGDB was built for Typhimurium using the SL1344 genome sequence and annotation. This strain contains 40 pseudogenes annotated in the correct frames and the annotation used standard qualifiers, so no additional input file formatting was required. The same PGDB build procedures of enzyme and multimer assignment, pathway prediction and hole filling were completed. All hole filling-candidates were assessed as for StyCyc, except that comparisons were made against both Typhi and *E. coli*.

2.3 Results

2.3.1 Generation of StyCyc 1.0

Sequence and annotation information from Typhi strain CT18 was used to generate a Pathway/Genome Database (PGDB), named StyCyc to follow the convention of EcoCyc and MetaCyc. An automated build procedure resulted in the importation (from MetaCyc) of all pathways containing a reaction catalysed by an enzyme in Typhi. Some pathways were subsequently deemed to have insufficient evidence (e.g. only one reaction with an enzyme assigned) and immediately removed from the PGDB. Following this, a number of manual refining steps were performed. This included scrutinising over 400 ‘probable enzymes’ which had been missed in the initial name-matching exercise for possible enzymatic assignments. Approximately 1/3 had names that were too generic to be assigned to reactions (e.g. putative hydrogenase). A minority of enzyme names contained typographical errors and were easily assigned while the remainder were searched in EcoCyc and MetaCyc to determine function.

Each Typhi protein coding gene was assigned a corresponding polypeptide, of which 1,033 were predicted to be enzymes and over 100 formed protein complexes. Together, these catalyse over 1,200 enzymatic reactions containing 925 compounds. A further 184 transporters were also predicted, manually confirmed and assigned to 145 transport reactions (Table 2-1).

A fully-automated run of the Pathway Hole Filler was completed for StyCyc 1.0, where top candidates (according to the inbuilt Bayes classifier - see Methods), for pathway holes were accepted without manual intervention. In total, 474 holes were identified in predicted Typhi metabolic pathways and the Hole Filler identified candidates from the genome to fill 217 of these, resulting in 59 intact pathways out of a total of 312 (Table 2-1). This indicated that the hole-filling algorithm was capable of finding novel functions for Typhi gene products.

Table 2-1 Statistics for StyCyc 1.0

Pathways	312
Enzymatic reactions	1264
Transport reactions	145
Polypeptides	4404
Protein complexes	103
Enzymes	1033
Transporters	184
Compounds	925

Total numbers in database categories after an automated build process, manual curation of unassigned enzymes and a fully-automated run of Pathway Hole Filler.

2.3.1.1 Issues and resolutions

Upon initial examination of StyCyc 1.0, it became apparent that pseudogenes were not given associated gene products and hence were not assigned to any metabolic pathways. As part of the intended utility of the Typhi metabolic map was to determine the location of pseudogenes within pathways, this needed to be resolved. The ‘/pseudo’ qualifier that identified pseudogenes to the software was removed from the input file. As an alternative identification mechanism, the string ‘(pseudogene)’ was added to the name of the product of each pseudogene.

In addition, for pseudogenes to be considered as candidates by the Hole Filler, accurate protein sequences were required. The initial annotation of Typhi displayed all pseudogenes in the same frame as the start codon and therefore did not take into account the effects of frameshift mutations. Out of the 204 pseudogenes identified in CT18 (Parkhill et al. 2001a), 96 were caused by frameshifts, so these were re-annotated in the appropriate frames and used to update the input file (Table 2-2).

Table 2-2 Classification of pseudogenes in Typhi CT18

Genetic lesion	Number
Frameshift	96
In-frame stop codon	76
Insertion	4
Fragment/remnant	32
Truncation	9
Internal deletion	4
Fusion	1

In accordance with the annotation of Typhi CT18, all 204 pseudogenes were examined for genetic lesions in ACT (Carver et al. 2005). In a few cases, more than one genetic lesion was recorded. Only frameshift mutations required re-annotation.

A key requirement of the Pathway Tools software was for each gene to be given a unique identifier which would be taken from the annotation. Typhi genes were identified by their ‘STY’ numbers, but these were not propagated through to StyCyc 1.0 due to the software not recognising the qualifier ‘/systematic_id’ in the original annotation file. The qualifier was renamed to ensure the software correctly recognised the information it was being given.

Before rebuilding StyCyc, the enzyme assignment decisions made early in the creation of StyCyc 1.0 were transferred into the locally stored enzyme mapping file, to be used by all subsequent *Salmonella* PGDB builds.

2.3.2 StyCyc 2.0 and onwards

2.3.2.1 Improvements from StyCyc 1.0

As well as the improvements detailed above, new data on Typhi pseudogenes were included in the updated input file. A high-throughput sequencing study of 19 Typhi strains had generated a list of pseudogenes that were either core (i.e. present in all) or variable (absent in some) in the sequenced population (Holt et al. 2008). The variable pseudogenes were given the suffix '(VPT)' (for variable pseudogene in Typhi) in the gene product name. The updated Typhi annotation was then used to build StyCyc 2.0.

2.3.2.2 New pathways and transport reactions

Typhi expresses the Vi antigen, synthesised by genes encoded upon SPI-7. This pathogenicity island is found in only a few *Salmonella* serovars and in some *Citrobacter*. Vi antigen biosynthesis was therefore not present in MetaCyc and could not be imported into StyCyc. Using experimental evidence from the literature, this pathway and its transport reaction was reconstructed using the editing tools available in Pathway Tools (Virlogeux et al. 1995; Zhang et al. 2006) (Figure 2-1).

