

Comparison of the Small Molecule Metabolic Pathways in *Escherichia coli* and *Saccharomyces cerevisiae*: Non-orthologous displacements, Gene Fusions and Protein Interactions

Oliver Jardine
Dept. Crystallography,
Birkbeck College,
Malet Street,
London WC1E 7HX
UK

o.jardine@mail.crysl.bbk.ac.uk

Sarah A. Teichmann
Dept. Biochemistry & Molecular Biology,
University College London
Darwin Bldg., Gower Street,
London WC1E 6BT
UK

sat@biochem.ucl.ac.uk

ABSTRACT

We determined the domain structure and protein families of the enzymes in small molecule metabolism in *Escherichia coli* and *Saccharomyces cerevisiae* (yeast) using a combination of structural assignments and sequence comparisons. This allowed us to compare the evolutionary relationships between the proteins in the two organisms to determine the extent of conservation in pathways that are present in both yeast and *E. coli*. Among the 48 pathways and 232 enzymes shared between the two organisms, we identified twelve cases of non-orthologous displacement, where the enzymes carrying out identical functions belonged to entirely different protein families. Among the majority of enzymes which are conserved, we studied whether the subunit composition and gene structure were the same, looking for cases of gene fusions or fissions. We found fourteen cases where there is a multifunctional enzyme in one organism carrying out functions that are catalysed by several proteins in the other organism. Ten of the multifunctional enzymes were in the eukaryote yeast as expected, but another four of these multifunctional enzymes were in *E. coli*. In multifunctional enzymes, one of the advantages is the co-localisation of multiple active sites. In order to gain an insight into the role of protein interactions in metabolic pathways, we analysed the experimental data on protein-protein interactions available for the enzymes in yeast. In agreement with the small number of gene fusions identified in yeast, the extent of physical protein interactions between the enzymes is also limited, and most of the cases observed are between enzymes that are within a few reaction steps of each other within a pathway.

1. INTRODUCTION

The prokaryote *Escherichia coli* and the unicellular eukaryote *Saccharomyces cerevisiae* (yeast) are separated by roughly three billion years of evolution, when the bacterial and archae/eukaryote division is thought to have taken place. During this time, there have been countless chances for the genes in the two organisms to diverge by mutation, to change gene structure by gene fusion or fission and to acquire new genes for a function

by horizontal transfer or functional displacement of one gene by another within a genome.

Until now, investigations of these evolutionary processes have been limited to individual instances or small sets of occurrences, mostly identified by sequence comparison methods [2,10]. Here we investigate, and to some extent quantify, the frequency of these processes in a complete set of pathways between two distantly related organisms. The large amount of information available about the pathways, functions and structures of enzymes in these organisms allows us to study the evolutionary processes in small molecule metabolism in *E. coli* and yeast. Such a comparison would be much less successful in any other pair of organisms due to lack of knowledge of their enzymes and pathways, but because *E. coli* and yeast are model organisms, they have been subject to very extensive experimental characterization of their genes and proteins, including the determination of their complete genome sequence.

Our approach is to use sequence and structural information to characterise the domains and evolutionary relationships of shared enzymes. The use of structural information together with powerful multiple sequence comparison methods provides us with a very complete picture of the protein families that the enzymes belong to, including very distant evolutionary relationships: with these methods, at least one domain could be assigned to 96% (*E. coli*) and 85 % (yeast) of the protein chains in each organism. These domains belong to 385 sequence and structural protein families in *E. coli* and 317 families in yeast. 48 out of the roughly 60 pathways in each organism are shared, and 232 enzymes, which can represent multiple polypeptide chains, are present in both organisms. After providing an overview of the enzymes, protein families and metabolic pathways in the two organisms, we will compare this set of orthologs to each other to find differences in protein families and gene structure. Having ventured into the co-localisation of enzymes in analysing the cases of gene fusion, we will discuss the pattern of protein-protein interactions that we see in the enzymes in yeast.

1. PATHWAYS, ENZYMES AND PROTEIN FAMILIES IN *E. COLI* AND YEAST

1.1 The set of small molecule metabolic pathways

The small molecule metabolic pathways and enzymes in *E. coli* and yeast were modified from those present in the KEGG database [5]. The format of the KEGG database makes it easy to draw parallels between organisms, but the reason we did not simply adopt the precise set of KEGG pathways and proteins is that the reconstruction of pathways in KEGG takes place purely on the basis of Enzyme Classification (EC) numbers [12]. The Enzyme Classification system describes the function of an enzyme in terms of four numbers organised in a hierarchical manner, and there can be several enzymes that have the same EC number but are different proteins present in different pathways. This means that enzymes can be assigned to erroneous pathways or potentially also the wrong position within a pathway. Therefore, we post-processed the set of KEGG pathways and enzymes through a combination of automatic and manual analysis and comparison to other databases such as MetaCyc [6] and ERGO [3] by removing some pathways all together and modifying others.

