An Integrated Approach to Function Annotation in the Histidine Phosphatase Superfamily

Christopher John Roberts

MRes Advanced Biological Sciences
Institute of Integrative Biology
University of Liverpool

14th September 2015

An Integrated Approach to Function Annotation in the Histidine Phosphatase Superfamily

ABSTRACT

The histidine phosphatase superfamily is a large functionally diverse group of homologous proteins that during reactions become phosphorylated at a conserved catalytic core centred on a histidine. Although a diverse list of activities have already been discovered, the superfamily certainly houses uncharacterised novel functions. Many superfamily sequences cannot be annotated reliably with any known functions as they have little similarity to identified members and superfamilies present significant, unresolved challenges to automated genome annotation methods. Clustering experiments had previously revealed well defined groups that likely share the same function, for which no functional data are available. This study aimed to identify novel functions for members of the HP superfamily, focussing on large groups of hypothetical proteins. In this study structure function relationship knowledge was combined with a variety of bioinformatics methods, including genomic context, homology modelling and small molecule docking, to uncover novel functional annotations.

INTRODUCTION

The Histidine Phosphatase Superfamily Background

The histidine phosphatase (HP) superfamily is a large functionally diverse group of proteins, whose range of activities include contributing to metabolic pathways, both anabolic and catabolic, signalling and regulation. During reactions they become phosphorylated at a conserved catalytic core centred on a histidine (Rigden, 2008). The earliest discovered and most intensively studied member of the family, cofactor-dependent phosphoglycerate mutase (dPGM), is anomalous in catalysing a mutase reaction, as the rest of the family is composed of phosphatase activities. This history has caused problems of mis- and over-annotation by automated annotation tools as well as human interpretation of sequence relationships being biased to mutase rather than phosphatase activity (Rigden, 2008).

The superfamily is split into two branches with a distant evolutionary relationship that share limited sequence similarity. A RHG motif, in which the histidine is phosphorylated during catalysis, is shared near the beginning of the sequence (Bazan *et al*., 1989). The functions of the larger branch 1 with more bacterial sequences are much more diverse than that of the smaller branch 2, with more eukaryotic proteins containing mainly acid phosphatases (APs) and phytases. Branch 1 family members are found in the cytoplasm and nucleus, while branch 2 proteins appear to be secreted or remain in the endoplasmic reticulum, at the cell surface, periplasm or cell wall (Rigden, 2008).

The specificities of the enzymes deviate greatly from the small substrate of dPGM, phosphoglycerate (M_r = 187) to the substrate of SixA, the histidine-containing phosphotransfer (Hpt) domain of ArcB (M_r = 9050) (Rigden, 2008).

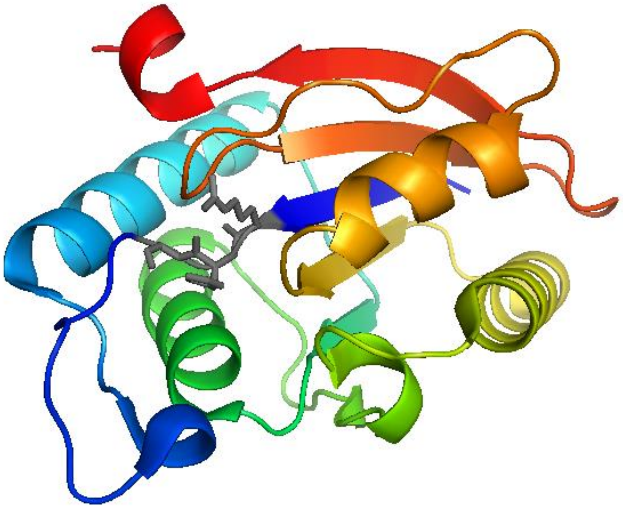


Figure 1: Cross-eyed stereo cartoon of *Escherichia coli* SixA (PDB code 1UJC). Structure is shown as a cartoon coloured rainbow except for grey showing the conserved RHG domain. Image generated using the PyMOL Molecular Graphics System, Version 1.3 Schrödinger, LLC.

SixA (Figure 1) has a near minimal core α/β domain (Hamada *et al*., 2005). Functions differentiate between lineages as the substrate binding site is defined by different insertions into the fold. The largest HP structure is almost three times bigger than the smallest due to significant N- and C-terminal extensions and the fact that the insertions can be as long as the core domain. The SixA structure can be hypothesised as resembling the ancestor of present members of the HP superfamily, with evolution adding insertions producing formations containing suitable cavities for binding smaller substrates. Rigden (2008) suggests that twice independently early on in evolutionary history the decoration of the basic fold occurred, due to the lack of structural or sequence similarity between insertions in the two branches and both branches being widely represented in eukaryotes and bacteria.

In the superfamily catalytic activity centres on phosphorylation and dephosphorylation of a histidine residue that follows the first β-strand (β1) of the fold (His8 in *E. coli* SixA). In the regions after strands β1, β2 and β4 the other residues contributing to the conserved catalytic core can be found. Proton donors are found following β1, β3 or β4. The catalytic cycle mechanism is shown in Figure 2.

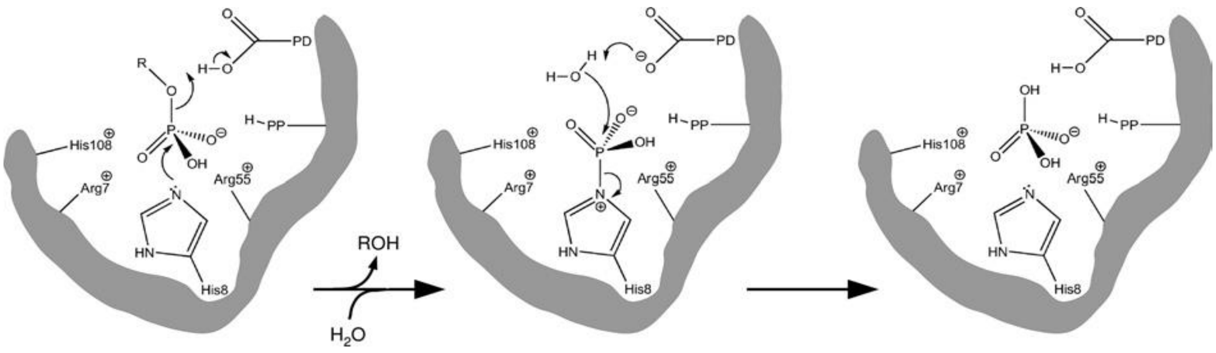


Figure 2: Histidine phosphatase superfamily catalytic mechanism. Showing the four invariant residues of the catalytic core as numbered in SixA of *E. coli*. During the course of the reaction His8 is phosphorylated. The "phosphate pocket" is formed from the other three residues that electrostatically interact with the phospho group before, during and after its transfer. PP represents additional residues (neutral or positive), which vary to a surprising extent, that may hydrogen bond to the phospho group contributing to the "phosphate pocket". PD shows the proton donor (an aspartate or glutamate) whose position varies in different families. From Rigden (2008).

Varying between lineages, additional "phosphate pocket" interactions are added that are contributed by residues lying in the β1–β2 loop. The residues that form the "phosphate pocket" hydrogen bond the phosphate group aiding the inline transfer of the phospho group from substrate to the enzyme (Wang *et al*., 2006). These include a pair of arginine residues (positions 7 and 55 in SixA) and another histidine, His108, which in known active members of the superfamily are completely conserved. Other residues are also highly conserved, between otherwise very diverse sequences, including Gly9 (SixA numbering) which forms part of the characteristic RHG motif and a L[S/T]XXG motif in the region between β1 and β2 (Rigden, 2008).

Scattered throughout the superfamily sequences, clustered in the loops following β1, β3 and β4, are varying specificity determining residues that bind the rest of the substrate. San Luis *et al.* (2013) identified an additional residue (Arg383) within the Sts (suppressor of T-cell signalling) phosphatases that is critical for catalytic activity and essential for Sts-1 to function as a negative regulator of T-cell receptor (TCR) signalling pathways. They found Sts-1 Arg383 is conserved in all Sts homologues and branch 2 APs however is lacking in the majority of proteins in branch 1, suggesting that despite the same catalytic core present in the two branches, there may be significant differences between the catalytic apparatus between the two branches (San Luis *et al*., 2013).

While the majority of the superfamily are phosphatases carrying out the mechanism in Figure 2, there are also the anomalous mutases: dPGMs and BPGMs. These catalyse three reactions, a phosphatase reaction follows the method in Figure 2, as well as a synthase and mutase reaction. The mutases can be distinguished from the phosphatases by their capacity to preserve intermediates bound while allowing their reorientation with the active site (Rigden, 2008).

Mutations or deficiencies in many human members of the HP superfamily lead to diseases, therefore there are important medical and clinical applications altering the activity of the enzymes (Hadjigeorgiou *et al*., 1999; Lemarchandel *et al*., 1992). HP superfamily members are found in various parasites and their inhibition is also of therapeutic interest (Slavin *et al*., 2002; Shakarian *et al*., 2003). Novel antibiotic production efforts may be aided by a better understanding of the roles of HPs involved in the synthesis of different classes of antibiotics (Huang *et al*., 2005; Palanichelvam *et al*., 2000). There are also current and potential future applications of several phosphatases in agriculture, with the addition of phytase to animal feed already in use (Kim *et al*., 2006).

Although the HP superfamily possesses a diverse list of activities, it certainly houses novel functions yet to be discovered. Many superfamily sequences cannot be annotated reliably with any known functions as they have little similarity to identified members. Well defined groups have been shown by clustering experiments (Rigden, unpublished) that probably share the same function, for which no functional data are available. With structure function relationship knowledge bioinformatics tools could be used to predict their functions (Rigden, 2008).

Computational Methods for Predicting Function

There has been a continuation in the exponential rate of growth of protein sequences in databases with the arrival of third generation sequencing technology, enhancing the sequence diversity of superfamilies such as the HPs (Mello *et al*., 2010). It would be a costly and time consuming task to experimentally determine the functions of all these proteins (Radivojac *et al*., 2013; Lee *et al*., 2007). Bioinformatics methods provide a helpful stopgap and direct laboratory experiments in annotating protein sequences with functions, from broad functional categories to specific catalytic activities (Mello & Rigden, 2012).

Protein function can be described at three different levels (Loewenstein *et al*., 2009), making it difficult to predict (Radivojac *et al*., 2013). Activity at the molecular level, such as catalysis, is described by molecular function, which is commonly predicted through identifying homologues or orthologues. Biological process considers broader functions, such as particular metabolic pathways, that are carried out by assemblies of molecular functions. Indirect functional associations and direct physical protein-protein interactions found in biological processes can be identified by genomic inference methods. The compartment of a cell in which a protein performs its function is described by the cellular component, which can be predicted by methods that predict post translational modifications, membrane association, residue composition or signal sequence (Lee *et al.,* 2007). Predicting protein function is also complicated as functions are context dependent and proteins are often promiscuous and multifunctional (Radivojac *et al*., 2013).

Inheritance through homology, the knowledge that proteins with similar sequences often have similar functions, is the most common approach to computational protein function prediction, although many of the mis- and over-annotations in databases are the result of inheritance through homology being used liberally (Lee *et al*., 2007). If they have descended from a common ancestral sequence protein sequences are considered homologous (Fitch, 1970). They are likely to perform a similar function at the molecular level as they have a similar three-dimensional (3D) structure (Teichmann, 2002).

A traditional approach for computational functional inference is searching databases of proteins with algorithms such as BLAST (Basic Local Alignment Search Tool) (Altschul *et al.*, 1990; Altschul *et al*., 1997), detecting homologues whose functions have been determined experimentally (Gabaldón & Huynen, 2004; Radivojac *et al*., 2013). By 2004 homology search sensitivity had more than doubled due to PSI-BLAST and other profile based methods (Gabaldón & Huynen, 2004). Another iterative search tool that uses Hidden Markov Models is Jackhmmer which is more sensitive and selective than PSI-BLAST (Li *et al*., 2012). In this project sequence based methods like these were useful for identifying new clusters but were not suitable for assigning functions as the aim was to identify novel functions. In this respect genomic context and modelled structures were used.

Evidence of Function from Genomic Context

Except for parasitic or symbiotic cases, being encoded in the same genome is required for proteins to interact. Regardless of their sequence similarity genes that are part of the same biological process have a tendency to occur in each other’s genomic context. Analysing sequences in their genomic context allows the identification of interacting proteins as well as the biological process in which they play a role, and thus provides information about their functional context (Gabaldón & Huynen, 2004).

The most direct form of genomic context are gene fusion events. Functionally this fusion may result in an enhancement of an interaction between their activities biochemically (Gabaldón & Huynen, 2004). Marcotte *et al*. (1999a) and Enright *et al*. (1999) showed that many gene fusions involved genes known to functionally interact

(Enright *et al*., 1999; Marcotte *et al*., 1999a). Gene fusion allows expanding the functional association of larger groups of interconnected genes.

Genomes are rapidly shuffled and rearranged over the course of evolution, although in prokaryotes some gene clusters are conserved (Gabaldón & Huynen, 2004). These genes tend to be part of the same operon (Moreno-Hagelsieb *et al*., 2001) and encode proteins that functionally interact (Dandekar *et al*., 1998; Overbeek *et al*., 1998). Gene neighbourhood methods use this organisation and extend to eukaryotes in which co-regulated interacting genes are occasionally found to cluster in the genome (Blumenthal, 1998; Teichmann & Babu, 2002). In 2003 Lee and Sonnhammer made the observation that genes involved in the same biochemical pathway tend to be clustered together in a variety of eukaryotic genomes (Lee & Sonnhammer, 2003).

Functional predictions can also be made when two genes are encoded in a large number of genomes, yet both are lacking from other genomes, using co-occurrence techniques, as studies have shown proteins that are distributed across species in a similar manor have a high propensity to interact functionally (Pellegrini *et al*., 1999). Another powerful sign of functional associations between proteins is conservation of co-expression, a proxy for co-regulation (Marcotte *et al*., 1999b), therefore functional interactions can also be predicted from genome wide microarray expression data (Gabaldón & Huynen, 2004).

Sequences can be visualized in their genomic context using the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) web server (Szklarczyk *et al*., 2011). STRING provides a critical assessment and integration of protein-protein interactions on a global scale including functional (indirect) and physical (direct) associations (Szklarczyk *et al*., 2015). STRING reports predicted functional partners for query proteins with a confidence score (Szklarczyk *et al*., 2011), based on interactions combined from several sources including *de novo* predicted interactions from algorithms using co-expression as well as genomic information, pathway connections imported from manually curated databases, experimental evidence derived from primary databases and automated text-mining of publications (Szklarczyk *et al*., 2015). STRING not only reports the level of genomic association between genes and therefore the intensity of the functional association between their products, but also additional information about the genes in their genomic context such as the gene order (Snel *et al*., 2000).

It is common for genomic context to provide clues of reaction type, but when used alone it is often not sufficient to specify the identity of the substrates. For this task homology modelling and subsequent *In silico* ligand docking are used (Gerlt *et al*., 2012).