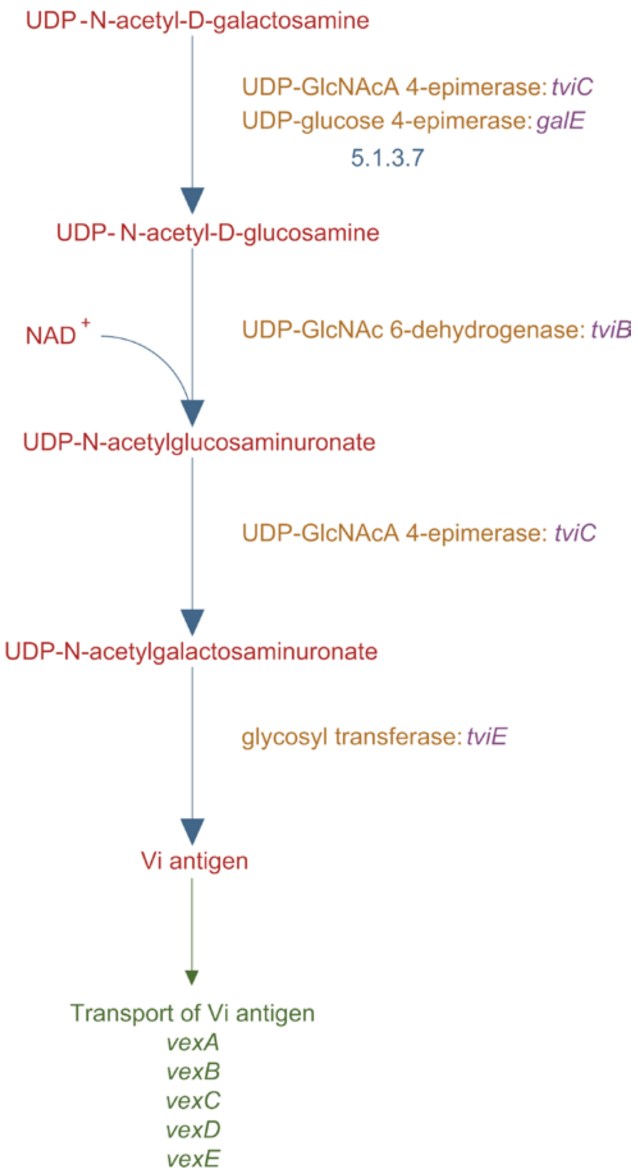
UDP-N-acetyl-D-galactosamine
UDP-GlcNAcA 4-epimerase: *tviC*
UDP-glucose 4-epimerase: *galE*
5.1.3.7
UDP- N-acetyl-D-glucosamine
NAD +
UDP-GlcNAc 6-dehydrogenase: *tviB*
UDP-N-acetylglucosaminuronate
UDP-GlcNAcA 4-epimerase: *tviC*
UDP-N-acetylgalactosaminuronate
glycosyl transferase: *tviE*
Vi antigen
Transport of Vi antigen
vexA
vexB
vexC
vexD
vexE

Figure 2-1 Vi antigen biosynthesis

Pathway Tools depiction of the Vi antigen biosynthetic pathway. Pathway direction is indicated by blue arrows. Gene names given in purple; enzyme names in gold; compound names in red. Blue number, enzyme commission (E.C.) number for the reaction. Transport reaction and related genes shown in green.

Using EcoCyc as a standard for presentation of central metabolic pathways, links between glycolysis, pyruvate dehydrogenase and the TCA cycle were created to form a 'superpathway' (Figure 2-2). In Pathway Tools, superpathways simply represent a 'bigger picture' in order to enable the user to better understand how highly related processes are connected. Similarly, a superpathway was created to visualise how the glyoxylate cycle relates to the TCA cycle.

Initially, predictions from the inbuilt transport identification module, based upon gene annotation, were used to assign transport reactions. Information from TransportDB was then used to identify other potential transport reactions based on protein sequence and seventeen transporters were added to the database (Table 2-3). Multidrug efflux systems which act in co-ordination with TolC were also added based upon evidence from the literature (Nishino et al. 2006).

Table 2-3 Predicted transporter proteins in Typhi CT18

Type	Number	Example substrates
Channel-type facilitators	11	nickel; calcium; glycerol; formate
Secondary transporters	47	glutamate; rhamnose; citrate; fucose
ATP-driven transporters	37	thiosulfate; glutathione; maltose; methionine
PEP-driven transporters	11	cellobiose; mannose; glucose; fructose
Unknown mechanism	27	serine; xanthine; gluconate; glucarate

PEP, phosphoenolpyruvate.