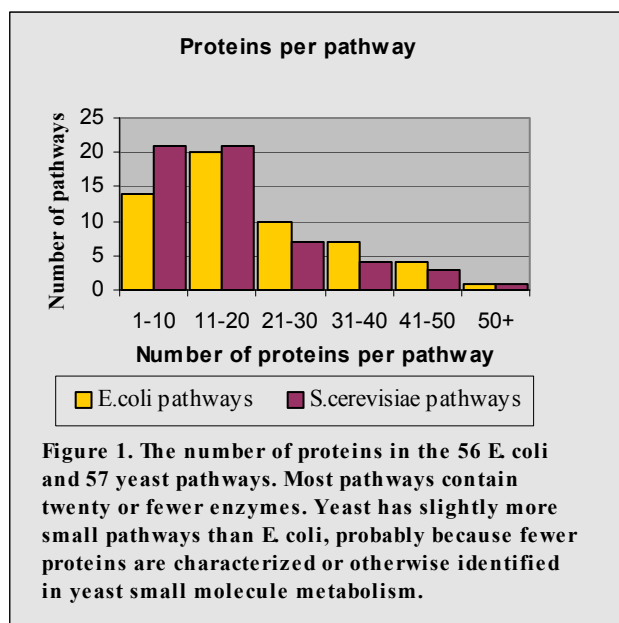
The final number of KEGG pathways was 56 in *E. coli* and 57 in yeast, as shown in Table 1. 48 of these pathways are shared, and this is the set of pathways we used in comparing the two organisms. The number of proteins in pathways varied between 2 and 72, and the distribution is shown in Figure 1. Most pathways contain under twenty proteins, and yeast has more small pathways, probably as a result of its metabolic pathways being less well characterised and hence fewer proteins are identified in the pathways.

Table 1. Overview of pathways, proteins and families

No. pathways		
<i>E.coli</i> : 56	<i>S.cerevisiae</i> : 57	Both: 48
No. proteins		
<i>E.coli</i> : 716	<i>S.cerevisiae</i> : 599	
No. total protein families ¹		
<i>E.coli</i> : 385	<i>S.cerevisiae</i> : 317	Both: 215
No. SCOP protein families		
241	199	159
No. PFAM protein families		
111	93	40
No. sequence families		
33	25	16
No. completely assigned sequences ²		
<i>E.coli</i> : 566	<i>S.cerevisiae</i> : 392	

¹ This includes structural families, PFAM families and clustered sequence families

² Protein sequences with no unassigned region greater than 30 residues in length



1.2 Determining the domain structure and family membership of enzymes

1.2.1 Structural domains

Most of the domains identified in the enzymes belong to structural protein families. The domain definitions and evolutionary relationships of the proteins of known structure are described in the Structural Classification of Proteins (SCOP) database [9]. In SCOP, domains are structural, but also evolutionary units, so a domain has to be observed on its own in a structure, or combined with several different domains, in order to be classified as a domain.

Gough *et al.* [4] used the domains from SCOP version 1.53 as seed sequences to build Hidden Markov Models. The iterative Hidden Markov Model method used is SAM-T99, described in Karplus *et al.* [7], with parameters optimised using SCOP as the gold standard for detection of distant evolutionary relationships. The database of Hidden Markov Models is available at: <http://stash.mrc-lmb.cam.ac.uk/SUPERFAMILY/>

These models were then scanned against the *E. coli* and yeast enzymes to identify domains. The family membership of the *E. coli* and yeast domains was inferred from the SCOP superfamily membership of the homologous SCOP domain.

In our set of 716 *E. coli* and 599 yeast proteins, 685 and 513 respectively had at least one structural domain assignment. Many proteins consist of multiple domains, so 868 domains in *E. coli* and 726 domains in yeast were found all together. The *E. coli* domains belong to 241 SCOP superfamilies and the yeast domains to 199 superfamilies, as shown in Table 1. The sizes of the SCOP superfamilies vary from one domain to 57 and 42 domains in *E. coli* and yeast

respectively belonging to the family of NAD(P)-binding Rossmann domains.

1.2.2 Sequence domains

The sequence regions in the enzymes that remained unassigned after the identification of SCOP domains were scanned against the Pfam database [1]. The Pfam database is a collection of Hidden Markov Models based on sequence alignments. Some of the Pfam families have a homolog of known structure, but many do not. 178 additional domains in *E. coli* and 135 domains in yeast could be identified in this way. The Pfam domains belong to 111 families in *E. coli* and 93 families in yeast, as shown in Table 1. 40 Pfam families are present in both *E. coli* and yeast.

Even after the Pfam search, some sequence regions longer than thirty residues remained without a domain assignment. We compared these sequences to each other with FASTA [14] and clustered them into families in the manner described in Park & Teichmann [13]. This way, another 33 families were identified in *E. coli* and 25 in yeast, with 16 of the sequence families present in both organisms. Because the evolutionary relationships in sequence families are not as distant as evolutionary relationships that can be detected on the basis of three-dimensional structure, a larger fraction of the Pfam and sequence families are specific to one organism.