Structure Based Functional Inference

As homologous proteins evolve their structure is frequently more conserved than their sequence (Chothia & Lesk, 1986). When sequences diverge beyond a degree of similarity that can be detected reliably using sequence comparison methods, structural information can be used to reveal proteins with similar function (Lee *et al*., 2007; Loewenstein *et al*., 2009).

Proteins that display structural similarity along their entire length of amino acid sequence are likely to have similar or the same function (Lee *et al*., 2007). It is paramount to take into account the number of residues in the alignment and the quality of the superposition when evaluating the significance of the similarity between two proteins (Lee *et al*., 2007). Transferring function from one protein to another should be undertaken cautiously as two proteins may have a similar fold but different functions (Loewenstein *et al*., 2009).

In highly variable superfamilies that have significant structural divergence, such as the HP's, new functions can evolve through insertion of secondary structure components as mentioned above (Reeves *et al*., 2006). They frequently accumulate producing a bigger structural motif or feature on the surface modifying the active site geometry or encouraging novel protein-protein interactions (Lee *et al*., 2007).

Using structure to predict function often uses global structure comparison, comparing the query protein structure to structure database domains (Loewenstein *et al*., 2009). Global structure comparison methods do not differentiate overall fold conservation and functionally relevant regions of the protein (Loewenstein *et al*., 2009) as small changes in an active or binding site can cause a divergence of function (Lee *et al*., 2007), and so were not relevant to this particular project. In this study it was more appropriate to analyse localised structural regions such as pockets and active site clefts to produce details on potential small molecule binding to suggest functions (Lee *et al*., 2007; Loewenstein *et al*., 2009).

To conserve a certain function through evolution the functional sites local environment must be conserved, even if other areas of the fold are modified. Enzymes create a specific chemical environment by isolating substrates in binding pockets or active site clefts, where catalysis is performed by a limited number of residues (Lee *et al*., 2007). It is the identity and special arrangement of these active site residues that determine substrate specificity (Gerlt *et al*., 2012). Some pocket centred approaches detect the preservation of physico-chemical properties, such as hydrophobicity and charge, of the binding sites amino acids in similar 3D conformations to describe protein ligand interactions allowing the identification of authentic functional homologues (Lee *et al*., 2007; Loewenstein *et al*., 2009).

In silico ligand docking methods capture information about differences in substrate specificity from the structures of the binding sites and residue substitutions within them (Gerlt *et al*., 2012). Interaction prediction through docking procedures predict where and how proteins interact rather than which proteins interact (Gabaldón & Huynen, 2004).

Previous Studies

The computational strategies proposed for assigning novel functions to members of the HP superfamily, such as genomic context, homology modelling and metabolite docking, have been previously successfully reported in the study of the enolase superfamily by Babbitt and colleagues (Gerlt *et al*., 2012; Lukk *et al*., 2012; Zhao *et al*., 2013). The enolase superfamily is also functionally diverse containing more than 8000 members (Gerlt *et al*., 2012) and more than 20 distinct substrates have been identified (Lukk *et al*., 2012). The active sites could be considered analogous to the phosphatase of the HP superfamily.

As the majority of the enolase superfamily are found in microbes (Neidhart *et al*., 1990) information provided by operon context was used for annotating some members with previously unknown functions (Gerlt *et al*., 2012). Homology modelling was used to obtain “dockable” structures using the template of an experimentally determined ligand structure. *In silico* ligand docking methods were then used to infer substrate specificity from the binding site structure. Libraries of potential and known metabolites were docked to predict which substrates are likely to bind along with their “poses” in the binding site. A rich diversity of substrate specificity for several uncharacterised groups was predicted and used to direct experimental testing of substrates. The biochemical studies subsequently confirmed that most of the key interactions between the active site and substrate were correctly predicted, resulting in the assignment of novel functions to members of the enolase superfamily (Gerlt *et al*., 2012; Lukk *et al*., 2012; Zhao *et al*., 2013).

Aims of the Current Project

The aim of the present study was to identify novel functions for members of the HP superfamily, focussing on large groups of hypothetical proteins. To achieve this, the first phase of the investigation was the collection of a complete set of HP sequences using the iterative search program Jackhmmer (Eddy, 1998; Finn *et al*., 2011). The second phase was to partition the collected sequences, using sequence similarity, into “clusters” that likely share the same function by utilising CLANS (Frickey & Lupas, 2004; Frickey & Weiller, 2007). As the emphasis was the prediction of novel functions, the next stage was to eliminate clusters of HPs with known functions by performing database searches. The final phase was to predict the functions of the remaining large clusters of HPs, using a variety of bioinformatics methods including genomic context, homology modelling and metabolite docking. STRING (Szklarczyk *et al*., 2015) was proposed for visualising genomic context, identifying predicted functional partners and determining potential ligands for subsequent docking, using the RosettaLigand (Combs *et al*., 2013), into homology modelling created with RosettaCM (Combs *et al*., 2013).

METHODS

Collecting a Complete Set of Histidine Phosphatase Sequences

The Jackhmmer program, from the Hmmer v3.1b1 package (Eddy, 1998; Finn *et al*., 2011), was used to iteratively search query protein sequences against the UniRef90 protein sequence database downloaded from http://www.uniprot.org/uniref/ on 12/02/15 (Bateman *et al*., 2015; Suzek *et al*., 2011). UniRef90 was chosen to remove redundancy, as it is built from clustering UniRef100 sequences at a 90% sequence identity level.

Jackhmmer's default of 5 iterations was increased to a maximum number of 20 iterations, to search UniRef90 with 8 query sequences from branch 1 representatives (Appendix 1). For each query searched, the number of sequences collected was plotted against the number of iterations. The files containing the collected sequences were compared, to assess if the sequences were the same in each collection, and manually searched for the presence of branch 2 proteins.

Clustering Sequences

CLANS software (Frickey & Lupas, 2004; Frickey & Weiller, 2007) was used to divide the collected sequences into clusters that likely share the same function. As the FASTA file generated by Jackhmmer contained a relatively large number (29,295) of sequences and was too big for CLANS to process, CD-HIT v4.6.1 (Fu *et al*., 2012; Li & Godzik, 2006) was first used (word size of 4) to generate a file with a 65% sequence identity cut-off, again removing redundancy, resulting in a FASTA file containing 10,548 sequences. For full functionality of CLANS, NCBI's BLAST 2.2.18 (Altschul *et al.*, 1990; Altschul *et al.*, 1997) complete with blastall and formatdb executables was installed. The FASTA file containing sequences generated by CD-HIT was used as an input for CLANS using an E-value of 1 and P-value of 0.1. In CLANS a cut-off value of 1e-40 was selected (specifying up to what E-value to take BLAST-hits into account for clustering) and clustering was run for 32739 rounds. CLANS was used to automatically detect clusters using an iterative "neural network based" approach (minimum sequences per cluster 2, maximum rounds 100).

Eliminating Clusters of Proteins with Known Functions

For the largest automatically detected clusters, database searches were performed to eliminate those with known functions. The HHpred server (Biegert *et al*., 2006; Söding *et al*., 2005) was used to search a representative query sequence from each cluster against the PDB (Berman *et al*., 2000; Bourne *et al*., 2004) and CD (Marchler-Bauer *et al*., 2015) databases, detecting clusters of sequences that have close similarities to protein structures where the function is better known. Clusters with relatively high sequence identity (>30%) to a reliable family or structure of known function, were not of interest in this project and so were discarded, as the same or similar function was assumed. The clusters of interest to this study had relatively low sequences identities (≤30%) to any family or structure of known function. As they were not clearly identifiable as anything in particular, no conclusion could be made about their functions. These clusters were retained for further investigation to determine novel functions by genomic context, homology modelling and small molecule docking. Other clusters were dismissed as they contained multiple large insertions in the sequence compared to the structure matches, therefore subsequent structure modelling in these sequence gap regions would be not much more than guess work

To detect activities not represented by structures or domain entries the blastp executable from BLAST-2.2.30+ (Camacho *et al*., 2009) with an E-value threshold of 0.01 was used to search the query sequences from each cluster against the reviewed Swiss-Prot database downloaded from http://www.uniprot.org/downloads on 17/06/15 (Bairoch & Apweiler, 2000; Bairoch *et al*., 2004; Bateman *et al*., 2015). Again clusters whose sequences had relatively high sequence identity (>30%), were dismissed as they were assumed to have the same or similar function and assigned the Swiss-Prot activity.

For the remaining clusters of interest, were no conclusion could be made about function, each sequence identifier was entered at http://www.uniprot.org/uploadlists/ (Bateman *et al*., 2015).The corresponding UniProt records were retrieve and examined to determine if they were reviewed (high quality manually annotated proteins, with information extracted from the literature, experimental results and curator-evaluated computational analysis) or unreviewed (automatically computationally annotated proteins awaiting full manual annotation) and find out what species each clusters sequences are from.

Detecting Domain Fusions

To detect any other domains fused to the HPs the RPS BLAST (Reversed Position Specific BLAST) executable from BLAST-2.2.30+ (Camacho *et al*., 2009) with an E-value threshold of 0.01 was used to search the query sequences from each cluster against the CD database of domains (Cdd_LE.tar.gz downloaded from ftp://ftp.ncbi.nih.gov/pub/mmdb/cdd/little_endian/ last modified 18/05/15; Marchler-Bauer *et al*., 2015).

STRING for Genomic Context

STRING v10 (Szklarczyk *et al*., 2015) was used to identify predicted functional partners and visualise genomic contexts. Protein sequences from each cluster were used to search the STRING database, restricting the organism to search for each cluster to the general groups identified from UniProt records described above (cluster 6564, insects; clusters 6563 and 6552 fungi; clusters 6560, 6545 and 6544 bacteria; cluster 6541, yeast). "Interactors wanted" was set to "COGs". Prediction methods selected for use were neighbourhood, gene fusion, co-occurrence, co-expression, experiments, databases and textmining. The confidence score was set to medium (0.400).

A bi-directional / reciprocal genome BLAST (Altschul *et al*., 1990; Altschul *et al*., 1997) was used to confirm the orthology of each clusters sequences (implying likely same function) to the most similar protein identified in the STRING database.

Detecting Signal Peptides and Predicting Localisation

Signal peptides were sought in three representative sequences for cluster 6564 (Accession numbers A0A026VZI3, L7M867 and W8BG64) using the SignalP 4.1 (Emanuelsson *et al*., 2007; Petersen *et al*., 2011) and Phobius (Käll *et al*., 2004; Käll *et al*., 2007) servers. The predicted subcellular localisation of the proteins was determined using the TargetP 1.1 (Emanuelsson *et al*., 2000; Emanuelsson *et al*., 2007) and PSORTII (Nakai & Horton, 1999) servers.

For cluster 6560 signal peptides were sought for a representative sequence (Accession number E5CGZ1) using the SignalP 4.1, Phobius, PrediSi - PREDIction of SIgnal peptides (Hiller *et al*., 2004) and Signal-3L (Shen & Chou, 2007) servers. To predicted the subcellular localisation of the proteins, and specifically distinguish bacterial secreted extracellular proteins from those localised and retained in periplasm, the following servers were used: PSORTb v3.0.2 (Yu *et al*., 2010), Gneg-mPLoc (Shen & Chou, 2010), SOSUIGramN (Imai *et al*., 2008), SLP-local (Matsuda *et al*., 2005), LocTree3 (Goldberg *et al*, 2014) and SRTpred (Garg & Raghava, 2008).

Homology Modelling with RosettaCM

Homology models were created for a representative member of each cluster using RosettaCM (Combs *et al*., 2013). To determine the optimum number of templates to use for modelling, preliminary homology models were created for a representative member of cluster 6541 (Accession Number G3AXW8) using 3, 5, 10, 15, 20, 30 and 50 templates. For this the HHpred server (Biegert *et al*., 2006; Söding *et al*., 2005) was used to obtain alignments of sequences with homolog PDB templates for the comparative model building (HMM database: PDB, Maximum number of hits set to 3, 5, 10, 15, 20, 30 and 50).

Rosetta's REscore (Combs *et al*., 2013) as well as Qmean (Benkert *et al*., 2008; Benkert *et al*., 2009) were used to predict the best models. The number of templates used against Rosetta's REscore (Appendix 2) and number of templates used against Qmean score (Appendix 3) were plotted, determining that the models produced from more templates are of lower quality. Based on these preliminary results from cluster 6541, models were only created using 3, 5 and 10 templates for the other clusters.

RosettaCM was then used to produce homology models for a representative member of the other clusters of interest. The HHpred server was used to obtain alignments of sequences with homolog PDB templates for the comparative model building (HMM database: PDB, Maximum number of hits set to 3, 5 and 10). For all of the clusters, the 20 models produced using 3 templates, gave overall better Qmean and Rosetta REscore results than those models produced using 5 and 10 templates. To select the models to use for docking, from the 20 produced for each cluster using 3 templates, the Rosetta REscores were plotted against the Qmean Scores. The models used for docking for each cluster were those that had the lowest Rosetta REscore and highest Qmean score. PyMOL (The PyMOL Molecular Graphics System, Version 1.3 Schrödinger, LLC) was used to view and align the top scoring models from each cluster. For all clusters notable differences were observed between the top scoring models at the N-terminus. Cluster 6552 models also had notable differences at the C-terminal.

It was decided to remodel the structures, using 3 templates (Table 1), after editing the sequences to remove the N-terminal sequence before the start of the first domain of the HP (everything before the domain proceeding the RHG), which included the nudix domain in clusters 6544 and 6545. In the case of cluster 6552 the C-terminal region that exhibited significant differences was also removed for remodelling.

Table 1: Sequences and templates used for final homology modelling. The representative sequence from each cluster, that was edited to remove the N-terminal region before the start of the first domain (and the C-terminal region that exhibited significant differences in cluster 6552), and the three homolog PDB templates identified by HHpred, used to produce the final 20 models for each cluster using RosettaCM.

Cluster	Representative Sequence (Accession Number)	Templates (PDB IDs)
6564	Q9W438	2GFI 1QWO 1QFX
6563	R8BHM4	1H2E 4IJ5 4PZA
6560	E5CGZ1	4FDT 2GFI 1QWO
6552	B2VS43	1ND6 3IT3 4JOB
6545	K9ADW1	4IJ5 4PZA 1H2E
6544	W5IGJ5	4PZA 4IJ5 1H2E
6541	G3AXW8	1H2E 4IJ5 4PZA

Mapping Sequence Conservation To Models

The CONSURF server (Ashkenazy *et al*., 2010; Celniker *et al*., 2013; Glaser *et al*., 2003; Landau *et al*., 2005) was used to map sequence conservation on to the top scoring model for each cluster. For this Multiple Sequence Alignment (MSA) files were prepared by aligning the sequences from each cluster using MUSCLE v3.8.31 (Edgar, 2004) which were corrected to remove the N-terminal sequence before the start of the first domain of the HP (In the case of cluster 6552 the C-terminal region that exhibited significant differences was also removed) using Jalview 2.8.1 (Waterhouse *et al*., 2009). The CONSURF results were visualised in PyMOL (using Colour > Spectrum > B-factors) to displaying sequence conservation on top of the models as a spectrum of colours from blue, indicating conservation among relatives of this protein, to red indicating lack of conservation.