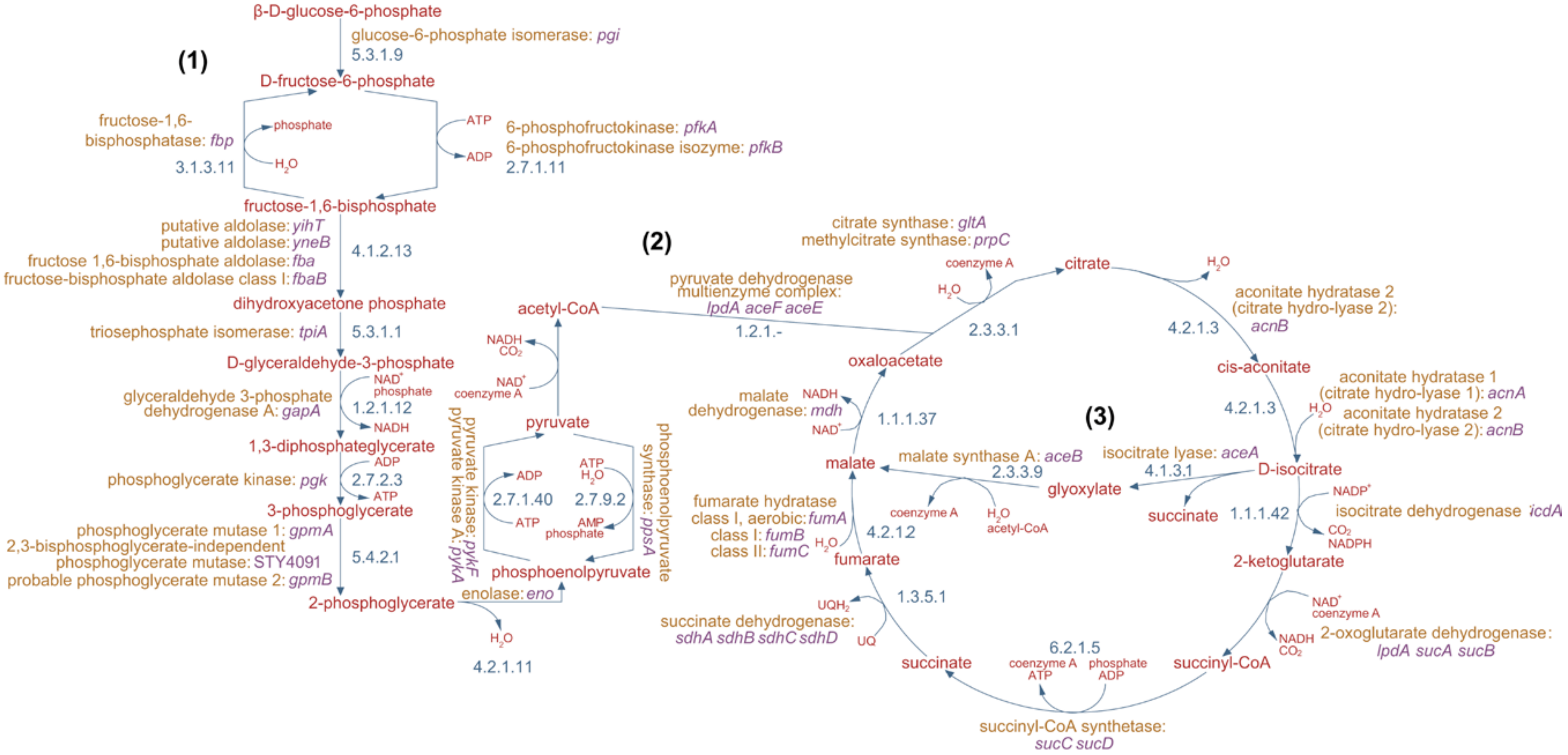

Figure 2-2 Superpathway of glycolysis, pyruvate dehydrogenase and TCA cycle

Representation of (1) glycolysis, (2) pyruvate dehydrogenase and (3) the TCA cycle as an interconnected ‘superpathway’. Pathway direction is indicated by blue arrows. Gene names given in purple; enzyme names in gold; compound names in red. Blue numbers, enzyme commission (E.C.) numbers for reaction

2.3.2.3 Pathway Hole Filling

The first stage of the hole-filling process identified 85 metabolic pathways with hole-filling candidates. Between 1 and 500 protein sequences were retrieved from enzymes performing the missing functions in other organisms. The majority of the incomplete pathways contained only one hole; the remainder between 2 and 5 (Table 2-4). A summary of how many pathways were completely filled is given in Table 2-5.

Table 2-4 Incomplete pathways in StyCyc

Holes	Number of pathways	Examples
1	54	coenzyme A biosynthesis; ethylene glycol degradation
2	12	lipoate biosynthesis and incorporation; ECA biosynthesis
3	13	histidine biosynthesis; menaquinone biosynthesis
4	4	4-hydroxyphenylacetate degradation; thiamine biosynthesis
5	2	lipid A core biosynthesis; methylerythritol phosphate pathway

StyCyc pathways contained between 1 and 5 pathway holes before assignment of hole-filling candidates. ECA, enterobacterial common antigen.

Table 2-5 Summary of StyCyc pathway outcomes

	Pathways			
Holes	Deleted	Completely filled	Partly filled	Not filled
1	26	21	0	7
2	3	6	2	1
3	2	9	0	2
4	1	2	0	1
5	0	2	0	0

Pathways without sufficient evidence were deleted from the PGDB. Hole-filling candidates were manually assessed for remaining pathways and assigned only with evidence as described in the methods.

2.3.2.4 PGDB statistics

The current version of StyCyc is StyCyc 7.0 which has undergone both literature-based and direct curation. The direct curation came from the subsequent creation of a PGDB for Typhimurium and is described in greater detail in section 2.3.3 below. StyCyc 7.0 contains 200 predicted pathways and 133 transport reactions. A total of 1,024 enzymatic reactions, containing 821 compounds, are catalysed by 1,052 enzymes (Table 2-6). The changes from previous StyCyc versions are for the most part reductions in the number of pathways, reactions and compounds predicted to be present in Typhi. This is due to curation aimed at identifying and eliminating false pathways and their associated reactions and compounds. However, both the number of polypeptides and protein complexes has increased from StyCyc 1.0 to 7.0. The former is due to the recognition of 204 pseudogenes and 2 genes whose corrupted information has been repaired and the latter is an increase from 103 to 139, due to the identification of new protein complexes based upon literature searches.

Table 2-6 Statistics from StyCyc 7.0

Category	Number (change from StyCyc 1.0)	
Pathways	200	(112 fewer)
Enzymatic reactions	1011	(253 fewer)
Transport reactions	133	(12 fewer)
Polypeptides	4610	(206 more)
Protein complexes	139	(36 more)
Enzymes	1052	(19 more)
Transporters	150	(34 fewer)
Compounds	821	(104 fewer)

These statistics reflect the status of StyCyc 7.0 after an automated build process, manual curation and a manually supervised run of Pathway Hole Filler.

In the Pathway Tools software, pathways are classified in a hierarchy that at the highest level divides into broad areas such as biosynthesis, degradation and generation of precursor metabolites and energy. The breakdown of the highest two levels is given in Table 2-7. This hierarchy recognises classical pathways such as glycolysis and fermentation alongside some unique to this software. For example, aminoacyl-tRNA charging is not strictly a metabolic pathway but being classed as such allows all 20 tRNA-charging reactions to be grouped together.

Table 2-7 Pathway hierarchy in Pathway Tools

Class (Sty/Stm)	Sub-class	Sty	Stm
Biosynthesis (140/139)	Amines and polyamines	8	8
	Amino acids	40	40
	Aminoacyl-tRNA charging	1	1
	Aromatic compounds	2	2
	Carbohydrates	8	8
	Cell structures	11	10
	Cofactors, prosthetic groups, electron carriers	43	43
	Fatty acids and lipids	20	20
	Metabolic regulators	1	1
	Nucleosides and nucleotides	10	10
	Other	1	1
	Siderophore	1	1
Degradation/utilisation/ assimilation (76/81)	Alcohols	5	5
	Aldehyde	5	5
	Amines and polyamines	7	7
	Amino acids	15	16
	Aromatic compounds	1	1
	Carbohydrates	15	15
	Carboxylates	7	7
	Other	1	1
	Fatty acids and lipids	1	1
	Inorganic nutrients	9	9
	Nucleosides and nucleotides recycling	4	4
	Secondary metabolites	8	12
Detoxification (3/3)	Acid resistance	1	1
	Methylglyoxal	3	3
Generation of precursor metabolites and energy (27/27)	Chemoautotrophic energy	1	1
	Fermentation	2	2
	Glycolysis	2	2
	Pentose phosphate pathways	3	3
	Respiration	11	11
	TCA cycle	5	5
Superpathways		*43*	*43*

Only classes present in StyCyc 7.0 and StmCyc 4.0 are shown. Some pathways may be present in more than one sub-class category. Sty, Typhi; Stm, Typhimurium; numbers indicate how many pathways of this class/sub-class are present in each database. Superpathways represent overviews of connected pathways.