1.3 Overview of domains and families in pathways

Taking together the SCOP superfamilies, Pfam families and sequence families, there are 566 (79%) out of 716 sequences in *E. coli* and 392 (65%) out of 599 in yeast that are completely covered by domain assignments. 119 (17%) sequences in *E. coli* and 121 (20%) in yeast have at least one domain assigned, but contain an unassigned stretch of residues longer than thirty. The number of single- and multi-domain proteins is shown in Table 2.

The single domain proteins are the proteins that consist of one domain that matches the whole sequence. There are 334 (47%) such proteins among the *E. coli* enzymes and 204 (34%) such proteins among the yeast enzymes. Therefore, the yeast enzymes tend to be slightly more complex in terms of their domain combinations than the *E. coli* enzymes. However, it should be noted that more than half of the *E. coli* enzymes consist of two or more domains, so even these ancient proteins in a prokaryote are the product of a combination of evolutionary units.

Table 2. Numbers of domains in proteins

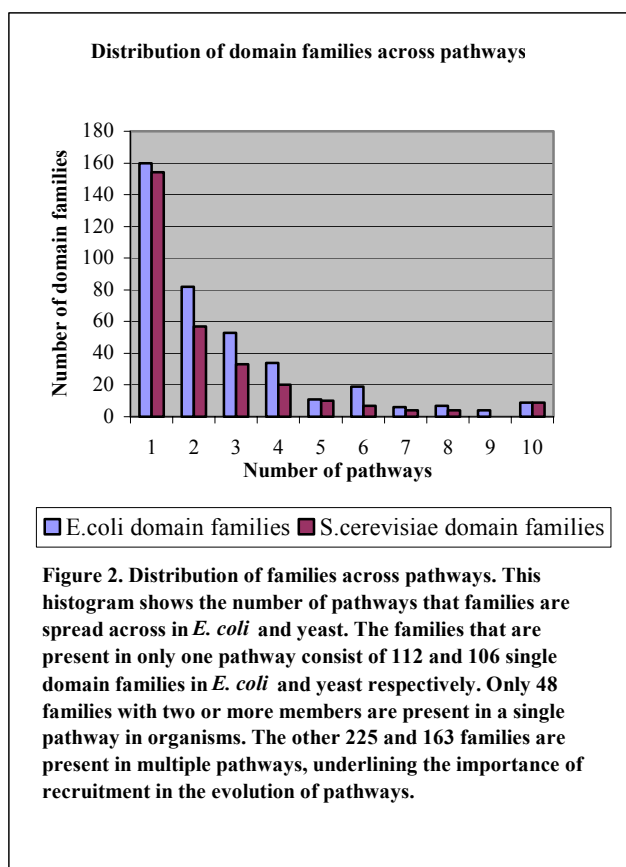
No. domains	<i>E. coli</i>		Yeast	
	Complete match	Partial match	Complete match	Partial match
1	334	74	204	80
2	152	37	123	25
3	50	5	46	8
4	19	3	10	3
5	7		5	4
6	3		3	
7	1			1
12			1	
Total	566	119	392	121

As mentioned above, the domains belong to families ranging in size from 1 to 42 (yeast) and 57 (*E. coli*) members within one organism. The distribution of the sizes of all three types of families is shown in Table 3.

Table 3. Family size distributions

Family size in no. domains	<i>E. coli</i>		<i>S. cerevisiae</i>	
	<i>E. coli</i>	<i>S. cerevisiae</i>	<i>E. coli</i>	<i>S. cerevisiae</i>
1	178	163	15	2
2	92	68	16	1
3	44	25	18	1
4	15	16	19	2
5	17	7	20	1
6	8	12	22	1
7	2	6	24	1
8	7	3	26	1
9	6		27	1
10	1	1	29	1
11		3	42	1
12	5	1	57	1
14	3	2		

Studying the way that the members of families are distributed across pathways within one organism would tell us about the evolution of the individual pathways in the last common ancestor. A detailed analysis of the patterns of families across pathways and the evolution of pathways in *E. coli* is provided by earlier work by Teichmann *et al.* [15]. Here we simply want to point out that of the families with more than one domain in either *E. coli* or yeast, the large majority of families are present in more than one pathway, as shown in Figure 2. This means that the recruitment of families across pathways, as proposed by Jensen in 1976 [8], was a major mechanisms in the evolution of these pathways.