Docking To Homology Models

The RosettaLigand tool (Combs *et al*., 2013) at the ROSIE Server (Lyskov *et al*., 2013) was used to dock a library of conformers for each ligand (SDF file) into active sites of the best scoring protein models (PDB files) for each cluster. First the ligands to dock into each clusters top scoring model, identified by STRING above, were obtained from PDB entries in the RCSB PDB database (Berman *et al*., 2000). They were opened in PyMOL and the 3D coordinates for each ligand saved as .mol files. The Molecular formats converter (http://www.webqc.org/molecularformatsconverter.php) was then used to convert each .mol file to a .mol2 file, adding hydrogens. Frog2 (Miteva *et al*., 2010) at RPBS's Mobyle (Alland *et al*., 2005; Néron *et al*., 2009) was used to generate the library of conformers for each ligand (output as an SDF file) using the mol2 files as input (All other settings Default).

As RosettaLigand cannot perform binding site detection the approximate location of the binding site within 5 Å must be input to ROSIE. To do this the ligands were manually docked into the binding sites of the models using PyMOL to determine the x, y and z coordinates of the central atom. The PDB file containing the best scoring protein model without the ligand present, the SDF file containing the conformers of the ligand to be docked, and the X, Y, Z cartesian coordinates of where the ligand should be placed initially prior to docking were used as input for the RosettaLigand server. The number of structures to generate (number of docking predictions to create) was set to 1000 to produce high quality protein-ligand docking poses (All other settings Default). The generated structures were rank ordered according to interface_delta_X scores (the difference between the total Rosetta energy score with the ligand bound, and the ligand unbound). The ten lowest scoring poses were visually evaluated using PyMOL.

RESULTS & DISCUSSION

Collection of a Complete Set of Histidine Phosphatase Sequences

Jackhmmer (Eddy, 1998; Finn *et al*., 2011) was used to collect the HP superfamily sequences, by iteratively searching eight branch 1 representative sequences (Appendix 1) against the UniRef90 (Bateman *et al*., 2015; Suzek *et al*., 2011) database. A preliminary test using Jackhmmer's default of a maximum of 5 iterations revealed that not all the HP superfamily sequences were collected, so the maximum number of iterations was increased to 20.

The number of sequences collected from each of the eight queries tended to plateau at around 30,000 sequences (Appendix 4). After further iterations some of the collections became contaminated with unrelated proteins and proteins with glutamine rich regions, climbing to over 100,000 sequences. The sequences collected from two of the mutase queries (Accession numbers P00950 and P07738) and the *E. coli* SixA query (Accession number P76502) were not contaminated even at higher iterations, remaining stable at around 30,000 sequences.

A comparison was made between the three collections that plateaued and remained stable at around 30,000 sequences, revealing the sequences were largely the same in each collection. As the queries used to collect the sequences were all from the larger branch 1, the collections were also manually checked for the presence of branch 2 proteins. All three collections contained branch 2 APs and phytases. The fact that branch 2 proteins were present and the sequences were largely the same in each collection and instilled confidence that the entire HP superfamily had been collected. The collection containing 29,295 sequences generated from the *E. coli* SixA query was arbitrary chosen for subsequent sequence analysis and clustering.

Sequence Analysis and Clustering

Redundancy was removed from the collected sequences, using CD-HIT v4.6.1 **(**Fu *et al.*, 2012; Li & Godzik, 2006) to apply a 65% sequence identity cut-off, reducing the number of sequences from 29,295 to 10,548. Using CLANS software (Frickey & Lupas, 2004; Frickey & Weiller, 2007) the sequences were then partitioned using sequence similarity into “clusters” that likely share the same function.

The sequence similarity network generated by CLANS shown in Figure 3 depicts the functional relationship among the sequences, where each node (dot) represents a protein sequence. Many of the nodes are connected with lines (edges) which represent the BLAST hits between the sequences. Dark lines symbolize highly significant BLAST hits (i.e. relatively low E-values) indicating close relationships between sequences, while lighter lines signify less significant BLAST hits (i.e. higher E-values).

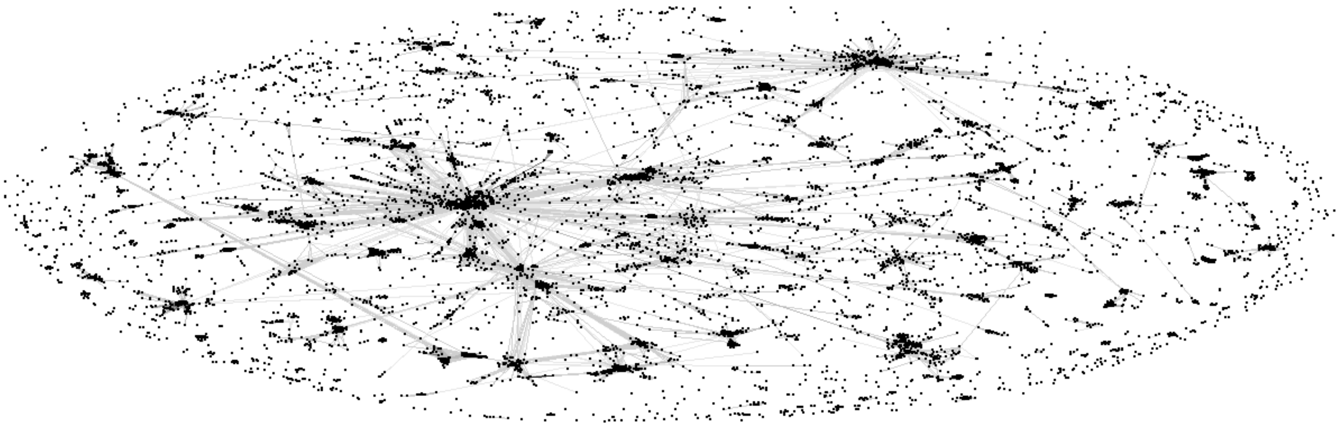


Figure 3: Sequence similarity network for the HP superfamily members. A 3D graph layout partitioning 10,548 HP superfamily sequences into "clusters". The various sequences (nodes) are represented as dots. The lines (edges) connecting the nodes represent connections with BLAST E-values more stringent than the cut-off value ≤1e-40. Image generated using CLANS (Frickey & Lupas, 2004; Frickey & Weiller, 2007).

The relationships between the sequences were visualised at different levels of granularity, to observe how the various groups associate or dissociate, by using different E-value cut-offs to represent the network. As the cut-off value specifies up to what E-value to take BLAST hits into account for the clustering, more stringent values displayed the network with more numerous smaller groups. The network shown in Figure 3 was generated from 32739 rounds of clustering using an E-value cut-off value of ≤1e-40. Sequence similarity networks offer the advantages of being able to visualise relationships among huge numbers of sequences, such as the 10,548 represented here, with moderate computational expense (Atkinson *et al*., 2009).

Eliminating Clusters of Proteins with Known Functions

Clusters were automatically detected in CLANS using an iterative "neural network based" approach. Database searches were then performed for the largest clusters, to eliminate those with known functions. First a representative sequence for each cluster was used as a query for a HHpred search (Biegert *et al*., 2006; Söding *et al*., 2005) against the PDB (Berman *et al*., 2000; Bourne *et al*., 2004) and CD (Marchler-Bauer *et al*., 2015) databases, to detect clusters containing sequences with close similarities to protein structures where the function is better known. The same or similar function was assumed for clusters with relatively high sequence identity (>30%) to a reliable family or structure of known function. An example was cluster 6572 that contained 216 sequences. It had a top hit of 85% sequence identity to 6-phosphofructo-2-kinase/fructose-2-6-bisphosphatase 3

(PDB ID 5AK0; Boyd *et al*., 2015). At levels of sequence identity greater than 30%, we could be confident our cluster had the corresponding activity, so the clusters were discarded.

For the remaining clusters activities not represented by structures or domain entries were detected using each clusters sequences as a query for a BLAST search (Camacho *et al*., 2009) against the Swiss-Prot database (Bairoch & Apweiler, 2000; Bairoch *et al*., 2004; Bateman *et al., 2015*). Again clusters whose sequences had relatively high sequence identity (>30%), were unambiguously assigned the Swiss-Prot activity and dismissed. An example was cluster 6566 that contained 61 Sequences. All representatives had 3-phytase as the top hit with a maximum sequence identity of 49%.

The clusters of interest to this study, that returned relatively low sequences identities (≤30%) from the HHpred and BLAST searches, are shown in Table 2. These clusters were retained for further investigation to determine novel functions by genomic context, homology modelling and small molecule docking.

Table 2: Clusters of Interest. Clusters were no conclusion could be made about their function, as they returned relatively low sequences identities (≤30%) from database searches. The number of sequences found in each cluster is shown along with the accession number of the representative sequence used to query the PDB and CD databases using HHpred, the highest sequence identity detected by HHpred, the results from searching all the sequences in the cluster against Swiss-Prot using BLAST, the orthologous groups (COGs and NOGs) that the cluster corresponds to in the STRING database and the organisms the clusters sequences are found in.

Cluster Number	**Number of Sequences in Cluster**	**Accession Number of Query for HHpred Search**	**HHpred Highest Sequence Identity**	**BLAST against Swiss-Prot Highest Sequence Identities**	**Corresponding Orthologous Group in STRING**	**Organism**
6564	59	D3TLX8	19%, phytase (PDB 1QWO; Xiang *et al*., 2004)	Majority of representatives had multiple inositol polyphosphate phosphatase 1 as the top hit with sequence identities generally ranging between 23-30% Five representatives had Regulatory-Associated Protein of mTOR (RAPTOR) as the top hit with sequence identities ranging between 41-43%	Majority of sequences correspond to NOG30599 containing 452 proteins in 184 species Five of the sequences correspond to NOG03700 containing 327 proteins in 214 species	Insects
6563	53	A0A074XP64	22%, glucosyl-3-phosphoglycerate phosphatase (PDB 4PZA; Zheng *et al*., 2014)	All representatives had phosphomutase-like protein 3 as the top hit with sequence identities generally ranging between 30-35%	NOG54269 containing 272 proteins in 107 species	Fungi
6560	38	A0A096AB40	27%, putative multiple inositol polyphosphate HP (PDB 4FDT; Stentz *et al*., 2014)	Majority of sequences yielded no hits 11 had multiple inositol polyphosphate phosphatase	NOG14402 containing 48 proteins in 45 species	Bacteria

				1 as the top hit, with sequence identities ranging between 19-24%		
6552	27	A0A010RPL5	24%, prostatic AP (PDB 1ND6; Ortlund *et al*., 2003)	Majority of representatives had lysosomal, prostatic and testicular APs as their top hits, with sequence identities generally ranging between 22-24%. 3 representatives had no hits found.	NOG23976. Containing 62 proteins in 52 species	Fungi
6545	17	E3GZT0	19%, Phosphatase (PDB 1H2E; Rigden *et al*., 2003)	A nudix family protein probable 8-oxo-dGTP diphosphatase 1 consistently had highest sequence identity generally ranging between 32-34% No HP domains found	Majority of sequences correspond to COG0494 containing 10042 proteins in 1869 species Four of the sequences correspond to COG0406 containing 5118 proteins in 1590 species	Bacteria
6544	17	A0A087BAT4	20%, Phosphatase (PDB 1H2E; Rigden *et al*., 2003)	Only one representative contained a HP (SixA with a 26% sequence identity A nudix family protein probable 8-oxo-dGTP diphosphatase 1 consistently had highest sequence identity generally ranging between 31-34%	Majority of sequences correspond to COG0494 containing 10042 proteins in 1869 species 1 exception (E6K2B8) corresponds to COG0406 containing 5118 proteins in 1590 species	Bacteria
6541	16	G3AXW8	22%, glucosyl-3-phosphoglycerate phosphatase (PDB 4PZA; Zheng *et al*., 2014)	All representatives had phosphomutase-like protein 3 as the top hit with sequence identities generally ranging between 30-34%	NOG54269 containing 272 proteins in 107 species	Yeast

Domain Fusion Detection

To detect any other domains fused to the HPs in the clusters of interest an RPS BLAST (Reversed Position Specific BLAST) (Camacho *et al*., 2009) against the CD database of domains (Marchler-Bauer *et al*., 2015) was performed using the sequences from each cluster as the query. Clusters 6541, 6552, 6560 and 6563 had no domains fused to the HP.

In cluster 6564 the majority of sequences had no other domain fused to the HP, although five had a fused RAPTOR N-terminal CASPase like domain (pfam14538; Ginalski *et al*., 2004). As the same domain architecture was expected for all members of the same cluster, and only a minority of the cluster had the RAPTOR domain fused there were doubts to its reliability (explained in detail below).

All sequences from clusters 6544 and 6545 contained domains from the nudix hydrolase superfamily fused to the HP domain. The strongest specific match was to cd03673, Ap6A_hydrolase, Diadenosine hexaphosphate (Ap6A) hydrolase.

Cluster 6560

Searching each of the 38 cluster 6560 sequence identifiers at UniProt (Bateman *et al*., 2015) revealed that the vast majority of the sequences were from medically important bacteria living on or inside humans, including various species from the gram negative *Bacteroides* and *Prevotella* genera.

In order to utilise the non-homology functional information contained in the STRING database, which deals with computationally predicted Clusters of Orthologous Groups (COG), it was determined that the sequences in cluster 6560 correspond to Nonsupervised Orthologous Group (NOG) 14402. There are 48 proteins in 45 species in NOG14402, as at STRING v10 (Szklarczyk *et al*., 2015). STRING revealed a strong connection of NOG14402 to the following predicted functional partners; NOG04515 containing the regulatory protein SusR (Score 0.592), NOG00966 containing SusD/RagB family proteins (Score 0.513) and NOG00099 containing TonB linked outer membrane receptor proteins of the SusC/RagA family (Score 0.442). These connections arise from genomic context evidence. As Figure 4 shows NOG14402 (corresponding to cluster 6560) genes are frequently found neighbouring these functional partners, encoding Sus (starch utilization system) family proteins. The Sus loci and other polysaccharide utilization loci (PULs), termed Sus-like systems, are responsible for the acquisition of starch and other glycans in prokaryotes.

Bacteria that reside in the human intestine forage for a broad diversity of complex glycans and polysaccharides, including those derived from dietary animal and plant tissues as well as host mucosal secretions (Koropatkin *et al*., 2012). Many of the animal and plant derived dietary glycans cannot be degraded by human genome encoded enzymes, so microbial fermentation, transforming indigestible complex glycans into products such as short-chain fatty acid that humans can absorb, has an crucial symbiotic role in helping humans access calories from otherwise indigestible nutrients (Flint *et al*., 2008; Koropatkin *et al*., 2012; McNeil, 1984).