2.3.2.5 Metabolic map: Typhi

A ‘cellular overview’ of Typhi metabolism was generated for the automatically created StyCyc 1.0 (Figure 2-3) and, following extensive pathway curation and pathway hole filling, for StyCyc 7.0 (Figure 2-4). These depict a Gram negative cell with inner and outer membranes, across which transport reactions are shown. By EcoCyc convention, biosynthetic pathways are drawn on the left, then energy pathways, with degradation pathways on the right. Independent metabolic reactions that do not form known pathways are shown on the far right. Figure 2-4 therefore, represents the current metabolic map of Typhi CT18.

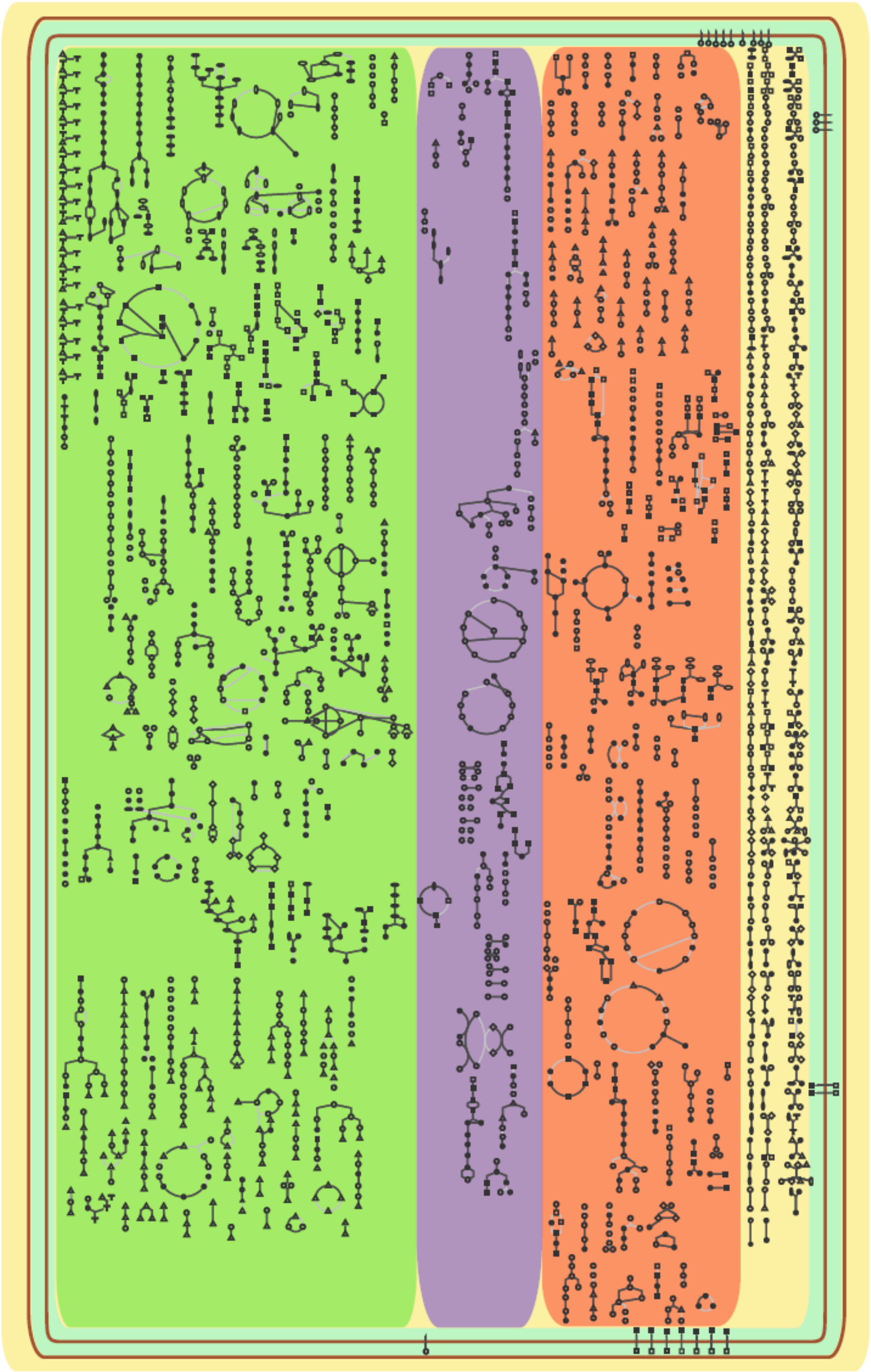

Figure 2-3 StyCyc 1.0: an automatically generated metabolic map of Typhi

Brown lines indicate bacterial cell membrane; black lines represent metabolic reactions. Green background, biosynthetic pathways; dark blue, energy pathways; red, degradation pathways; light blue, transport reactions. Symbols: upward-pointing triangle, amino acids; square, carbohydrates; diamond, proteins; vertical oblong, purines; horizontal oblong, pyrimidines; downward-pointing triangle, cofactors; T, tRNAs; open circle, other; enclosed circle, phosphorylated