2. COMPARING EQUIVALENT ENZYMES IN *E. COLI* AND *S. CEREVISIAE*: NON-ORTHOLOGOUS DISPLACEMENTS AND GENE FUSIONS

2.1 Identification of equivalent enzymes

The distribution of families across pathways gives an insight into the evolution of pathways, but here we want to focus on the conservation of pathways between the two very distantly related organisms *E.coli* and yeast. For the purpose of comparing the two organisms at the level of individual enzymes equivalent genes were identified and organised into pairs. Enzymes from *S.cerevisiae* and *E.coli* were allocated into pairs matching up genes by EC number and pathway. The domain architectures for the chains belonging to each enzyme of an equivalent pair were retrieved and compared to the domain architectures of the chains of the enzyme in the other organism. The results for the 232 equivalent enzyme pairs in *E. coli* and yeast are given in Table 4. Part of the reason for creating this set was to highlight where genes were functionally identical but structurally different. This set was also used to identify enzymes where the two organisms had different numbers of chains implicated in their function, therefore suggesting that gene fusion or fission may have taken place.

Table 4. Number of domains shared between the chains of the 232 equivalent enzymes in *E. coli* and yeast.

No. domains shared	No. ortholog pairs
0	20
1	131
2	59
3	13
4	5
5+	4
Total	232

2.2 Non-orthologous displacement

In Table 4, there are 20 cases where pairs of equivalent enzymes share no structural or sequence domains. These 20 cases were investigated more closely for evidence of non-orthologous displacement. This involved the retrieval of extra information from resources such as EcoCyc [6], MetaCyc [6], MIPS [11] and ERGO [3] to establish that the genes were genuinely functionally identical. Provided this was true, the protein chains involved were checked to ensure the absence of large unassigned sequence regions or confusions over equivalent domains with alternate names. In some cases, a chain in one organism was assigned a SCOP domain, whilst the chain in the other organism had a Pfam assignment. Whilst these chains appeared superficially to be unrelated, cross checking the SCOP assigned region with Pfam often found the two chains to be structurally equivalent and therefore evolutionarily related.

Table 5 lists the twelve cases of possible non-orthologous displacement that were found across all pathways out of a total of 232 ortholog pairs. They occur in 11 different pathways, representing just less than a quarter of all the pathways shared by the two organisms.

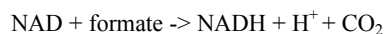
Table 5. Instances of non-orthologous displacements

No.	Pathway	EC number	Gene name(s) in <i>E.coli</i>	Gene name(s) in <i>S.cerevisiae</i>
1	Glycolysis/Gluconeogenesis	2.7.1.2	glk	glk1
2	Glyoxylate and dicarboxylate metabolism	1.2.1.2	fdoI, fdoH, fdoG	YPL275W, YPL276W
3	Sulfur metabolism	2.7.7.4	cysN, cysD	met3
4	Purine metabolism	4.6.1.1	cyaA	cyr1/cdc35/hsr1/sra4
5	Arginine and proline metabolism	4.1.1.17	speF/speC	spe1
6	Glycine, serine and threonine metabolism	4.2.1.13	sdaA, sdaB, sdhY	sdl1/cha1
7	Glycine, serine and threonine metabolism	3.1.3.3	serB	ser2
8	Aminosugars metabolism	2.7.7.23	glmU	qri1
9	Glycerolipid metabolism	3.1.4.46	glpQ, ugpQ	YPL206C
10	Phospholipid degradation	3.1.1.5	tesA	plb1
11	Porphyryn and chlorophyll metabolism	1.3.3.4	hemG	hem14
12	Riboflavin metabolism	2.7.7.2	ribF	fad1

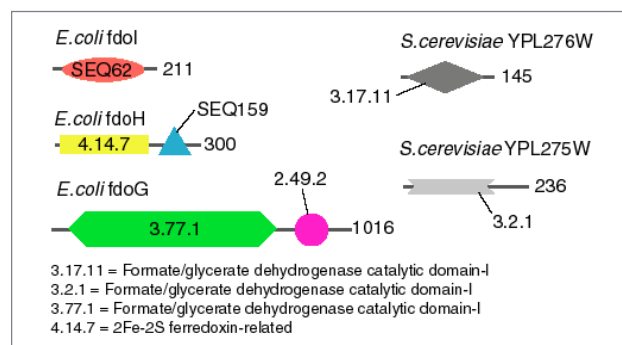
The structural details and biological explanations of two cases of non-orthologous displacement are given below as illustrations. The numbering of examples corresponds to that in Table 5

2. Formate dehydrogenase in Glyoxylate and dicarboxylate metabolism

This enzyme is involved in the metabolism of formate under anaerobic conditions. The reaction catalysed is:



The yeast chains originate from genes adjacent on the same yeast chromosome and make up a putative enzyme complex. The *E.coli* chains are subunits of the formate dehydrogenase complex.

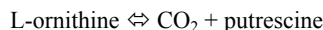


¹For explanation of diagram, see footnote.