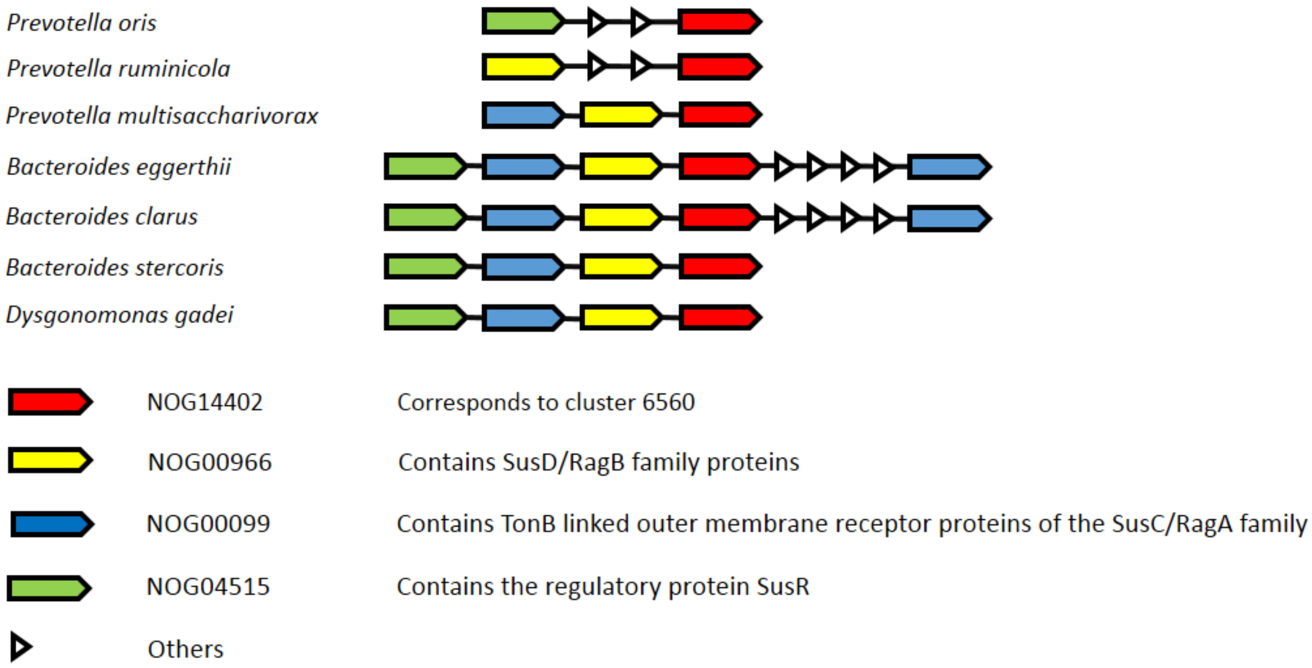

Prevotella oris
Prevotella ruminicola
Prevotella multisaccharivorax
Bacteroides eggerthii
Bacteroides clarus
Bacteroides stercoris
Dysgonomonas gadei
NOG14402
Corresponds to cluster 6560
NOG00966
Contains SusD/RagB family proteins
NOG00099
Contains TonB linked outer membrane receptor proteins of the SusC/RagA family
NOG04515
Contains the regulatory protein SusR
Others

Figure 4: NOG14402 neighbourhood view. Diagram shows genes that occur repeatedly in close neighbourhood to NOG14402 (corresponding to cluster 6560) in prokaryotic genomes. Different coloured arrows represent NOGs. The direction of arrows indicates the direction of transcription. Genes located together are linked with a black line (maximum allowed intergenic distance is 300 base pairs). Small white triangles represent other neighbouring genes. The diagram is based on relationships discovered in STRING v10 (Szklarczyk *et al*., 2015).

Bacteroides thetaiotaomicron, a gram-negative obligate anaerobe, is one of the best studied representatives of the *Bacteroides* genus, which are the most abundant microbes found in the human bowel (Hooper *et al*., 2002; Xu *et al*., 2003). Salyers and colleagues performed seminal work (Cho & Slayers, 2001; D'Elia & Salyers, 1996), discovering the multi-protein cell envelope-associated Sus system that provides the mechanism of how *B. thetaiotaomicron* metabolises starch, revealing a paradigm for acquisition of glycan that is universal in bacteroidetes (Martens *et al*., 2009).

SusR, SusC and SusD, homologs of the genes consistently found neighbouring NOG14402 (corresponding to cluster 6560), are three of the eight adjacent genes comprising the Sus locus (*susRABCDEFG*) required to metabolise starch in *B. thetaiotaomicron* (Martens *et al*., 2009; Tancula *et al.*, 1992). As shown in Figure 5, the Sus products are located in the periplasm and the outer membrane of the bacterium. They act by binding starch to the surface of cell before sequentially, degrading it into smaller oligosaccharides which they transport to the periplasmic space. They oligosaccharides are then degraded further to glucose and other simple sugars before they are imported into the cell (Koropatkin *et al*., 2012).

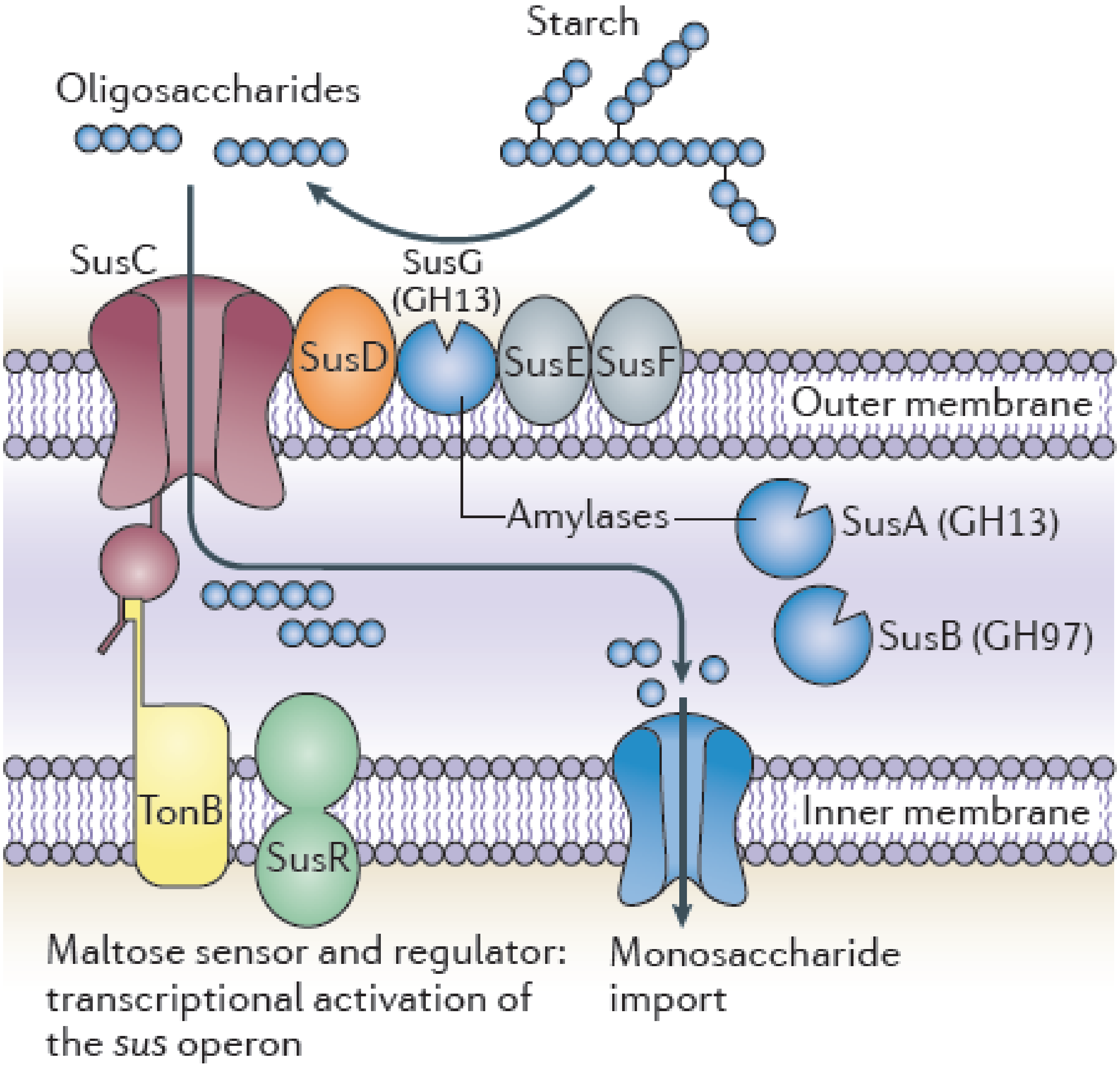


Figure 5: A model of the Sus system in *Bacteroides thetaiotaomicron*. The TonB-dependent transporter SusC works in concert with the starch-binding lipoproteins SusD, SusE, SusF and SusG that localise to the outer membrane (Shipman *et al*., 2000). SusD, SusE and SusF initiate starch binding. SusG, an outer membrane α-amylase, hydrolyses surface-bound starch (Reeves *et al*., 1997), generating cuts, releasing oligosaccharides larger than maltotriose (Martens *et al*., 2009). The oligosaccharides are transported into the periplasm via SusC in concert with the inner-membrane protein TonB. SusA and SusB are glycoside hydrolases that remain in the periplasm (Shipman *et al*., 2000), and further degrade the sequestered oligosaccharides into their component sugars prior to final transport to the cytosol (Martens *et al*., 2009). The presence of liberated maltose is sensed in the periplasm by the inner-membrane-spanning regulator SusR, which activates expression of the other Sus proteins. From Koropatkin *et al.* (2012).

SusR is an inner-membrane-spanning regulator, whose C-terminus, that contains a DNA-binding motif, remains in the cytoplasm, while it's N-terminus extends into the periplasmic space (D'Elia & Salyers, 1996). SusR activates the transcription of the other seven Sus proteins when the presence of starch, as small as the disaccharide maltose, is sensed by the periplasmic domain (D'Elia & Salyers 1996; Koropatkin *et al*., 2012). SusC and SusD localise to the outer membrane (Shipman *et al*., 2000). SusD initiates starch binding (Koropatkin *et al*., 2012) while SusC, in concert with the inner-membrane protein TonB, transports oligosaccharides into the periplasmic space (Koropatkin *et al*., 2012).

B. thetaiotaomicron has the ability to utilize a wide range of glycans and polysaccharides, not just starch (Hooper *et al*., 2002; Xu *et al*., 2003). Sequencing the genome of *B. thetaiotaomicron* revealed genes for over 80 PULs (Sonnenburg *et al*., 2010), whose products have been termed Sus-like systems as they work in a similar manner to Sus, but possess enzymes targeting glycans other than starch. Sus-like systems are common among Bacteroidetes (Koropatkin *et al*., 2012), with homologs of SusC and SusD present in every Sus-like PUL (Ravcheev *et al*., 2013), although the other genes within Sus-like PULs frequently share little or no homology to the archetypical Sus locus (Martens *et al*., 2009). Sus-like systems have been identified for all the glycans that are common in tissues of animals and plants that enter the human intestine, with the exception of cellulose (Koropatkin *et al*., 2012; Martens *et al*., 2011; Sonnenburg *et al*., 2010). Sus-like systems have evolved broadly, with the complexity of the glycan target correlated directly to the number of enzymes in a given system (Koropatkin *et al*., 2012). As well as encoding enzymes involved in breaking glycosidic linkages, some PULs also encode enzymes for removal of glycan modifications a prerequisite step in degrading the underlying backbone (Martens *et al*., 2009). These enzymes could potentially include phosphatase activities, such as the HPs from cluster 6560 that are frequently found in the genomic neighbourhood of SusR, SusC and SusD homologs.

In animals and plants glucan phosphatases are essential for the metabolism of glycogen and starch respectively (Gentry *et al*., 2013; Kotting *et al*, 2010; Silver *et al*., 2014). Humans contain a single identified glucan phosphatase called laforin that dephosphorylates glycogen and is conserved in vertebrates (Gentry *et al*., 2007; Gentry *et al*., 2013; Worby *et al*., 2006). Plants contain two know glucan phosphatases called Starch EXcess4 (SEX4) and Like Sex Four2 (LSF2) that dephosphorylate starch (Kotting *et al*., 2005; Santelia *et al*., 2011).

In plants reversible phosphorylation solubilises the outer surface of starch permitting access to hydrolytic enzymes for processive degradation, however β-amylase the main enzyme responsible for degradation, is unable to degrade glucan chains past a phosphate group (Kotting *et al*., 2009; Meekins *et al*., 2015; Takeda & Hizukuri

1981). For cyclical starch degradation to proceed these phosphate groups must be removed by the SEX4 and LSF2 glucan phosphatases (Kotting *et al*., 2009; Santelia *et al*., 2011).

Starch is phosphorylated at both the C3- and C6-positions, while glycogen is phosphorylated at the C2-, C3-, and C6-positions. Carbohydrate substrates are dephosphorylated by glucan phosphatases in a position specific manor, which differs significantly between the three glucan phosphatases. Laforin preferentially dephosphorylates the C3- position of glycogen, SEX4 shows a C6 preference of starch glucose and LSF2 is C3 specific (Meekins *et al*., 2015).

A hypothesis is that the HPs from cluster 6560 in bacteria could be carrying out a similar role to the glucan phosphatases in animals and plants. In the intestine, where an animal is digesting food composed of plant and animal tissues, the intestinal bacteria may have evolved mechanisms to exploit otherwise indigestible glucans. By dephosphorylating glucans, the HPs of cluster 6560 could permit the Sus (and Sus-like) enzymes in their genomic neighbourhood to further metabolise the glucans.

As the Sus and Sus-like products are located in the outer membrane and the periplasm of bacteria, a representative sequence (Accession number E5CGZ1 from a *Bacteroides*) was submitted to subcellular localisation prediction servers to specifically distinguish between bacterial secreted extracellular proteins from those proteins that are localised and retained in periplasm, to better determine where the HPs fit into the putative Sus-like pathway. The length of the saccharide the HPs encounter also depends on whether the protein is periplasmic or extracellular. Periplasmic proteins would most likely encounter smaller substrates, while secreted proteins would most likely encounter larger substrates.

A signal peptide was detected by SignalP 4.1 (Emanuelsson *et al*., 2007; Petersen *et al*., 2011), PrediSi (Hiller *et al*., 2004), Signal-3L (Shen & Chou, 2007) and PSORTb v3.0.2 (Yu *et al*., 2010), however Phobius (Käll *et al*., 2004; Käll *et al*., 2007) did not detect a signal peptide. For PSORTb the final predicted location of the protein is unknown, while PrediSi, LocTree3 (Goldberg *et al*., 2014) and SRTpred (Garg & Raghava, 2008) predicted it is secreted. In contract the SOSUIGramN (Imai *et al*., 2008) and SLP-local (Matsuda *et al*., 2005) servers both predict the protein is localised in the periplasm. Gneg-mPLoc (Shen & Chou, 2010) predicted both cytoplasm and periplasm, although cytoplasmic localisation cannot be correct as the general consensus is there is a signal peptide present.