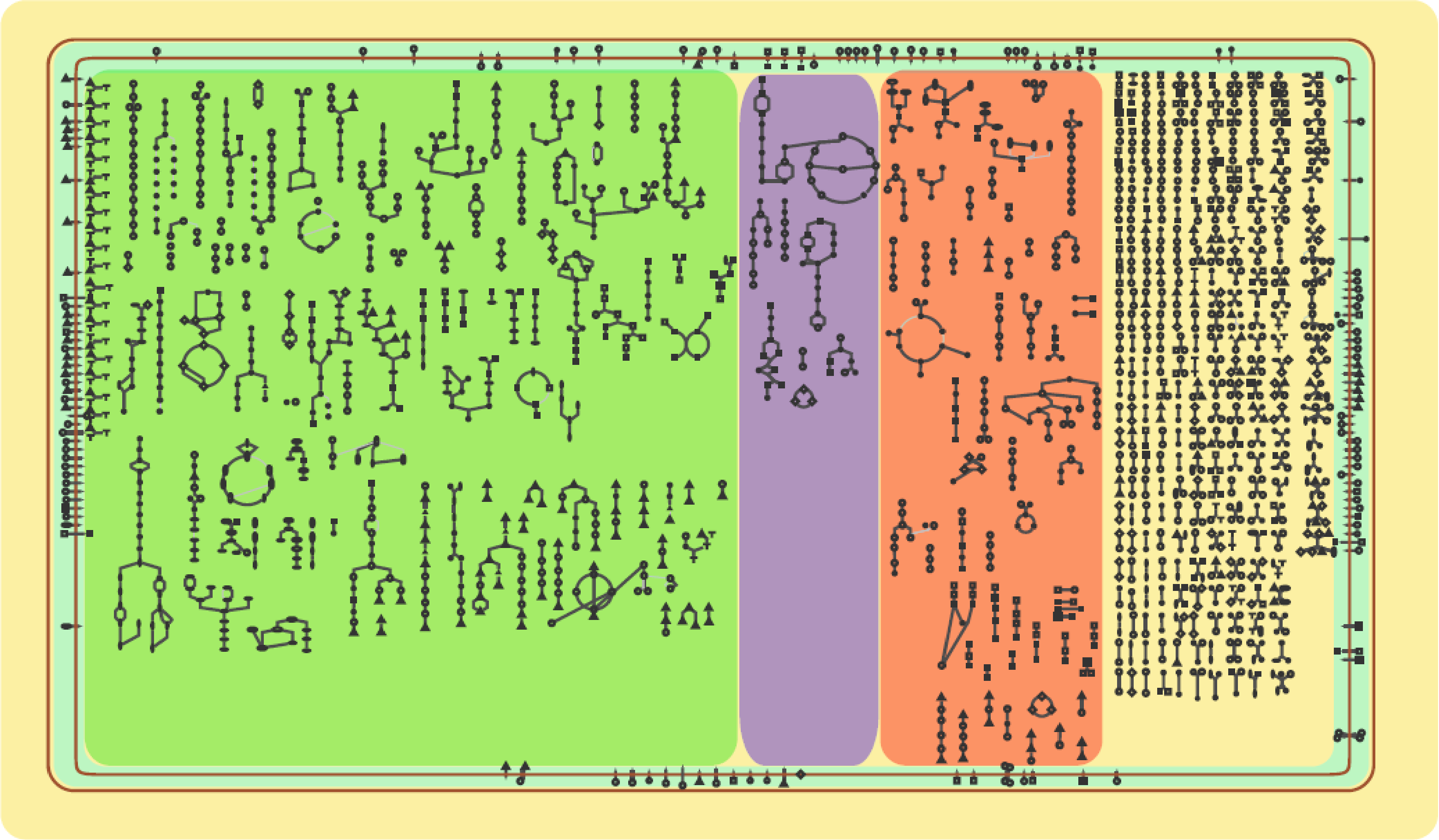

Figure 2-4 StyCyc 7.0: a manually curated metabolic map of Typhi

Brown lines indicate bacterial cell membrane; black lines represent metabolic reactions. Green background, biosynthetic pathways; dark blue, energy pathways; red, degradation pathways; light blue, transport reactions. Symbols: upward-pointing triangle, amino acids; square, carbohydrates; diamond, proteins; vertical oblong, purines; horizontal oblong, pyrimidines; downward-pointing triangle, cofactors; T, tRNAs; open circle, other; enclosed circle, phosphorylated.

2.3.2.6 Pseudogenes in Typhi

After comparison across multiple strains of Typhi, 211 pseudogenes were identified in Typhi CT18 alone, seven more than noted in the original annotation. However, the availability of 19 fully sequenced Typhi strains allowed pseudogenes from this serovar to be classified as ‘core’ (present in all) or ‘variable’ (present in one or more strains). In total, 274 pseudogenes were identified in all Typhi sequences. Using the metabolic map as a framework, the enzymatic reactions performed by these pseudogene ‘products’ were highlighted (Figure 2-5). Fifty seven appear in the overview, including 6 in biosynthetic pathways, 10 in degradation pathways and 20 in standalone reactions. Pseudogenes also account for 20 out of 133 transport reactions. The percentage of metabolic reactions attributed to pseudogenes is shown in Table 2-8.

Table 2-8 Classification of Typhi pseudogenes among metabolic reactions

Class	Total reactions	Pseudogene ‘products’			Possible complementary enzyme(s)
		Core	Variable	% of total	
Biosynthesis	590	4	2	1	1
Energy	103	1	0	1	1
Degradation	211	5	5	4.7	2
Standalone reactions	282	11	9	7.1	7
Transporters	133	18	2	15	-
Regulators*	238	7	4	4.6	-

A reaction was considered to have complementary enzyme(s) when the pseudogene ‘product’ was not the only enzyme assigned to that reaction. -, not determined. * Regulators are not displayed in the metabolic overview; these were determined from genome annotation.