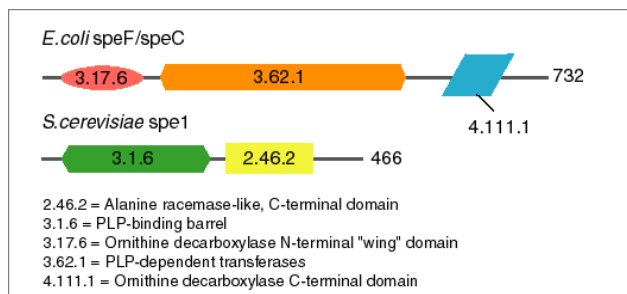
¹ The scale of the diagrams is consistent within but not across diagrams, due to limits of space. The horizontal grey lines are the polypeptide chains, with the coloured blocks representing the structural and sequence domains. The shapes and colours of blocks used do not conform to any greater key so the same structural domains can have two different shapes across diagrams. The lengths of the assigned domains are proportional to the coverage of the polypeptide chain that each represents. Each protein is labeled with its length in residues, organism and gene name (where alternate names exist they are separated by slashes). There is a key on each diagram to explain the domain labels. All SCOP superfamilies have their three number SCOP superfamily identifier, while PFAM families are given their name in PFAM and the sequence families are labeled with a number.

5. Ornithine decarboxylase in Arginine and proline metabolism

The reaction for this enzyme is:



The *E. coli* genes *speF* and *speC* are isozymes so share the same structure, but differ in their regulation. *SpeF* is the degradative



form and *speC* is biosynthetic.

Of the twelve cases we identified listed in Table 5, eight seem to be very likely examples of non-orthologous displacement whilst four are less sure due to incomplete structural assignments. It is clear from the absolute numbers involved (12 cases from 232 ortholog pairs – 5%) that the process of displacement by a non-orthologous enzyme is not a frequent occurrence in metabolic pathways. However, the fact that it occurs at all is in itself quite surprising, given the centrally important role played by the pathways of small molecule metabolism. A considerable degree of preservation of these systems is expected in the majority of organisms, and the interconnected nature and substrate specificity of the enzymes involved is likely to preclude their usurpation by unrelated proteins.

2.3 Gene Fusion

In the previous section, we have seen that there is extensive conservation of the enzymes in small molecule metabolism between *E. coli* and yeast. Now we turn our attention towards the conservation of gene structure and enzyme subunits in these two organisms. In *E. coli*, enzymes can be co-regulated through the use of operons, while in yeast operons do not exist, so that tight co-regulation could be achieved through gene fusion. There are also other advantages of having a single polypeptide chain that carries out multiple functions. The multifunctional enzyme usually carries out consecutive steps of a chain of reactions, so the product of one active site of the enzyme is immediately adjacent to the active site of the next step of the pathway, thus providing kinetic advantages, particularly with labile intermediates.

Since the majority of the enzymes in small molecule metabolism are single function enzymes, wherever a multifunctional chain exists, there is the possibility of a gene fusion event. Therefore, the set of orthologs were investigated to find the numbers of chains involved with each enzyme that was assigned multiple EC numbers, and the chains were compared in the two organisms.

We identified fourteen cases of gene fusion/fission with this system, listed in Table 6. The first four cases involve a single *E. coli* protein and multiple *S. cerevisiae* proteins. In these cases, the operon argument given above obviously does not apply. Instead, it seems more likely that gene fission of the multifunctional enzyme occurred in the eukaryote because it became advantageous to be able to regulate the individual components separately. In the other ten cases, the yeast enzyme is the multifunctional enzyme. In six of these cases, the *E. coli* enzymes are adjacent or close to each other on the bacterial chromosome, suggesting that they are co-regulated in *E. coli* in some way as well. In the four cases where the *E. coli* enzymes are far apart on the chromosome, the fingerprint of gene fusion is lost and it is not clear to what extent the individual enzymes are co-regulated.

Table 6. Gene fusion events ordered by gene structure

E. coli single multifunctional gene, *S. cerevisiae* multiple genes

No.	<i>E. coli</i>	<i>S. cerevisiae</i>
1	thrA	hom6, hom3
2	metL	hom6, hom3
3	pheA	pha2, aro7/osm2
4	igpd	his2, his3

S. cerevisiae single multifunctional gene, *E. coli* multiple genes adjacent on chromosome

No.	<i>S. cerevisiae</i>	<i>E. coli</i>
5	arg5,6/arg5/arg5 6	argC, argB
6	acc1/fas3	accB, accC, accD, (accA)
7	fas1, fas2	fabI, fabD, fabG, fabH/F/B
8	ura2	carA, carB, pyrB

S. cerevisiae single multifunctional gene, *E. coli* multiple genes close (within 10 genes) on chromosome

No.	<i>S. cerevisiae</i>	<i>E. coli</i>
9	thi6	thiE, thiM
10	gal10	galE, galM

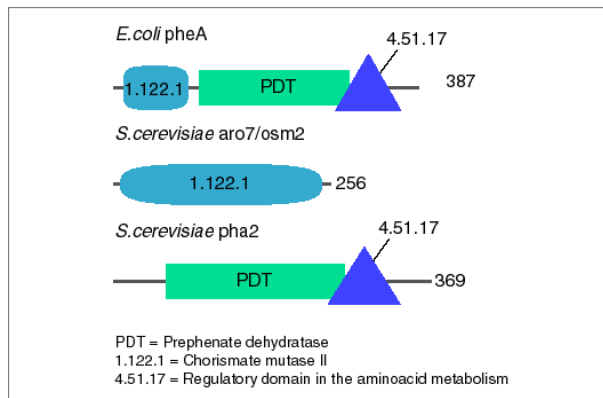
(continued overleaf)