In gram-negative bacteria proteins containing signal peptides are localised to the periplasmic compartment by default, and a different mechanism is required for them to be secreted (Pugsley *et al*., 1997). The server results reveal the HPs of cluster 6560 do not remain in the cytoplasm, as the general consensus is that they contain a signal peptide, although a confident distinction between secretion and periplasmic localisation was not made. There are lots of possibilities of different substrates of various sizes for these HPs. Homology modelling combined with conservation mapping was used to explore the size of a predicted conserved binding site, to shed light on the size of potential ligands.

Twenty homology models were constructed for a representative of cluster 6560 (Accession number E5CGZ1 – sequence edited to remove the N-terminal region before the start of the first domain), with RosettaCM (Combs *et al*., 2013), using three homologs (PDB IDs: 4FDT, 2GFI and 1QWO) identified by HHpred (Biegert *et al*., 2006; Söding *et al*., 2005) as templates. The best scoring model (Figure 6A) obtained a Rosetta REscore (Combs *et al*., 2013) of -130.152 and Qmean score (Benkert *et al*., 2008; Benkert *et al*., 2009) of 0.671, initially indicating a high quality model. When the top scoring models were viewed and aligned in PyMOL, they were similar throughout their structures. This consistency between the models increased confidence in the modelled binding site that a ligand was subsequently docked into.

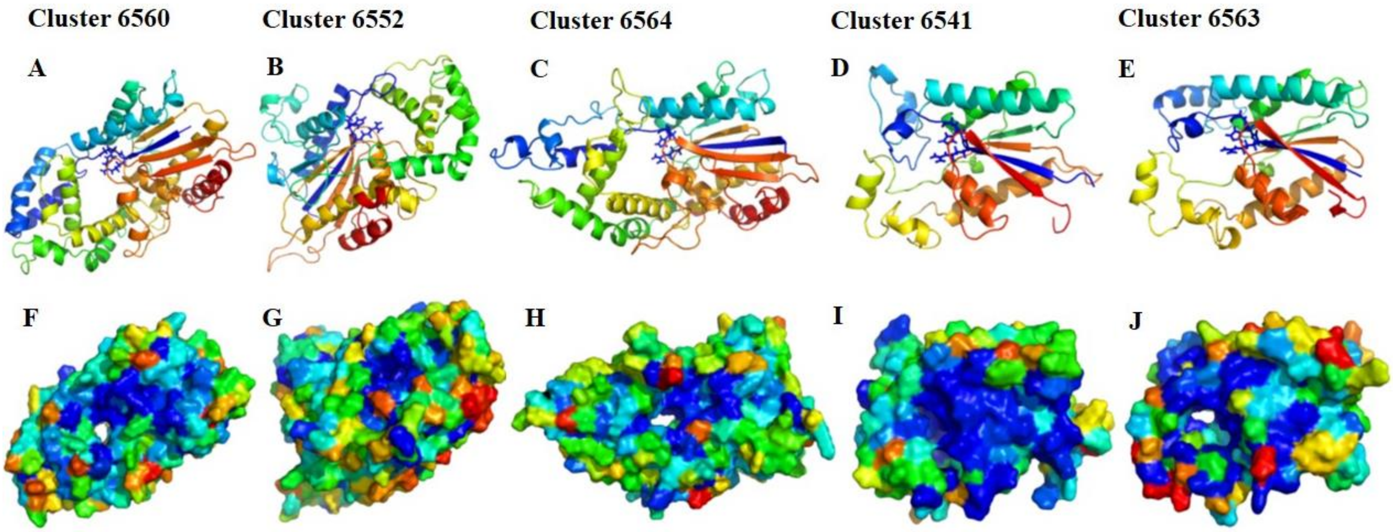


Figure 6. Two views of the top scoring model from each cluster. **A-E** show cartoon representations coloured blue (N-terminus) to red (C-terminus), with sticks for the conserved RHG residues. **F-K** show sequence conservation mapping onto the molecular surfaces with CONSURF (Ashkenazy *et al*., 2010; Celniker *et al*., 2013; Glaser *et al*., 2003; Landau *et al*., 2005). Sequence conservation is displayed as a spectrum of colours from blue (indicating conservation) to red (indicating lack of conservation). Images generated using the PyMOL Molecular Graphics System, Version 1.3 Schrödinger, LLC.

Sequence conservation was mapped onto the top scoring model with CONSURF (Ashkenazy *et al*., 2010; Celniker *et al*., 2013; Glaser *et al*., 2003; Landau *et al*., 2005), using a MSA of the sequences in cluster 6560 prepared with MUSCLE v3.8.31 (Edgar, 2004) in Jalview 2.8.1 (Waterhouse *et al*., 2009). Conservation mapping revealed a large strongly conserved patch (Figure 6F). In the model the RHG motif residues, which include the histidine that is phosphorylated during catalysis, were conserved and surrounded by other conserved residues. Viewing the molecular surface revealed a hole through the centre of the protein, which is extremely unusual and likely a localised modelling error. One side of this binding site pocket was more conserved than the other, and was closer to the RHG motif, so it was assumed the true site for binding, for manual docking of a ligand prior to docking with the RosettaLigand tool (Combs *et al*., 2013) to identify the optimum ligand binding pose. The size of the pocket comfortably accommodated the manual docking of alpha-maltose 1-phosphate, a phosphorylated glyan, so a library of conformers for this ligand where used for docking at the ROSIE server (Lyskov *et al*., 2013).

The library of conformers was prepared using Frog2 (Miteva *et al*., 2010). ROSIE produced 1000 docking prediction structures, that were rank ordered according to interface_delta_X scores (the difference between the total Rosetta energy score with the ligand bound, and the ligand unbound). This interface energy metric discriminates between well and poorly modelled ligand binding poses based on score, identifying conformations and relative orientations that minimises the Rosetta score function. The ten lowest scoring poses were visually evaluated using PyMOL.

When the docking results were compared against the conservation mapping from CONSURF it was revealed that the best scoring ligand binding pose predictions docked alpha-maltose 1-phosphate to the less conserved side of the hole through the model. In all ten lowest scoring poses the phosphate from the alpha-maltose 1-phosphate ligand was oriented away from the RHG motif and was generally not touching the conserved regions. The localised modelling errors that resulted in the hole through the centre of the structure were likely reducing the quality of the modelled binding site. It appeared the poorly modelled areas in the model were preventing the ligands phosphate from binding in the correct position. Although the docking did not successfully place the ligand phosphate at the heart of the catalytic site in the phosphate pocket, it did reveal that due to the size of the binding site and conserved regions, glycans larger than alpha-maltose 1-phosphate could not be accommodated.

Cluster 6552

Searching each of the 27 cluster 6552 sequence identifiers at UniProt (Bateman *et al*., 2015) revealed that the proteins are from various fungi including species of agricultural importance such as *Magnaporthe oryzae* (Rice blast fungus), *Gibberella zeae* (Wheat head blight fungus) and *Pyrenophora tritici-repentis* (Wheat tan spot fungus).

The sequences in cluster 6552 correspond to NOG23976. There are 62 proteins in 52 species in NOG23976 as at STRING v10. STRING revealed a strong connection of NOG23976 to the predicted functional partners riboflavin synthase alpha chain (COG0307, Score 0.889) and 3-phosphoadenosine 5-phosphosulfate sulfotransferase (PAPS reductase)/FAD synthetase and related enzymes (COG0175, Score 0.859). These connections arise from database evidence, as the NOG23976 proteins and both of the predicted functional partners are pathway neighbours involved in the manually curated riboflavin metabolism annotated KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway (imported from KEGG (July 2014)) (Kanehisa & Goto, 2000; Kanehisa *et al*., 2014).

Riboflavin (Vitamin B2) is the substrate precursor for the biosynthesis of the essential flavin coenzymes, flavin adenine dinucleotide (FAD) and flavin mononucleotide (FMN) (Birkenmeier *et al*., 2014; Hasnain *et al*., 2013; Hiltunen *et al*., 2012), which are essential for all living organisms and have roles in diverse redox reactions and other processes such as repair of DNA, bioluminescence and light sensing (Fischer & Bacher, 2005). Plants, fungi, archaea and bacteria are able to synthesize riboflavin *de novo* (Birkenmeier *et al*., 2014; Hasnain *et al*., 2013; Hiltunen *et al*., 2012), however mammals and other animals cannot so they must obtain it from a dietary supply (Bacher *et al*., 2000; Powers, 2003). Riboflavin is commercially highly valuable and is used in pharmaceuticals, cosmetics, animal feed supplements and the food industry (Shi *et al*., 2009; Stahmann *et al*., 2000).

The biosynthetic pathway responsible for synthesising riboflavin is similar in yeast, bacteria and plants (Bacher *et al*., 2000; Bacher *et al*., 2001). The filamentous fungus *Ashbya gossypii* overproduces riboflavin naturally and is now commercially one of the top producers by microbial fermentation (Stahmann *et al*., 2000). Osiewacz (2002) describes the synthesis of riboflavin in *A. gossypii* in which all enzymes involved in the process are known, except for the dephosphorylation of the 5-amino-6-ribitylamino-2,4 (1H,3H)-pyrimidinedione-5'-phosphate (ArPP), which is catalysed by a phosphatase that has yet to be characterised (Osiewacz, 2002). Starting with ribulose-5-phosphate (Ribu5P) and guanosine triphosphate (GTP) (Bacher, 1991), the reactions proceed as shown in Figure 7, resulting in riboflavin which is then phosphorylated to FMN and adenylated to FAD (Hiltunen *et al*., 2012). In *A. gossypii,* reduction of the ribosyl residue occurs before deamination of diamino pyrimidinone, in contrast to bacteria where the order of the reactions is reversed (Bacher, 1991; Burrows & Brown 1978).

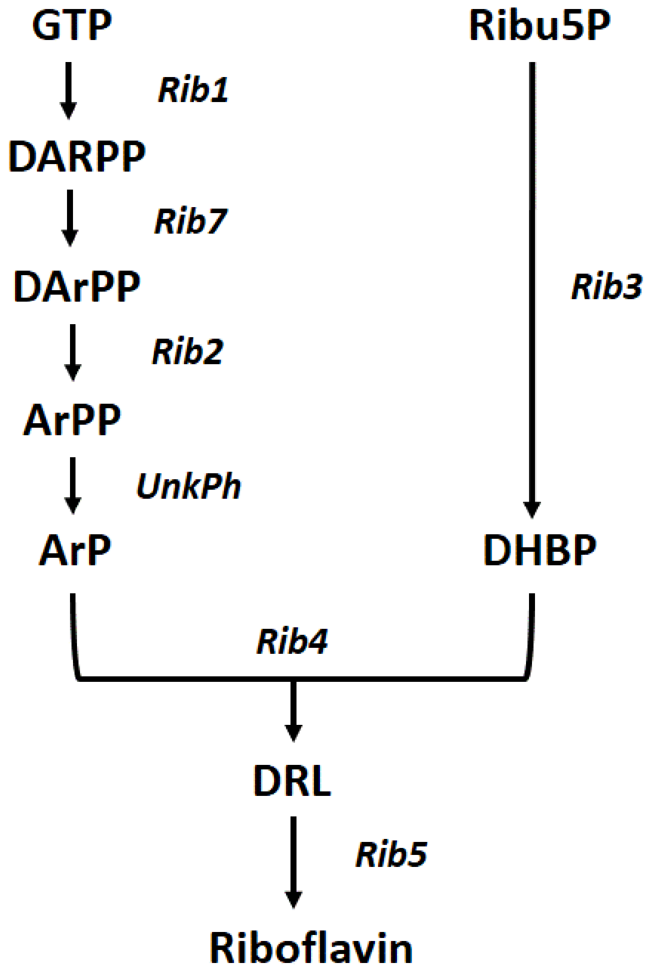


Figure 7: Riboflavin biosynthesis pathway in *Ashbya gossypii*. An enzyme catalysing a dephosphorylation of ArPP in not characterised. Abbreviations: UnkPh (unknown phosphatase), GTP guanosine triphosphate, Ribu5P (ribulose-5'- phosphate), DARPP 2,5-diamino-6-ribosylamino-4 (3H)-pyrimidinone-5'-phosphate / (2,5-diamino-6-hydroxy-4-(5'-phosphoribosylamino)-pyrimidine), DArPP (2,5-diamino-6-ribitylamino-4(3H)-pyrimidinone-5'-phosphate), ArPP 5-amino-6-ribitylamino-2,4 (1H,3H)-pyrimidinedione-5'-phosphate / (5-amino-6-(5'-phosphoribitylamino)uracil), ArP 5-amino-6-ribitylamino-2,4 (1H,3H)-pyrimidinedione / (4-(1-D-ribitylamino)-5-amino-2,6-dihydroxypyrimidine), DHBP (3,4-dihydroxy-2-butanone 4-phosphate), DRL 6,7-dimethyl-8-ribityllumazine / (6,7-dimethyl-8-(1-D-ribityl)lumazine, Rib1 GTP cyclohydrolase II, Rib2 DArPP deaminase, Rib3 DHBP synthase, Rib4 DRL synthase, Rib5 riboflavin synthase, Rib7 DARPP reductase. Adapted from Osiewacz (2002) and Ledesma-Amaro *et al*. (2014).

The phosphatase responsible for ArPP to ArP has not been identified in any organisms (Birkenmeier *et al*., 2014; Gerdes *et al*., 2012; Hasnain *et al*., 2013; Roje, 2007). It is still not clear at present if the dephosphorylation is implemented by a single specific phosphatase or numerous enzymes that are less specific (Hiltunen *et al*., 2012). Abbas & Sibirny (2011) hypothesise that the phosphatase involved in the biosynthesis of riboflavin is most likely substrate specific, as a nonspecific phosphatase would not be able to distinguish between the phosphorylated products of GTP cyclohydrolase II, reductase and deaminase (Abbas & Sibirny, 2011).

As manually curated database evidence from KEGG indicated that NOG23976 (corresponding to the HPs in cluster 6552) proteins are responsible for the previously uncharacterised substrate specific catalysis dephosphorylating ArPP to ArP in the riboflavin biosynthesis pathway, homology models were created for a representative of this cluster to dock a library of ArPP conformers to explore how the substrate potentially binds. The reference that KEGG based this assignment on was not found. KEGG gene annotations are accomplished by assigning genes to KEGG Orthology IDs (KOs) as a functional identifier. KOs are manually curated based on the literature information as well as sequence similarity. It is possible that the coverage of the sequence similarity data is not highly accurate. At present the phosphatase implicated in the pathway is still illusive in published literature.