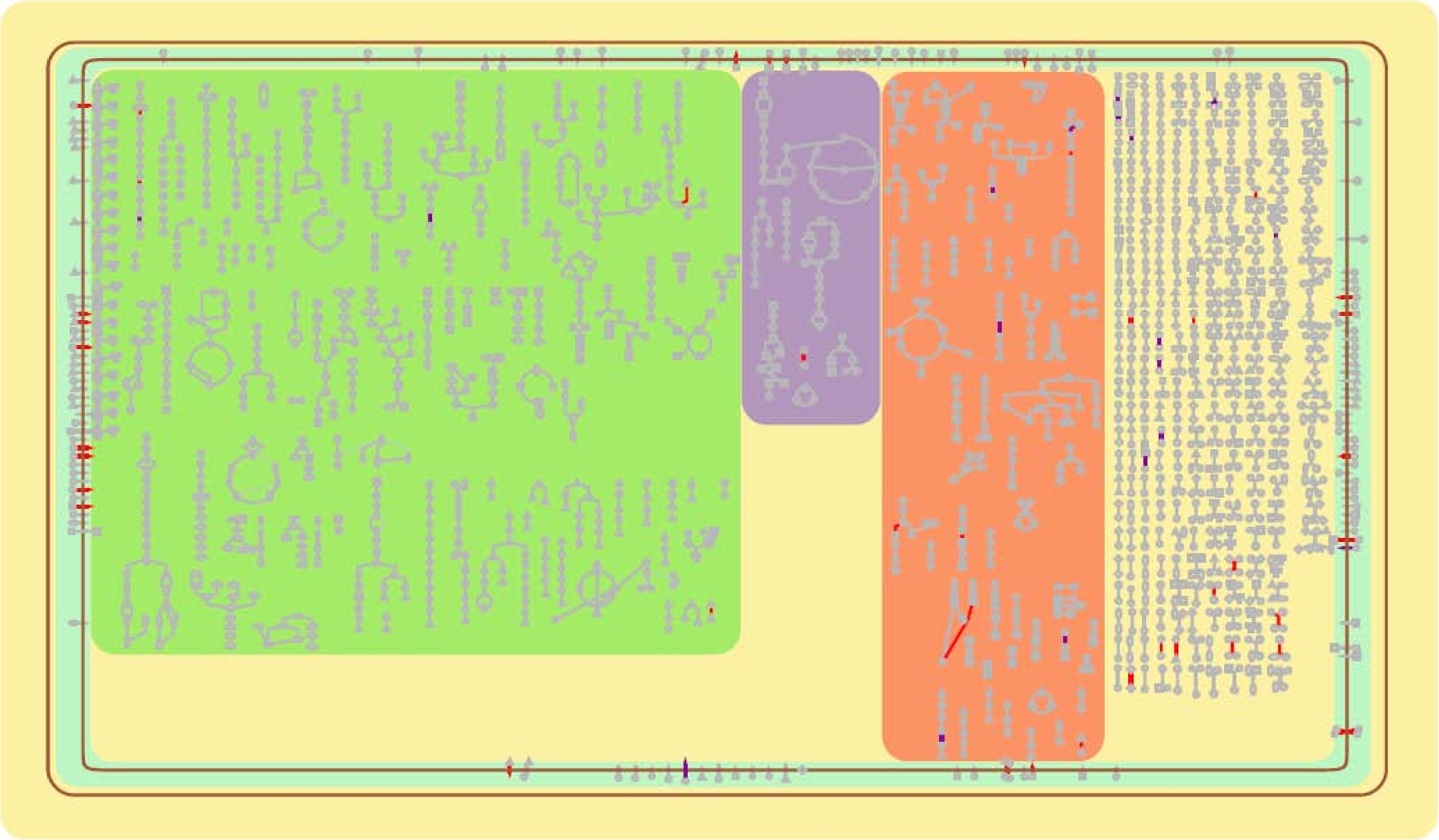

Figure 2-5 Pseudogenes interrupting metabolic pathways and transport reactions

Brown lines indicate bacterial cell membrane; red lines indicate core pseudogene ‘products’, purple lines are variable pseudogene ‘products’; greyed out lines represent metabolic reactions. Green background, biosynthetic pathways; dark blue, energy pathways; red, degradation pathways; light blue, transport reactions. Symbols: upward-pointing triangle, amino acids; square, carbohydrates; diamond, proteins; vertical oblong, purines; horizontal oblong, pyrimidines; downward-pointing triangle, cofactors; T, tRNAs; open circle, other; enclosed circle, phosphorylated.

2.3.3 StmCyc

2.3.3.1 Use of StyCyc to help assignments

The value of having already built StyCyc resulted in a much faster process for the initial stages of the Typhimurium database, StmCyc. The local enzyme mapping file contained extra *Salmonella*-specific information from StyCyc that would not have been found in EcoCyc or MetaCyc and hence the number of enzymes to manually assign was reduced from ~ 400 in StyCyc to ~280 in StmCyc.

This automated build took place approximately 11 months after StyCyc 2.0, during which time both EcoCyc and MetaCyc had been continuously updated with new metabolic pathways and reactions. Hence, more pathways were predicted to be present in StmCyc than had been in StyCyc 2.0. Further curation of StyCyc was performed to take this into account and is detailed in section 2.3.3.3.

2.3.3.2 Hole Filling

A semi-automated run of Pathway Hole Filler was performed, with hole-filling candidates predicted for 99 incomplete pathways. Sixty of these contained a single pathway hole, with two long pathways that had seven holes (Table 2-9). The ratio of deleted and filled pathways is shown in Table 2-10.

Table 2-9 Incomplete pathways in StmCyc

Holes	Number of pathways	Examples
1	60	arginine degradation; galactitol degradation
2	24	lyxose degradation; NAD biosynthesis
3	8	flavin biosynthesis; ketogluconate metabolism
4	5	ubiquinone biosynthesis; methylcitrate cycle
7	2	adenosylcobalamin biosynthesis; purine nucleotides biosynthesis

StmCyc pathways contained between 1 and 7 pathway holes before assignment of hole-filling candidates.

Table 2-10 Summary of StmCyc pathway outcomes

	Pathways			
Holes	Deleted	Completely filled	Partly filled	Not filled
1	25	26	0	7
2	10	13	0	1
3	2	6	0	2
4	0	4	1	1
7	0	2	0	0

Pathways without sufficient evidence were deleted from the PGDB. Hole-filling candidates were manually assessed for remaining pathways and assigned only with evidence as described in the methods.

2.3.3.3 Additional development

Many new pathways were predicted in StmCyc that were not present in the relevant databases at the time StyCyc was built. In order to have two *Salmonella* databases comparable in terms of the predicted pathways, these new pathways were assessed for their suitability for transfer into StyCyc. ACT comparisons and literature searches were used where necessary to inform this process. Nineteen new biosynthetic pathways and eight biosynthetic superpathways were transferred from StmCyc and assigned the relevant enzymes in StyCyc. A further ten degradation pathways, including one superpathway, were added, alongside eleven reactions relating to electron transfer.

A global comparison of pathways present in each PGDB highlighted some pathways that had not been predicted in StmCyc that were present in StyCyc. Some of these had been manually curated in StyCyc, and twelve were subsequently added to StmCyc. The two databases are now curated to the same level and are fully comparable across pathways present in both.

Of the forty pseudogenes annotated in Typhimurium SL1344, four appear in the cellular overview: *caiB* (degradation: carnitine), *cusA* (transport: cation), *mdaA* (standalone: nitroreductase) and *appC* (standalone: cytochrome oxidase subunit). None of these are pseudogenes in the laboratory strain Typhimurium LT2, thus they are potential ‘variable’ pseudogenes in Typhimurium.

2.3.3.4 PGDB statistics

The current version is StmCyc 4.0 which contains 204 metabolic pathways and encompasses 1,133 enzymatic reactions. The summary statistics are shown in Table 2-11.

Table 2-11 Statistics from StmCyc 4.0

Pathways	204
Enzymatic reactions	1133
Transport reactions	134
Polypeptides	4536
Protein complexes	137
Enzymes	1119
Transporters	169
Compounds	870

2.3.3.5 Metabolic map: Typhimurium

Following the side-by-side curation with StyCyc, a metabolic map for Typhimurium was generated (Figure 2-6). The map follows the same visual conventions as StyCyc and EcoCyc.