***S.cerevisiae* single multifunctional gene, *E.coli* multiple genes further than 10 genes away on chromosome**

No.	<i>S.cerevisiae</i>	<i>E.coli</i>
11	aro1	aroA, aroL/K, aroD, aroB, (aroE)
12	his4	hisD, hisI
13	fol1	folK, ygiG, folP
14	ade5,7/ade57	purM, purD

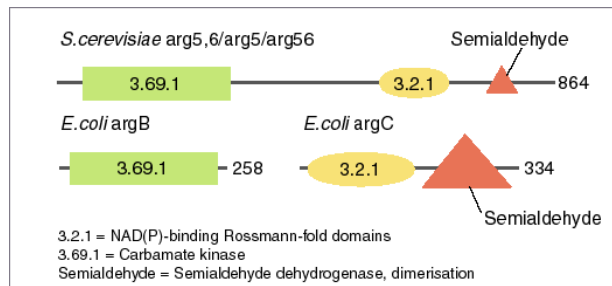
Below, we provide some examples of gene fusion events. The style of the diagrams is as for the cases of non-orthologous displacement above, and the numbering is as in Table 6.

3. The enzymes shown here are from the Phenylalanine, tyrosine and tryptophan biosynthesis pathway. The *E.coli* chain, pheA, has the functions of chorismate mutase-P and prephenate dehydratase. These functions are matched by the yeast chains - aro7 (chorismate mutase) and pha2 (prephenate dehydratase). The yeast chains are not known to physically interact although they are positioned consecutively in the pathway.

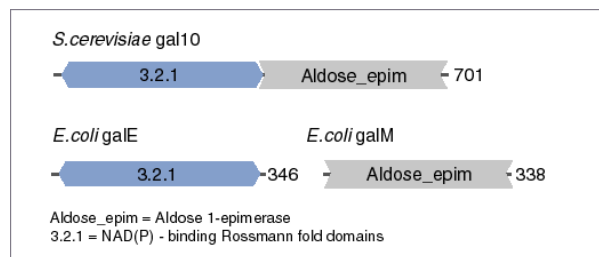


The discrepancy in the size (a difference of 165 residues) of the chorismate mutase domain between pheA and aro7 is interesting, suggesting it either became truncated during the fusion of the yeast chains, or possibly was expanded after the fission of the *E.coli* protein. The other domains involved have remained very similar in size.

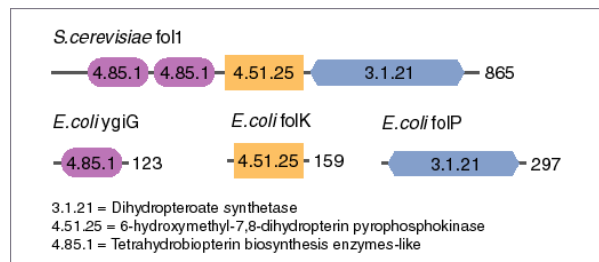
5. This is the first of the cases where the single, multifunctional chain is in yeast and the multiple chains are in *E.coli*. The yeast chain arg5,6 has the functions of N-acetyl-gamma-glutamylphosphate reductase and acetylglutamate kinase, while the *E.coli* proteins have the functions N-acetyl-gamma-glutamylphosphate reductase and acetylglutamate kinase for argB and argC respectively. All the enzymes are from the Urea cycle and metabolism of amino groups pathway. ArgB and argC are in the same operon in *E.coli* and are also binary relations (consecutive enzymes).



10. The enzymes here are present in three pathways - Glycolysis, Galactose metabolism and Nucleotide sugars metabolism. The functions of the enzymes are: gal10 - UDP-glucose 4-epimerase, galE - UDP-galactose-4-epimerase and galM - galactose-1-epimerase (mutarotase). The yeast enzyme is found in all three, however, unlike all the other cases given here, the two *E.coli* enzymes are never in the same pathway. GalE is only in Galactose metabolism and Nucleotide sugars metabolism whilst galM is only in Glycolysis. GalE and galM are close to each other on the bacterial chromosome, though not adjacent.



13. The enzymes in this example are all from folate biosynthesis. The yeast chain fol1 has the functions of dihydroneopterin aldolase, dihydro-6-hydroxymethylpterin pyrophosphokinase and dihydropteroate synthetase. YgiG is a putative kinase, folK is known as 7,8-dihydro-6-hydroxymethylpterin-pyrophosphokinase and folP is 7,8-dihydropteroate synthase. Given the structural similarity between ygiG and the first two domains of fol1 it seems likely that these two are functionally equivalent, making ygiG dihydroneopterin aldolase. The *E.coli* enzymes are consecutive in the pathway.