Twenty homology models were constructed for a representative of cluster 6552 (Accession number B2VS43 – sequence edited to remove the N-terminal region before the start of the first domain and the C-terminal region), with RosettaCM, using three homologs identified by HHpred (PDB IDs: 1ND6, 3IT3 and 4JOB) as templates. The best scoring model (Figure 6B) obtained a Rosetta REscore of -225.364 and Qmean score of 0.656, indicating a high quality model. When the top scoring models were viewed and aligned in PyMOL, they were generally similar throughout their structures although they differed slightly in the light blue region shown in figure 6B that is not near the catalytic site. This general consistency between the models increased confidence in the modelled binding site that the ArPP ligand was subsequently docked into.

Sequence conservation was mapped onto this top scoring model with CONSURF (Ashkenazy *et al*., 2010; Celniker *et al*., 2013; Glaser *et al*., 2003; Landau *et al*., 2005), using a MSA of the sequences in cluster 6552 prepared using MUSCLE in Jalview. Conservation mapping revealed a large strongly conserved patch on the molecular surface forming a binding pocket (Figure 6G). In the model the RHG motif residues, which includes the histidine that is phosphorylated during catalysis, were conserved and surrounded by other conserved residues. The catalytic core RHG motif was positioned at the base of the conserved binding pocket. The regions of the structures that differed between the aligned top scoring models (light blue in Figure 6B), was shown to not be particularly well conserved between the sequences of cluster 6552.

A library of conformers of ArPP was docked into the binding site of the model using the RosettaLigand tool (Combs *et al*., 2013) at the ROSIE Server (Lyskov *et al*., 2013). The library of conformers was prepared using Frog2 (Miteva *et al*., 2010). ROSIE produced 1000 docking prediction structures that were rank ordered according to interface_delta_X scores. The ten lowest scoring poses were visually evaluated using PyMOL, revealing that ArPP was predicted to be positioned “snuggly” in the conserved binding pocket identified by CONSURF. ArPP was touching and surrounded by conserved residues, however its phosphate was oriented away from the RHG motif. Although ArPP was located in the conserved binding cleft, the docking had not positioned ArPP’s phosphate in the phosphate pocket. Errors in the models structure around the binding site could explain this result.

Clusters 6541 and 6563

Cluster 6541 and 6563 are neighbouring clusters in the network generated in CLANS using an E-value cut-off value of ≤1e-40 (Figure 3). When less stringent E-value cut-offs are used these two clusters associate together into one cluster. Searching each of the sequence identifiers at UniProt revealed that the 53 proteins in cluster 6563 are from various fungi, and the 16 proteins in cluster 6541 are more specifically found in yeast.

The sequences in both cluster 6541 and cluster 6563 are most similar to PMU1 from *Saccharomyces cerevisiae* in the STRING database, corresponding to NOG54269. There are 272 proteins in 107 species in NOG54269 as at STRING v10. A bidirectional / reciprocal genome BLAST (Altschul *et al.*, 1990; Altschul *et al.*, 1997) verified the relationship of cluster 6541 and 6563 sequences to PMU1.

PMU1 encodes a putative phosphomutase found in *S. cerevisiae*. Elliot *et al*. (1996) named it PMU1 for Phospho-Mutase homologue, as the encoded protein contains a region homologous to the active site of phosphoglycerate mutases from various organisms (Elliot *et al*., 1996). STRING uncovered two substrates for PMU1 (corresponding to NOG54269), 5'-Phosphoribosyl-4-carboxamide-5-aminoimidazole (AICAR) and trehalose-6-phosphate. Both of these connections are based on experimental evidence (Rébora *et al*., 2005).

First STRING revealed a strong connection of NOG54269 to the predicted functional partners AICAR transformylase / IMP cyclohydrolase (COG0138, Score 0.914), formyltetrahydrofolate synthetase (COG2759, Score 0.839) and 5,10-methylene-tetrahydrofolate dehydrogenase / methenyl tetrahydrofolate cyclohydrolase (COG0190, Score 0.810). Overexpression of PMU1 (NOG54269 corresponding to clusters 6541 and 6563) suppresses the histidine auxotrophy of yeast ADE3 ADE16 ADE17 triple mutants. ADE16 ADE17 mutants lack AICAR transformylase activity (COG0138), while ADE3 mutants lack a trifunctional enzyme with methylenetetrahydrofolate dehydrogenase, methenyltetrahydrofolate cyclohydrolase (COG0190) and formyltetrahydrofolate synthetase (COG2759) activities, that synthesis 10-formyl tetrahydrofolate (THF), a co-substrate required for biosynthesis of IMP (Figure 8).

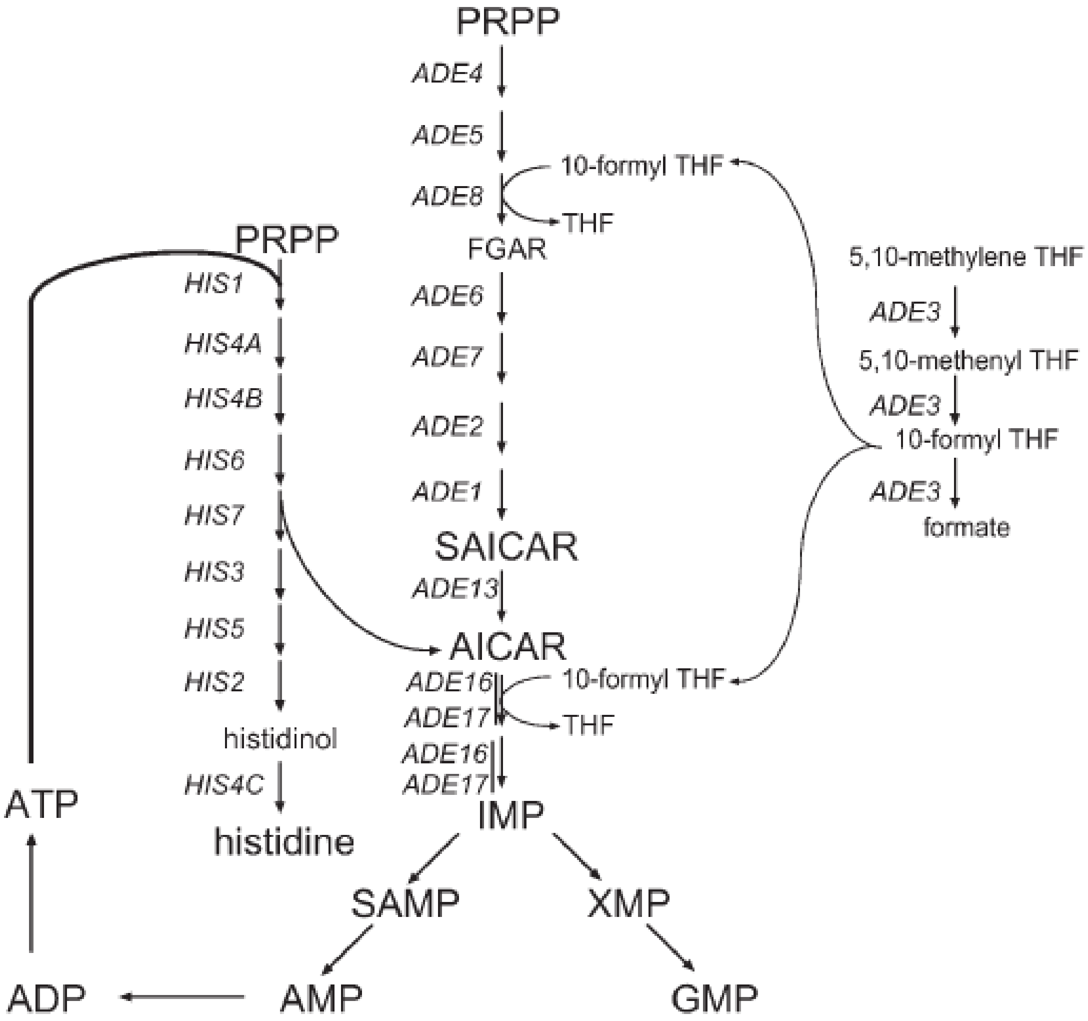


Figure 8: Histidine and IMP biosynthesis pathways. The reactions supply 10-formyl-THF for IMP biosynthesis. Abbreviations: FGAR 5’-phosphoribosyl *N*-formylglycinamide, IMP inosine 5’-monophosphate, PRPP 5-phosphoribosyl-1-pyrophosphate, SAMP adenylosuccinate, XMP xanthosine 5’-monophosphate. Gene names are *italicized*. From Rébora *et al*. (2005).

AMP & histidine biosynthesis pathways are co-regulated at the level of transcription in response to the availability of adenine (Daignan-Fornier & Fink, 1992), and significantly contribute to AICAR synthesis. As well as being an intermediate metabolite at the junction between these two pathways, AICAR is also a regulatory molecule and is critical in directly activating expression of ADE genes when adenine is absent (Rébora *et al*., 2005).

AICAR overproduction and accumulation is toxic for yeast cells (Rébora *et al*., 2005). Tibbetts & Appling (2000) hypothesise that the histidine requirement of ADE3 ADE16 ADE17 triple mutant strains is due to AICAR accumulation, which is responsible for the inhibition of a late step of histidine biosynthesis (Tibbetts & Appling, 2000). PMU1 bypass's the negative effect of AICAR accumulation. PMU1 overexpression suppresses the histidine auxotrophy of ADE3 ADE16 ADE17 triple mutants, by detoxifying AICAR. PMU1 presumably directly modifies AICAR which is a monophosphate nucleotide derivative, by transforming it into another product (Rébora *et al*., 2005).

STRING also revealed a strong connection of NOG54269 to the predicted functional partner trehalose-6-phosphate phosphatase (COG1887, Score 0.776). This connection arose from experimental evidence of a protein-protein interaction between the NOG54269 members and TPS2 (the gene encoding the 100-kDa phosphatase subunit of the trehalose-6-phosphate synthase/phosphatase complex) in *S. cerevisiae*. Overexpression of PMU1 in yeast suppresses the temperature sensitivity of mutants lacking TPS2 whose protein product is involved in the biosynthesis of trehalose (Elliott *et al*., 1996).

Synthesis of the disaccharide trehalose is well correlated with high temperatures (Hottiger *et al*., 1992; De Virgilio *et al*., 1991; Neves & Francois, 1992), and *in vivo* and *in vitro* evidence suggests it is a thermoprotectant, acting by stabilising proteins and preventing heat inactivation (Colaco *et al*., 1992; Hottiger *et al*., 1994). In yeast cells high concentrations of trehalose are well correlated with heat shock resistance (Neves & Francois 1992), and genetic evidence suggests it is important for thermotolerance (De Virgilio *et al*., 1994).

Figure 9 shows the two step reaction to synthesise trehalose in yeast (Elliott *et al*., 1996). TPS2 mutant strains that fail to make the thermoprotectant trehalose and accumulate trehalose-6-phosphate have been isolated. These mutants lacking TPS2 showed significant heat shock sensitivity (Elliott *et al*., 1996). In the stationary phase of TPS2 mutants, both the presence of trehalose-6-phosphate and the lack of trehalose contribute to heat shock sensitivity, while in the log phase temperature sensitivity is solely due to trehalose-6-phosphate accumulation. Elliott *et al*. (1996) suggest the accumulation of trehalose-6-phosphate in TPS2 mutants may be toxic to cells inducing sensitivity to heat shock (Elliott *et al*., 1996).

Overexpression of PMU1 in TPS2-deleted cells reduces the toxic trehalose-6-phosphate levels and suppresses the temperature sensitivity. PMU1 prevents the accumulation of toxic trehalose-6-phosphate by removing it somehow, possibly by transferring the phosphate to another molecule. PMU1 does not restore trehalose to wild type levels and so cannot restore heat shock resistance fully (Elliott *et al*., 1996).

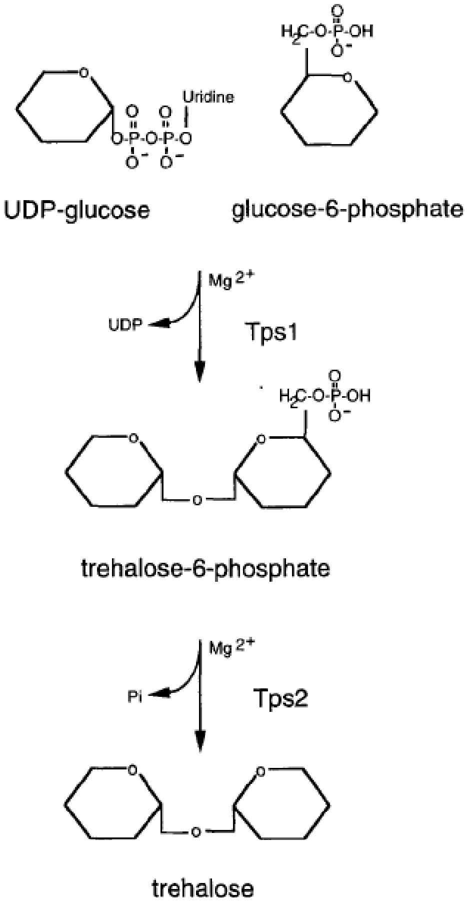


Figure 9: Trehalose biosynthesis in yeast. Step 1: Tps1 (trehalose-6-phosphate synthase) transfers the glucosyl residue from UDP-glucose to glucose-6-phosphate producing trehalose-6-phosphate. Step 2: Tps2 (trehalose-6-phosphate phosphatase) cleaves the phosphate from trehalose-6-phosphate yielding trehalose. From Elliott *et al*. (1996).

As STRING revealed strong connections of cluster 6541 and 6563 HPs to the substrates AICAR and trehalose-6-phosphate, based on the experimental evidence outlined above, these two substrates where docked into homology models of a representative from each of these clusters. As the experimental evidence is drawn from studies involving overexpression of the protein responsible for catalysis, AICAR and trehalose-6-phosphate are not likely to be the primary substrates of the HPs in these clusters. The docking results could however potentially reveal the likelihood of other sugar phosphates, as the targets of these enzymes.

RosettaCM was used to construct twenty homology models for a representative of cluster 6541 (Accession number G3AXW8) and cluster 6563 (Accession number R8BHM4). Both sequences used to construct the models were edited to remove the N-terminal region before the start of the first domain. The structures of the same three homologous proteins (PDB IDs: 1H2E, 4IJ5 and 4PZA) were used as model templates for both clusters, as identified by HHpred.

When the top scoring models of the cluster 6541 representative were viewed and aligned in PyMOL, they were generally similar throughout their structures although they differed slightly in the yellow and light blue regions shown in Figure 6D. The light blue region is not located near the catalytic site, but the potential poor modelling of the yellow region closer to the catalytic site may interfere with ligand docking. When the top scoring models of the cluster 6563 representative were aligned in PyMOL, they were also generally similar throughout their structures although there was a slight divergence in the light blue α-helix at the end furthest from the catalytic (Figure 6E). Inspecting the regions around the catalytic core of the RHG motif in the "cartoon" view, for the top scoring model from each cluster, revealed the α-helices, β-sheets and loop structures in these regions shared similarities in terms of orientation and proximity to the RHG motif. This was expected as the two clusters are neighbours in CLANS generated using an E-value cut-off value of ≤1e-40 (Figure 3), and associate together into one cluster when less stringent E-value cut-offs are used. This sequence similarity displayed in the CLANS network, indicating a functional relationship among the sequences, was confirmed when the HHpred server returned the same three PDB homologs to use as templates in modelling for the representative of both the clusters.