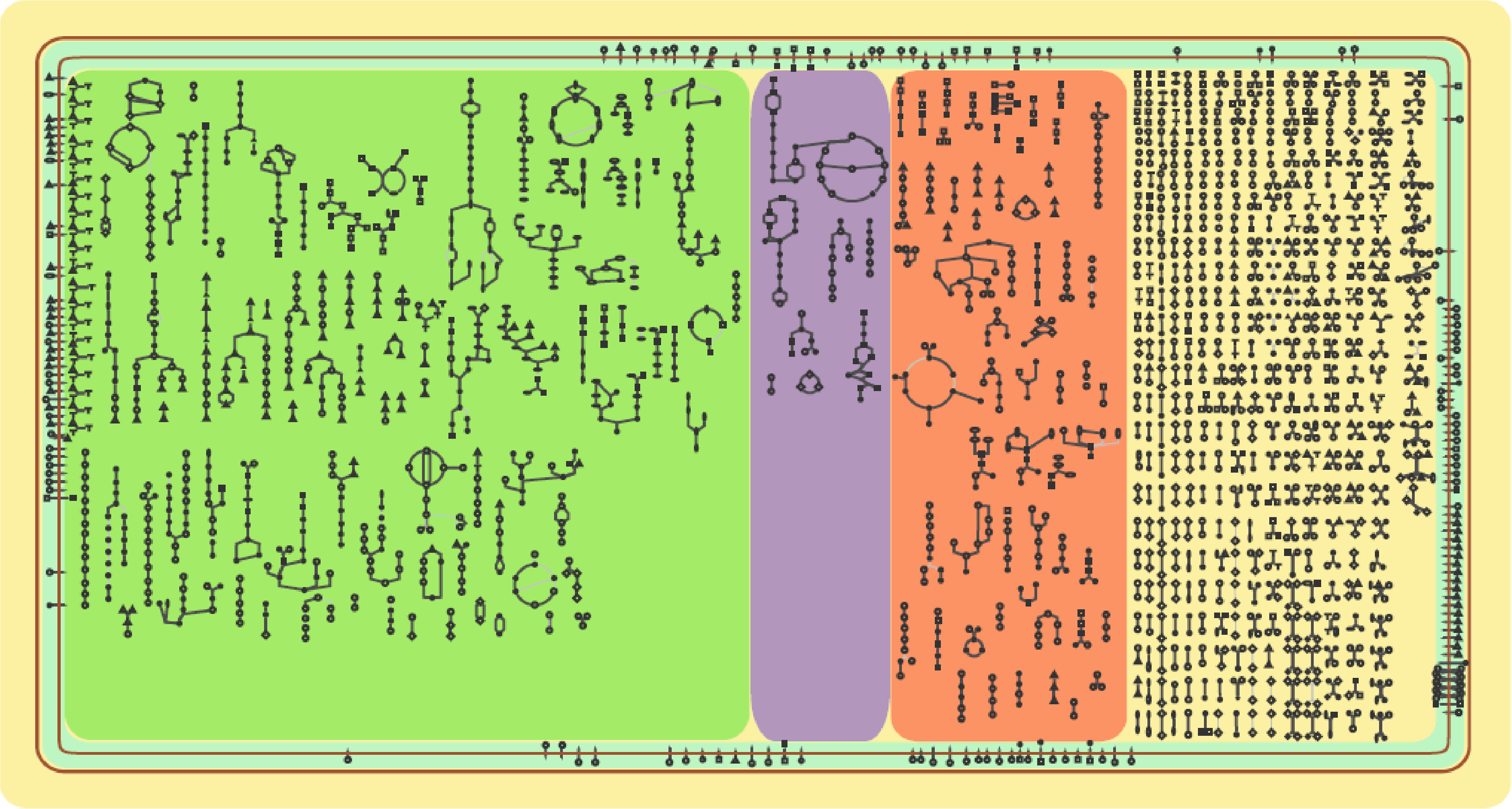

Figure 2-6 StmCyc 4.0: a metabolic map of Typhimurium

Brown lines indicate bacterial cell membrane; black lines represent metabolic reactions. Green background, biosynthetic pathways; dark blue, energy pathways; red, degradation pathways; light blue, transport reactions. Symbols: upward-pointing triangle, amino acids; square, carbohydrates; diamond, proteins; vertical oblong, purines; horizontal oblong, pyrimidines; downward-pointing triangle, cofactors; T, tRNAs; open circle, other; enclosed circle, phosphorylated.

2.3.4 Comparing StyCyc with StmCyc

Side-by-side curation of the two PGDBs facilitated the identification of pathway differences between them. Metabolically, they differ by the presence or absence of 5 pathways, one biosynthetic and four degradative (Table 2-12). They also differ in the number of intact pathways, with StmCyc maintaining 11 intact pathways that in StyCyc contain ‘core’ pseudogenes (Table 2-13).

Table 2-12 Metabolic pathway differences

Pathway	Present in StyCyc	Present in StmCyc	Genes
Vi antigen biosynthesis	✓	✗	*tvi*
L-idonate degradation	✗	✓	*idn*
Ketogluconate metabolism	✗	✓	*idn*
myo-inositol degradation	✗	✓	*dgo*
Glutamine degradation II	✗	✓	SL1455

Pathways present or absent in StyCyc and StmCyc as indicated by cross or tick.

Table 2-13 Metabolic pathways interrupted in Typhi

Pathway	Typhi pseudogene(s)
adenosylcobalamin biosynthesis (early cobalt insertion)	*cbiC*, *cbiJ cbiK*
allantoin degradation (anaerobic)	*allA*
arginine degradation / ethylene glycol degradation	STY1536
asparagine degradation	STY4203
hydrogen oxidation I (aerobic)	*hyaA*, *hyaB2*
4-hydroxyphenylacetate degradation	*hpaB*, *hpcC*
N-acetylneuraminate and *N*-acetylmannosamine degradation	*nanE*
putrescine biosynthesis	*speC*, *speF*
rhamnose degradation	*rhaD*
trehalose degradation	*treA*

Pathways intact in StmCyc but interrupted in StyCyc by the presence of single or multiple pseudogenes.

2.4 Discussion

2.4.1 Rationale for building a pathway database

Both human-restricted serovars Typhi and Paratyphi A display reduced metabolic profiles as well as significant genome degradation. While individual pseudogenes in these serovars have been previously identified, their global effect upon metabolism has not. Only by looking at pseudogenes in the context of metabolic pathways and transport reactions can wider comparisons be made and inactivating mutations of the same pathway be recognised. Building a metabolic pathway database provides the framework required for such a comparison, and having PGDBs customised for both Typhi and Typhimurium lends great scope for the comparison of the metabolic effect of pseudogenes, across host restricted *Salmonella*.