Taken as a whole, the cases of gene fusion identified in the data set indicate that, whilst the process is not common in metabolism, it does happen and usually involves consecutive enzymes in the same pathway. (There is only one exception to this in our data set, which involves two enzymes one step apart.) Furthermore,

where the component genes are in *E.coli* they are often in the same operon and therefore already adjacent on the chromosome. As to those cases where the component genes are not close on the *E.coli* chromosome, it may be that they have been relocated since the fusion event or were brought together in the eukaryote through transposition or another process.

The absence of operons in eukaryotes may explain the higher incidence of multifunctional chains in this group. Since operons are sets of co-regulated genes, the simplest way of creating this organisation in eukaryotes would be for the genes to fuse into one transcription unit. Interestingly, four of the cases involve single multifunctional chains in *E.coli* and multiple chains in yeast, the opposite to what might be expected. The existence of these enzymes suggests that gene fission may also be occurring, possibly with the advantage of independent regulation for enzyme subunits.

3. PROTEIN-PROTEIN INTERACTIONS IN YEAST

From the analysis above, gene fusions occur more frequently in *S. cerevisiae* compared to *E. coli*, and multi-subunit enzymes and protein complexes are known to be widespread in both *E. coli* and yeast. Using the large set of protein-protein interactions determined for yeast proteins experimentally, we surveyed the interactions between proteins within and across pathways in order to investigate to the role of protein interactions between the

enzymes in metabolic pathways aside from complexes and multi-subunit enzymes.

The raw data consisted of 7424 pairs of interacting yeast protein chains extracted from the MIPS database [11]. This database curates the published protein interactions in yeast from individual and large-scale experiments. The interactions are divided into genetic or physical interactions depending on the technique used to determine them. The genetic interactions relate to complementary mutations in two protein chains. The physical interactions are typically inferred from methods such as yeast-two-hybrid assays, affinity chromatography and so forth.

The set of interactions were parsed for protein identifiers that were in metabolic pathways. Of the original 7424 pairs, only 148 were found to consist of two protein chains where both were present in at least one pathway.

The 148 interactions where both chains were in pathways were subdivided into those where the participants were in the same pathway and those where they were in different pathways. These were then analysed at the level of EC number so the interactions could be compared to the pathways. For the interactions within pathways, there were 32 pairs of chains where both were part of the same enzyme or complex, or were isozymes or enzymes enzymes at the same step in a pathway of some other description. There were seven cases where the chains had different EC numbers and represented different steps in a pathway not known to be part of the same complex. A full summary of the results with the enzyme and pathway details is given in Table 7 below.

Table 7. Interactions between proteins at different steps within the same pathway not known to be members of a complex.

Consecutive enzymes (3 pairs):
<i>Glycoprotein biosynthesis:</i> G RHK1 (putative Dol-P-Man dependent alpha(1-3) mannosyltransferase ⇔ oligosaccharyl transferase glycoprotein complex
<i>Pyrimidine metabolism:</i> UPRTase ⇔ Uridine kinase
<i>Oxidative phosphorylation:</i> G ATP synthase subunit h ⇔ COX5B (Cytochrome-c oxidase chain Vb)
1 enzymatic step apart (3 pairs):
<i>Glycoprotein biosynthesis:</i> G dolichol-P-glucose synthetase ⇔ ... ⇔ components of oligosaccharyl transferase glycoprotein complex
<i>Pyruvate metabolism:</i> G carbon-catabolite sensitive malate synthase ⇔ ... ⇔ pyruvate carboxylase isozymes
<i>Oxidative phosphorylation:</i> G ATP synthase subunit h ⇔ ... ⇔ Ubiquinol cytochrome-c reductase subunit 8
2 enzymatic steps apart (1 pair):
<i>Starch and sucrose metabolism:</i> Hexokinase B ⇔ ... ⇔ ... ⇔ multifunctional trehalose-6-phosphate synthase/phosphatase complex

From Table 7, it is clear that the interactions between enzymes at different steps within the same pathway are all very close to each other, at most two reaction steps apart. The interactions across pathways also follow this pattern to some extent: five of the twelve genetic interactions across pathways are at junctions of pathways in the metabolic network, and hence are actually close

to each other in terms of reaction steps. The remaining seven genetic interactions across pathways are cases where the pathway assignment is unclear, so some of these may also be interactions between proteins close to each other in terms of reaction steps.

All together, there are 72 different combinations of interacting chains across pathways, but upon inspection many of the interactions determined by physical methods involved membrane proteins or proteins in different compartments. Due to the large number of potential errors, we only considered the twelve interactions found by genetic methods as mentioned above.