Cluster 6541's best scoring model (Figure 6D) obtained a Rosetta REscore of -186.412 and Qmean score of 0.629, while cluster 6563's best scoring model (Figure 6E) had similar scores of -185.25 for Rosetta REscore and 0.624 for its Qmean score. Sequence conservation was mapped onto these top scoring models with CONSURF, using a MSA of the sequences in each protein's respective cluster prepared with MUSCLE. Conservation mapping of cluster 6541's top scoring model revealed an extremely large strongly conserved patch covering almost an entire side of the molecular surface, including the suspected binding pocket (Figure 6I). In this model the RHG motif residues, positioned at the base of the binding pocket, are conserved and surrounded by other conserved residues. Viewing the molecular surface conservation map of cluster 6563's top scoring model revealed a hole through the centre of the protein (Figure 6J), and that the larger opening of this thoroughfare, suspected to be the binding site due to its proximity to the RHG motif, was not as conserved as the binding site in cluster 6541. This signalled a modelling error in the binding region of cluster 6563's model.

The RosettaLigand tool was used to dock a library of conformers of trehalose-6-phosphate into the binding sites of the top scoring model from cluster 6541 followed by the top scoring model from 6563. Next a library of conformers of AICAR was docked into the top scoring model from each cluster. The ten lowest interface_delta_X scoring docking predictions for each combination of dockings (trehalose-6-phosphate docked to the top scoring cluster 6541 model, trehalose-6-phosphate docked cluster 6563 model, AICAR docked to cluster 6541 model and AICAR docked cluster 6563 model) were visually evaluated using PyMOL.

From the four combinations of dockings, AICAR docked to top scoring cluster 6563 model yielded the best results. For all ten lowest scoring poses from this docking the AICAR substrate was positioned “snuggly” in the conserved binding pocket identified by CONSURF. The AICAR ligand was touching and surrounded by conserved residues with its phosphate pointing towards the histidine of the RHG motif, in the phosphate pocket (Figure 10).

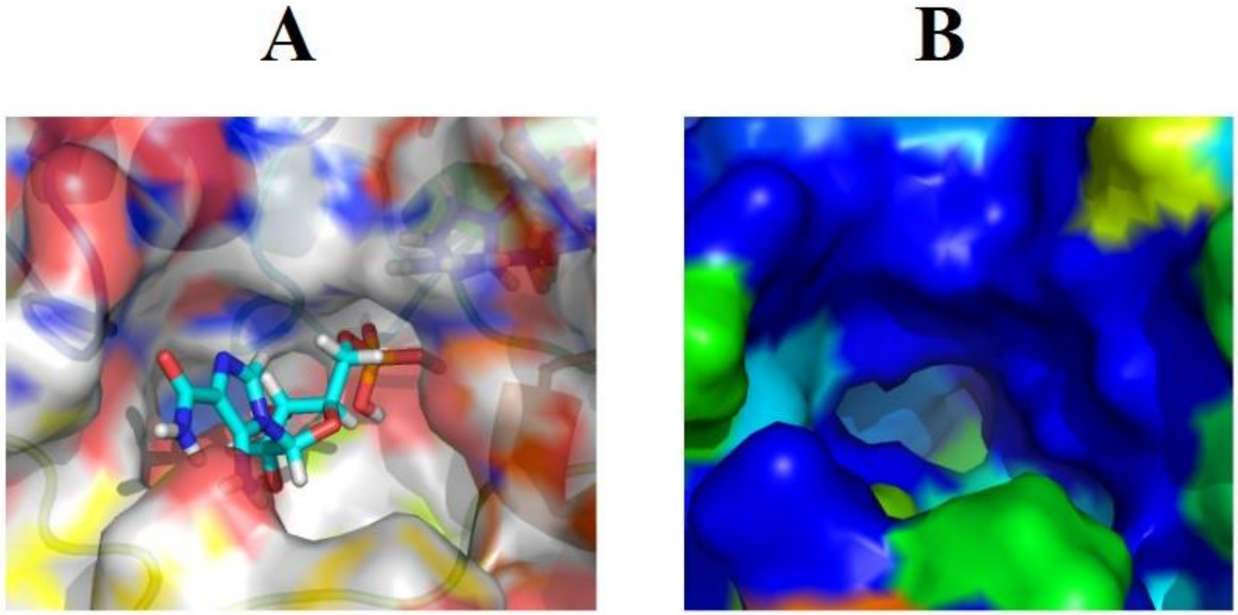


Figure 10. View of AICAR docked into the top scoring cluster 6563 homology model.
(A) Shows the docking pose of AICAR in the top scoring homology model for cluster 6563. The surface of the model is transparent revealing AICAR’s phosphate positioned in the phosphate pocket pointing towards the catalytic histidine residue. **(B)** Shows sequence conservation mapping onto a molecular surface with CONSURF (Ashkenazy *et al*., 2010; Celniker *et al*., 2013; Glaser *et al*., 2003; Landau *et al*., 2005) Sequence conservation is displayed as a spectrum of colours from blue (indicating conservation) to red (indicating lack of conservation). The area where AICAR is best predicted to dock is highly conserved. Images generated using the PyMOL Molecular Graphics System, Version 1.3 Schrödinger, LLC.

When trehalose-6-phosphate was docked into the same cluster 6563 homology model, the majority of the lowest ten scoring docking predictions position trehalose-6-phosphate in the same highly conserved binding site, however the is phosphate oriented closer to the arginine (R) than the histidine (H) in the RHG motif.

The results from docking the same two substrates to the top scoring homology model from clusters 6541 representative where not as promising. The majority of the lowest scoring poses predicted by the RosettaLigand tool position the substrates in the highly conserved binding site, however the phosphates are oriented away from the RHG domain of the model.

Cluster 6564

Searching each sequences identifier at UniProt revealed the sequences were found in an array of insects including *Drosophila melanogaster*, *Nasonia vitripennis* (Parasitic wasp), *Apis mellifera* (Honeybee), and various mosquitos (*Culex quinquefasciatus*, *Aedes aegypti* and *Anopheles gambiae*).

The majority of representatives had multiple inositol polyphosphate phosphatase 1 as the top hit when BLAST was used to search the sequences against Swiss-Prot database, with sequence identities generally ranging between 23-30%. Five representatives however had Regulatory-Associated Protein of mTOR (RAPTOR) as the top hit ranging between 41-43%. RAPTOR is not a HP so a domain fusion, could explain this result. The RPS BLAST against the CD database also revealed the majority of sequences had no other domain fused to the HP, although five had a top hit of a RAPTOR N-terminal CASPase like domain (pfam14538; Ginalski *et al*., 2004). As the same domain architecture was expected for all members of the same cluster, and only a minority of cluster 6564 have the RAPTOR domain fused there were doubts to its reliability.

In the STRING database the majority of sequences in cluster 6564 are most similar to a multiple inositol polyphosphate phosphatase 1-like protein (LOC725931) from the honey bee *A. mellifera*, corresponding to NOG30599. Five of the sequences were most similar to the GF20313 gene product, from transcript GF20313-RA involved in the mTOR signalling pathway in *Drosophila ananassae*, corresponding to NOG03700.

Results from three representative sequences (Accession numbers A0A026VZI3, L7M867 and W8BG64) submitted to signal peptide detection and subcellular localisation prediction servers determined that the RAPTOR connection was an annotation error rather than a legitimate connection, so STRINGS connection to NOG03700 was dismissed. The SignalP 4.1 (Emanuelsson *et al*., 2007; Petersen *et al*., 2011) and Phobius (Käll *et al*., 2004; Käll *et al*., 2007) servers both detected a signal peptide. TargetP 1.1 (Emanuelsson *et al*., 2000; Emanuelsson *et* *al*., 2007) predicted that the subcellular localisation of the protein is the secretory pathway, while PSORTII (Nakai & Horton, 1999) gave conflicting results predicting the endoplasmic reticulum, mitochondria and nucleus. Although the conflicting localisation results did not lead to a confident prediction of location, the general consensus is that there is signal peptide present, and the protein is therefore secreted or localised inside an organelle and not maintained in the cytoplasm. As RAPTOR is a cytosolic protein a fusion of the RAPTOR domain to the cluster 6564 HPs does not make sense so it is likely an annotation error.

The majority of the sequences did not have this fusion error and corresponded to NOG30599. There are 452 proteins in 184 species in NOG30599 as at STRING v10. STRING revealed a strong connection of NOG30599 to the predicted functional partners SPX domain-containing protein involved in vacuolar polyphosphate accumulation (COG5036, Score 0.997), vacuolar transporter chaperone (COG5264, Score 0.997) and NOG212389 (Score 0.997) which contains various transporters including phosphate transporters. These connections arise from co-expression in *S. cerevisiae, Schizosaccharomyces pombe* and *Plasmodium falciparum*.

COG5264's vacuolar transporter chaperone (VTC) complex is involved vacuolar polyphosphate accumulation (Ogawa *et al*., 2000) as well as several other membrane related processes including microautophagy (Uttenweiler *et al*., 2007), membrane trafficking (Müller *et al*., 2003) and non-autophagic vacuolar fusion (Müller *et al*., 2002). COG5036's SPX domain-containing proteins are also involved in vacuolar polyphosphate accumulation. Inorganic polyphosphate (polyP), a linear polymer of three to thousands of inorganic phosphate (Pi) residues linked by high-energy phosphoanhydride bonds, is found in all organisms throughout nature (Kornberg, 1999; Ogawa *et al*., 2000).

In *S. cerevisiae* the PHO pathway regulates Pi homeostasis. During normal and high Pi conditions the PHO pathway is inactive. Under Pi limiting conditions transcription of the PHO operon genes is activated, which act to optimize Pi uptake and utilization (Secco *et al*., 2012). These genes encode secreted APs and the related proteins that increase Pi recovery from the extracellular sources including high-affinity transporters (Dick *et al*., 2011; Mouillon & Persson, 2006; Oshima, 1997). A subset of these proteins contain the SPX domain, which has been shown to be key regulators of Pi homeostasis, and is conserved among all major eukaryotes (Secco *et al*., 2012).

The VTC complex, that is induced under Pi limiting conditions, functions as a heterodimer comprised of Vtc1, Vtc4 (the catalytic subunit) and either Vtc2 or Vtc3 (Secco *et al*., 2012). It synthesizes PolyP, using ATP as a substrate, and then transports the phosphate polymers to the lumen of the vacuole (Hothorn *et al*., 2009). Under Pi deficiency production of intracellular and extracellular protein phosphatases, that hydrolize phosphate esters releasing inorganic phosphate (Pi), is also induced in order to increase levels of free Pi by scavenging Pi from macromolecules (Baldwin *et al*., 2001; Duff *et al*., 1994; Dick *et al*., 2011). An example is the SPX domain containing Gde1, responsible for the hydrolysis of glycerophosphocholine into choline and glycerophosphate (Secco *et al*., 2012). The PHO5 gene (and its homologs PHO11and PHO12) encodes an AP which is localised to the periplasmic space (Ogawa *et al*., 2000).

The APs are nonspecific, and hydrolyze a variety phosphorylated substrates, including nucleic acids, phosphosugars, phospholipids, and phosphoproteins (Baldwin *et al*., 2001; Dick *et al*., 2011; Ogawa *et al*., 2000). The induction of phosphatase activity in response to Pi starvation is a common phenomenon among organism acquiring Pi from the environment (Dick *et al*., 2011).

A hypothesis is that in insects the HPs contained in cluster 6564 (corresponding to NOG30599) are releasing Pi from macromolecules, to be used in vacuolar polyphosphate accumulation by the VTC and other SPX domain-containing proteins that are co-expressed with the phosphatases. As the APs involved in scavenging Pi are non-specific there are many potential substrates for the proteins contained in cluster 6564.

Twenty homology models were constructed for a representative of cluster 6564 (Accession number Q9W438 - sequence edited to remove the N-terminal region before the start of the first domain), with RosettaCM, using three homologs identified by HHpred (PDB IDs: 2GFI, 1QWO and 1QFX) as templates. The best scoring model (Figure 6C) obtained a Rosetta REscore of -78.233 and Qmean score of 0.641. When the top scoring models were viewed and aligned in PyMOL, they were generally similar throughout their structures although they differed slightly in dark blue α-helix shown in Figure 6H. As this model is thought to be of a broad range AP involved in scavenging Pi there are potentially a variety of possible phosphorylated substrates. If the AP dephosphorylates small substrates the potential error in this varying loop, which is relatively far away from the catalytic site, should not affect substrate binding. Sequence conservation mapping onto the top scoring model revealed that this particular α-helix is not highly conserved between the cluster 6564 sequences. As with the top scoring model from cluster 6560, viewing the molecular surface revealed an unlikely hole through the centre of the protein (FIGX), indicating a probable confined modelling error. The molecular surface was particularly conserved in a large patch on one side of this hollow, suggesting the potential of this possible broad range phosphatase to catalyse large substrates. The strongly conserved patch lay above the catalytic RHG motif, pointing towards the binding site for substrates.

Due to the hole through the centre of the model that indicated a modelling error in the binding site combined with the literature research revealing that as the APs involved in scavenging Pi are non-specific (hydrolysing a variety of substrates including phosphosugars, phospholipids, phosphoproteins and nucleic acids), docking was not performed on the homology model from this cluster.

Clusters 6544 and 6545

Although the sequences from clusters 6544 and 6545 were identified as containing domains from the nudix hydrolase superfamily fused to the HP domain, due to time limitations their functions where not explored in depth. These HPs found in a variety of bacteria almost certainly house novel functions. Domain fusions are a potent source of clues to function. Future work would involve further literature research into the nudix domains to shed light on potential substrates for the fused HP domain, and subsequent docking of potential substrates identified into the created homology models.

CONCLUSION

In conclusion this study illustrates the combination of functional information form a variety of different bioinformatics sources. This project reports the successful collection of the complete set of HP superfamily sequences using an iterative database search strategy. Large groups of uncharacterised proteins with potential unknown novel functions where identified as planned. STRING was utilised to provide strong evidence of potential ligands. State of the art homology modelling and metabolite docking techniques were also employed, with mixed results. In general the dockings results did not provide strong evidence of function, as metabolite docking relies on the challenging task of predicting high quality homology models. Modelling can be temperamental as slight changes in a predicted structure can lead to loop formation that blocks the binding site preventing successful prediction of a substrate ligand pose completely. Although some of the regions in the models preventing phosphates binding the correct space, conserved patches identified by conservation mapping indicate the size of likely substrates.