2.4.2 Rationale for using Pathway Tools

Prior to building a metabolic map for Typhi, the various tools for pathway visualisation and prediction were assessed. KEGG was suitable for visualisation purposes, as it was possible to view the organism-specific set of metabolic pathways. However, there were a number of factors that led to the decision to discount KEGG as a possibility: firstly, when organism-specific pathways were visualised, this was in the context of all the other pathways in the reference diagram, hence there was no way of looking solely at a Typhi-specific metabolic map; secondly, the presence of pseudogenes in pathways was not clearly denoted and there was no way to overlay this data on the framework; lastly, and most importantly, KEGG generates static reference diagrams, and there were no editing tools available to tailor pathways in any way.

The second possibility was Cytoscape, which is a more abstract method for looking at metabolism as it is oriented towards interaction data. However, this is most powerful when various types of interaction data can be used and these weren’t available at the time for Typhi.

The methods implemented by other software tools often use either KEGG or BioCyc as references for their own predictions. Pathway Tools software is part of the BioCyc suite and appeared well-suited to the task. Unlike some software, it does requires an annotation file, but since Typhi had been annotated to a very high level, this was not an issue. Another positive was that this would mean a pathway database built on the same framework as the extensively curated EcoCyc. Where there was overlap between *E. coli* and *Salmonella*, EcoCyc could therefore be used to speed up curation. The knowledge that the software was already installed and functional at the Institute also factored into the decision to use Pathway Tools to build pathway databases for Typhi and later Typhimurium.

2.4.3 Presence of pseudogenes can explain metabolic capability

For the first time, all Typhi pseudogenes affecting metabolic functions could be viewed simultaneously as part of the metabolic map. Data from 19 Typhi sequences were used to determine whether pseudogenes were core to all strains or had only occurred in some (variable). All strains sequenced came from patients with typhoid fever, suggesting that any variable pseudogenes are giving information on genes and pathways whose functional status is not relevant during the infective process.

Of the core pseudogenes, there are a number that provide a genotypic basis for a known phenotype. In the API20E system, Typhi is negative for the metabolism of rhamnose. StyCyc shows that rhamnose degradation is carried out by the *rha* genes, which catalyse 3 reactions to generate L-lactaldehyde and dihydroxyacetone phosphate. The latter feeds into glycolysis, but the final enzymatic step required to produce this metabolite is catalysed by *rhaD*, which is a pseudogene in Typhi, due to an early frameshift. Hence Typhi cannot utilise rhamnose as a sole carbon source.

A weakly positive production of hydrogen sulphide during growth on thiosulphate is also used to identify Typhi in biochemical testing. StyCyc indicates that the genes *ttrABC* encoding tetrathionate reductase are all intact, so a search for regulators found that this is likely due to the pseudogene present in *ttrS*, the sensor element of the *ttrRS* 2-component regulator that controls *ttrABC*.

Production of vitamin B12 has been identified previously as inactive in both Typhi and Paratyphi A, and StyCyc confirms that multiple *cbi* pseudogenes interrupt the biosynthetic pathway. Thus, StyCyc has identified the genetic basis for known lesions in Typhi metabolism, validating the PGDB for use in determining the effects of other pseudogenes upon metabolic capability.

Perhaps the most striking effect seen in StyCyc is the number of Typhi pseudogenes that occur in transport enzymes and complexes. Fifteen percent of transport reactions are encoded by pseudogenes or complexes that contain pseudogenes. The majority of the affected transporters are predicted to promote the passage of sugars from the environment into the cell. This has implications for the range of sugars Typhi is able to take up and utilise, with the possibility that only a small subset of sugars is available during human infection, hence other transport systems are no longer required. A knock-on effect if a substrate can no longer be imported is that flux through the metabolic pathway for utilisation of that substrate may be vastly reduced, leading to compensatory flux changes elsewhere.

2.4.4 StyCyc versus StmCyc: a basis for comparison

The eventual side-by-side curation of StyCyc and StmCyc allowed any pathway differences to be easily identified and explained. Currently in StyCyc 7.0, there is only one biosynthetic pathway present that is not found in Typhimurium; Vi antigen biosynthesis. The *tvi* genes required to produce the Vi antigen are encoded by the *viaB* locus of SPI-7, which has never been found in Typhimurium. In contrast, StmCyc 4.0 contains four additional pathways for the degradation of secondary metabolites, with respect to Typhi. Two of these are related: L-idonate degradation and ketogluconate metabolism, both of which involve *idn* genes no longer found in Typhi, due to phage insertion. The ability to degrade myo-inositol is conferred upon Typhimurium by a genomic island just over 22 kb in length, encoding the *iol* gene cluster (Kroger and Fuchs 2009). This island is only otherwise found in Paratyphi B and two other rare *Salmonella* serovars. Another region unique to Typhimurium encodes five *dgo* genes that account for the transport and degradation of D-galactonate, via a different pathway to that found in non-enteric bacteria. Finally, Typhimurium encodes SL1455, which is predicted to function as a glutaminase catalysing the conversion of glutamine to glutamate, as an alternative to the glutamate synthase complex *gltBD* present in both Typhi and Typhimurium.

2.4.5 StyCyc and StmCyc as resources

Both PGDBs can be accessed via the Internet at http://pathways.genedb.org. At present, StmCyc is in use by external collaborators at the University of Birmingham to look at gene expression data in Typhimurium. StmCyc is also being used by internal collaborators to locate pseudogenes in metabolic pathways of host-adapted variants of Typhimurium, and to visualise gene expression.

In another internal collaboration, we have used the new genome sequence and annotation from *S. bongori* 12419 to generate a metabolic map modified from StyCyc 6.1. This provided a shorthand method to obtain a global view of metabolism in this *Salmonella* species.

2.5 Conclusions

The generation of metabolic maps for Typhi and Typhimurium has opened up new possibilities for the visualisation and interpretation of gene-based data in *Salmonella*. For the first time, the gene degradation apparent from the Typhi genome sequence was put into the context of metabolic pathways, allowing metabolic phenotypes to be associated with pseudogene genotypes. A side-by-side curation effort with both databases revealed which pathways were unique to each serovar, and showed that the major differences between the two were due partly to genes present in Typhimurium and not Typhi and partly to the interruption/inactivation of pathways and transport reactions in Typhi and not Typhimurium.

StyCyc and StmCyc provide a basis for metabolic comparisons not only between themselves, but also as a framework upon which data from multiple *Salmonella* serovars can be visualised and interrogated.