From the analysis of protein interactions within and across pathways in the yeast proteins, it seems as though the main function of these interactions is to co-localize proteins that are in reaction steps close in a pathway. The most likely reason for this is to decrease efficiency lost in diffusion of intermediates between enzymes. There are few genuine cross-pathway interactions, so it appears that physical protein interactions between enzymes are not a means for cross-talk or inter-pathway regulation of enzymes. Regulation of pathways can of course occur by gene regulation of enzymes, or regulation of enzyme activity at the protein level by activation or inhibition by small molecules or post-translational modifications.

4. CONCLUSIONS

Our comparison of yeast and *E. coli* small molecule metabolic pathways and enzymes shows that this central set of pathways is largely conserved in terms of pathways present and in terms of the domain architecture and gene structure of the enzymes. Among the set of 232 equivalent enzymes, there are only twelve cases of non-orthologous displacement and fourteen cases of gene fusion or fission.

Ten of the fourteen cases of gene fusion involve a multi-functional yeast enzyme and several individual *E. coli* enzymes. The imbalance toward the eukaryote may be due to the absence of operons, but the four cases of gene fission in yeast show that independent regulation of the enzymes in a pathway is sometimes advantageous.

Cases of gene fusion demonstrate that co-localization of enzymes in consecutive steps in a pathway is advantageous, and a survey of the protein interactions between yeast enzymes shows that a large fraction of these is also between enzymes that are either consecutive or very close to each other in a pathway in terms of reaction steps. Few protein interactions are apparent between enzymes where this is not the case, so enzyme regulation does not appear to take place by straightforward physical protein interactions between enzymes. Instead, control of pathways is known to occur by gene regulation, activation and inhibition by small molecules and post-translational modifications. Establishing the extent of conservation of small molecule metabolism at the level of gene regulation will soon be possible as more expression data for a variety of organisms becomes available.

5. ACKNOWLEDGEMENTS

We would like to thank J. Gough for structural assignments, and A. Shepherd and S. Rison for technical advice.

6. REFERENCES

- [1] A. Bateman, E. Birney, R. Durbin, S.R. Eddy, K.L. Howe and E.L. Sonnhammer. The Pfam protein families database. *Nucleic Acids Res.*, 28(1):263-266.
- [2] A.J. Enright, I. Iliouopoulos, N.C. Kyrpides and C.A. Ouzounis. Protein interactions maps for complete genomes based on gene fusion events. *Nature*, 402(6757):86-90, November 1999.
- [3] ERGO database. <http://igweb.integratedgenomics.com/WIT2/>
- [4] J. Gough, C. Chothia, K. Karplus, C. Barrett and R. Hughey (2000) Optimal Hidden Markov Models for all sequences of known structure. *Currents in Computational Molecular Biology*, p.124-125, eds Miyano, S., Shamir, R., Takagi, T, Universal Academic Press, Tokyo, Japan.
- [5] M. Kanehisa, and S. Goto, KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 28(1):133-135, January 2000.
- [6] P.D. Karp, M. Riley, M. Saier, I.T. Paulsen, S.M. Paley and A. Pellegrini-Toole. The EcoCyc and MetaCyc databases. *Nucleic Acids Res.*, 28(1):56-59.
- [7] K. Karplus, C. Barrett, and R. Hughey. Hidden Markov Models for detecting remote protein homologies. *Bioinformatics*, 14(10): 846-56, 1998.
- [8] Jensen, R.A. (1976) Enzyme Recruitment in Evolution of New Function. *Ann. Rev. Microbiol.*, 30, 409-425.
- [9] L. LoConte, B. Ailey, T.J. Hubbard, S.E. Brenner, A.G. Murzin and C. Chothia. SCOP: a structural classification of proteins database. *Nucleic Acids Res.*, 28(1): 257-259.
- [10] K.S. Makarova, L. Aravind, M.Y. Galperin, N.V. Grishin, R.L. Tatusov, Y.I. Wolf and E.V. Koonin. Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell. *Genome Res.*, 9(7):608-628, July 1999.
- [11] H.W. Mewes, D. Frishman, C. Gruber, B. Geier, D. Haase, A. Kaps, K. Lemcke, G. Mannhaupt, F. Preiffer, C. Schuller, S. Stocker and B. Weil. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, 28(1): 37-40, January 2000.
- [12] Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB) (1992) Enzyme Nomenclature, Academic Press, New York
- [13] J. Park and S.A. Teichmann. DIVCLUS: an automatic method in the GEANFAMMER package that finds homologous domains in single- and multi-domain proteins. *Bioinformatics*, 14(2):144-150, 1998.
- [14] W.R. Pearson and D.J. Lipman. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U.S.A.*, 85(8):2444-2448, April 1988.
- [15] S.A. Teichmann, S.C.G. Rison, J.M. Thornton, M. Riley, J. Gough and C. Chothia. The Evolution and Structural Anatomy of the Small Molecule Metabolic Pathways in *Escherichia coli*. *J. Mol. Biol.*, in press.