Each method has its own set of benefits and pitfalls, but by combining approaches for a holistic view, the likelihood of successfully recognising new functions increases. The ligands identified here highlight the potential of computational methods to narrow down possible metabolites for future experimental characterisation. Although predicting function remains a challenge the approaches used here for the HP superfamily are applicable to other superfamilies.

ACKNOWLEDGEMENTS

I am particularly indebted to my academic supervisor Dr Daniel Rigden for overall supervision of the project, critically reading the manuscript and providing feedback throughout my research. I would like to thank Jens Thomas for helping set up my Linux workstation and Angela Rosin for technical assistance with setting up X2Go Client for remote access to my machine. Many thanks also go to Felix Simkovic, Jens Thomas and Mara Kozic for providing helpful suggestions and illuminating discussions in monthly lab meetings. I am also grateful to Dr Tancred Frickey at the University of Konstanz for making the CLANS software publically available and for answering technical questions with regards to the program via email. Research was performed at the University of Liverpool. The author declares no affiliations with any of the tools used.

REFERENCES

Abbas, C. A., & Sibirny, A. A. (2011). Genetic control of biosynthesis and transport of riboflavin and flavin nucleotides and construction of robust biotechnological producers. *Microbiology and Molecular Biology Reviews* **75**, 321-36.

Alland, C., Moreews, F., Boens, D., Carpentier, M., Chiusa, S., Lonquety, M., Renaultm N., Wong, Y., Cantalloube, H., Chomilier, J., Hochez, J., Pothier, J., Villoutreix, B. O., Zagury, J. F. & Tufféry, P. (2005). RPBS: a web resource for structural bioinformatics. *Nucleic Acids Research* **33**, 44-49.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403-410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389-3402.

Ashkenazy, H., Erez, E., Martz, E., Pupko, T. & Ben-Tal N. (2010). ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Research* **38**, 529-533.

Atkinson, H. J., Morris, J. H., Ferrin, T. E. & Babbitt, P. C. (2009). Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PloS one* **4**, 1-14.

Bacher, A. (1991). Biosynthesis of Flavins. In *Chemistry and Biochemistry of Flavoenzymes Volume 1* (Müller, F., Ed.), pp. 215-259. CRC Press, Boca Raton.

Bacher, A., Eberhardt, S., Fischer, M., Kis, K. & Richter, G. (2000). Biosynthesis of vitamin B2 (riboflavin). *Annual Review of Nutrition* **20**, 153-167.

Bacher, A., Eberhardt, S., Eisenreich, W., Fischer, M., Herz, S., Illarionov, B., Kis, K. & Richter, G. (2001). Biosynthesis of riboflavin. *Vitamins & Hormones* **61**, 1-49.

Bairoch, A. & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res*earch **28**, 45-48.

Bairoch, A., Boeckmann, B., Ferro, S. & Gasteiger, E. (2004). Swiss-Prot: juggling between evolution and stability. *Briefings in Bioinformatics* **5**, 39-55.

Baldwin, J. C., Karthikeyan, A. S. & Raghothama, K. G. (2001). LEPS2, a phosphorus starvation-induced novel acid phosphatase from tomato. *Plant Physiology* **125**, 728-737.

Bateman, A., Martin, M. J., O'Donovan, C., Magrane, M., Apweiler, R., Alpi, E., Antunes, R., Arganiska, J., Bely, B., Bingley, M., Bonilla, C., Britto, R., Bursteinas, B., Chavali, G., Cibrian-Uhalte, E., Silva, A. D., De Giorgi, M., Dogan, T., Fazzini, F., Gane, P., Castro, L. G., Garmiri, P., Hatton-Ellis, E., Hieta, R., Huntley, R., Legge, D., Liu, W., Luo, J., MacDougall, A., Mutowo, P., Nightingale, A., Orchard, S., Pichler, K., Poggioli, D., Pundir, S., Pureza, L., Qi, G., Rosanoff, S., Saidi, R., Sawford, T., Shypitsyna, A., Turner, E., Volynkin, V., Wardell, T., Watkins, X., Zellner, H., Cowley, A., Figueira, L., Li, W., McWilliam, H., Lopez, R., Xenarios, I., Bougueleret, L., Bridge, A., Poux, S., Redaschi, N., Aimo, L., Argoud-Puy, G., Auchincloss, A., Axelsen, K., Bansal, P., Baratin, D., Blatter, M. C., Boeckmann, B., Bolleman, J., Boutet, E., Breuza, L., Casal-Casas, C., de Castro, E., Coudert, E., Cuche, B., Doche, M., Dornevil, D., Duvaud, S., Estreicher, A., Famiglietti, L., Feuermann, M., Gasteiger, E., Gehant, S., Gerritsen, V., Gos, A., Gruaz-Gumowski, N., Hinz, U., Hulo, C., Jungo, F., Keller, G., Lara, V., Lemercier, P., Lieberherr, D., Lombardot, T., Martin, X., Masson, P., Morgat, A., Neto, T., Nouspikel, N., Paesano, S., Pedruzzi, I., Pilbout, S., Pozzato, M., Pruess, M., Rivoire, C., Roechert, B., Schneider, M., Sigrist, C., Sonesson, K., Staehli, S., Stutz, S., Sundaram, S., Tognolli, M., Verbregue, L., Veuthey, A. L., Wu, C. H., Arighi, C. N., Arminski, L., Chen, C., Chen, Y., Garavelli, J. S., Huang, H., Laiho, K., McGarvey, P., Natale, D. A., Suzek, B. E., Vinayaka, C., Wang, Q., Wang, Y., Yeh, L. S., Yerramalla, M. S. & Zhang, J. (2015). UniProt: a hub for protein information. *Nucleic Acids Research* **43**, 204-212.

Bazan, J. F., Fletterick, R. J. & Pilkis, S. J. (1989). Evolution of a Bifunctional Enzyme: 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase. *Proceedings of the National Academy of Sciences of the United States of America* **86**, 9642-9646.

Benkert, P., Tosatto, S. C. E. & Schomburg, D. (2008). QMEAN: A comprehensive scoring function for model quality assessment. *Proteins: Structure, Function, and Bioinformatics* **71**, 261-277.

Benkert, P., Künzli, M. & Schwede, T. (2009). QMEAN Server for Protein Model Quality Estimation. *Nucleic Acids Research* **37**, 510-514.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research* **28**, 235-242.

Biegert, A., Mayer, C., Remmert, M., Söding, J. & Lupas, A. N. (2006). The MPI Toolkit for protein sequence analysis. *Nucleic Acids Research* **34**, 335-339.

Birkenmeier, M., Neumann, S. & Röder, T. (2014). Kinetic modeling of riboflavin biosynthesis in Bacillus subtilis under production conditions. *Biotechnology letters* **36**, 919-928.

Blumenthal, T. (1998). Gene clusters and polycistronic transcription in eukaryotes. *Bioessays* **20**, 480-487.

Bourne, P. E., Addess, K. J., Bluhm, W. F., Chen, L., Deshpande, N., Feng, Z., Fleri, W., Green, R., Merino-Ott, J. C., Townsend-Merino, W., Weissig, H., Westbrook, J. & Berman, H. M. (2004). The distribution and query systems of the RCSB protein data bank. *Nucleic Acid Research* **32**, 223-225.

Boyd, S., Brookfield, J. L., Critchlow, S. E., Cumming, I. A., Curtis, N. J., Debreczeni, J., Degorce, S. L., Donald, C., Evans, N. J., Groombridge, S., Hopcroft**,** P., Jones, N. P., Kettle, J. G., Lamont, S., Lewis, H. J., MacFaull, P., McLoughlin, S. B., Rigoreau, L. J., Smith, J. M., St-Gallay, S., Stock, J. K., Turnbull, A. P., Wheatley, E. R., Winter, J. & Wingfield, J. (2015). Structure-based design of potent and selective inhibitors of the metabolic kinase PFKFB3. *Journal of medicinal chemistry* **58**, 3611-3625.

Burrows, R. B. & Brown, G. M. (1978). Presence of Escherichia coli of a deaminase and a reductase involved in biosynthesis of riboflavin. *Journal of Bacteriology* **136**, 657-667.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC bioinformatics* **10**, 421-429.

Celniker, G., Nimrod, G., Ashkenazy, H., Glaser, F., Martz, E., Mayrose, I., Pupko, T. & Ben-Tal N. (2013). ConSurf: Using Evolutionary Data to Raise Testable Hypotheses about Protein Function. *Israel Journal of Chemistry* **53**, 199-206.

Cho, K. H. & Salyers, A. A. (2001). Biochemical analysis of interactions between outer membrane proteins that contribute to starch utilization by Bacteroides thetaiotaomicron. *Journal of Bacteriology* **183**, 7224-7230.

Chothia, C. & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *The EMBO Journal* **5**, 823-826.

Colaco, C., Sen, S., Thangavelu, M., Pinder, S. & Roser, B. (1992). Extraordinary stability of enzymes dried in trehalose: simplified molecular biology. *Nature Biotechnology* **10**, 1007-1011.

Combs, S. A., Deluca, S. L., DeLuca, S. H., Lemmon, G. H., Nannemann, D. P., Nguyen, E. D., Willis, J. R., Sheehan, J. H. & Meiler, J. (2013). Small-molecule ligand docking into comparative models with Rosetta. *Nature Protocols* **8**, 1277-1298.

Daignan-Fornier, B. & Fink, G. R. (1992). Coregulation of purine and histidine biosynthesis by the transcriptional activators BAS1 and BAS2. *Proceedings of the National Academy of Sciences* **89**, 6746-6750.

Dandekar, T., Snel, B., Huynen, M. & Bork P. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. *Trends in Biochemical Sciences* **23**, 324-328.

D'Elia, J. N. & Salyers, A. A. (1996). Effect of regulatory protein levels on utilization of starch by Bacteroides thetaiotaomicron. *Journal of Bacteriology* **178**, 7180-7186.

De Virgilio, C., Piper, P., Boller, T. & Wiemken, A. (1991). Acquisition of thermotolerance in Saccharomyces cerevisiae without heat shock protein hsp104 and in the absence of protein synthesis. *FEBS letters* **288**, 86-90.

De Virgilio, C., Hottiger, T., Dominguez, J., Boller, T. & Wiemken, A. (1994). The role of trehalose synthesis for the acquisition of thermotolerance in yeast I**.** Genetic evidence that trehalose is a thermoprotectant. *European Journal of Biochemistry* **219**, 179-186.

Dick, C. F., Dos-Santos, A. L. A. & Meyer-Fernandes, J. R. (2011). Inorganic phosphate as an important regulator of phosphatases. *Enzyme Research* **11**, 1-7.

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics* **14**, 755-763.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792-1797.

Elliott, B., Haltiwanger, R. S. & Futcher, B. (1996). Synergy between trehalose and Hsp104 for thermotolerance in Saccharomyces cerevisiae. *Genetics* **144**, 923-933.

Emanuelsson, O., Nielsen, H., Brunak, S. & von Heijne, G. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *Journal of Molecular Biology* **300**, 1005-1016.

Emanuelsson, O., Brunak, S., von Heijne, G. & Nielsen, H. (2007). Locating proteins in the cell using TargetP, SignalP, and related tools. *Nature Protocols* **2**, 953-971.

Enright, A. J., Iliopoulos, I., Kyrpides, N. C. & Ouzounis, C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**, 86-90.

Finn, R. D., Clements, J. & Eddy S. R. (2011). HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Research* **39**, 29-37.

Fischer, M. & Bacher, A. (2005). Biosynthesis of flavocoenzymes. *Natural Product Reports* **22**, 324-350.

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Systematic Zoology* **19,** 99-113.

Flint, H. J., Bayer, E. A., Rincon, M. T., Lamed, R. & White, B. A. (2008). Polysaccharide utilization by gut bacteria: potential for new insights from genomic analysis. *Nature Reviews Microbiology* **6**, 121-131.

Frickey, T. & Lupas, A. N. (2004). CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics* **20**, 3702-3704.

Frickey, T. & Weiller, G. (2007). Analyzing microarray data using CLANS. *Bioinformatics* **23**, 1170-1171.

Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. (2012). CD-HIT: accelerated for clustering the next generation sequencing data. *Bioinformatics* **28**, 3150-3152.

Gabaldón, T. & Huynen, M. A. (2004). Prediction of protein function and pathways in the genome era. *Cellular and Molecular Life Sciences: CMLS.* **61**, 930-944.

Garg, A. & Raghava, G. P. S. (2008). A machine learning based method for the prediction of secretory proteins using amino acid composition, their order and similarity-search. *In Silico Biology* **8**, 129-140.

Gentry, M. S., Dowen, R. H., Worby, C. A., Mattoo, S., Ecker, J. R. & Dixon, J. E. (2007). The phosphatase laforin crosses evolutionary boundaries and links carbohydrate metabolism to neuronal disease. *The Journal of Cell Biology* **178**, 477-488.

Gentry, M. S., Roma-Mateo, C. & Sanz, P. (2013). Laforin, a protein with many faces: glucan phosphatase, adapter protein, et alii. *FEBS Journal* **280**, 525-537.

Gerdes, S., Lerma-Ortiz, C., Frelin, O., Seaver, S. M. D., Henry, C. S., de Crécy-Lagard, V. & Hanson, A. D. (2012). Plant B vitamin pathways and their compartmentation: a guide for the perplexed. *Journal of Experimental Botany* **63**, 5379-5395.

Gerlt, J. A., Babbitt, P. C., Jacobson, M. P. & Almo, S. C. (2012). Divergent Evolution in Enolase Superfamily: Strategies for Assigning Functions. *Journal of Biological Chemistry* **287**, 29-34.

Ginalski, K., Zhang, H. & Grishin, N. V. (2004). Raptor protein contains a caspase-like domain. *Trends in Biochemical Science* **29**, 522-524.

Glaser, F., Pupko, T., Paz, I., Bell, R. E., Bechor, D., Martz, E. & Ben-Tal, N. (2003). ConSurf: Identification of Functional Regions in Proteins by Surface-Mapping of Phylogenetic Information. *Bioinformatics* **19**, 163-164.

Goldberg, T., Hecht, M., Hamp, T., Karl, T., Yachdav, G., Ahmed, N., Altermann, U., Angerer, P., Ansorge, S., Balasz, K., Bernhofer, M., Betz, A., Cizmadija, L., Do, K. T., Gerke, J., Greil, R., Joerdens, V., Hastreiter, M., Hembach, K., Herzog, M., Kalemanov, M., Kluge, M., Meier, A., Nasir, H., Neumaier, U., Prade, V., Reeb, J., Sorokoumov, A., Troshani, I., Vorberg, S., Waldraff, S., Zierer, J., Nielsen, H. & Rost, B. (2014). LocTree3 prediction of localization. *Nucleic Acids Research* **42**, 350-355.

Hadjigeorgiou, G. M., Bruno, C., Andreu, A. L., Sue, C. M., Shanske, S., DiMauro, S., Kawashima, N., Kawashima, A. & Rigden, D. J. (1999). Manifesting heterozygotes in a Japanese family with a novel mutation in the muscle-specific phosphoglycerate mutase (PGAM-M) gene. *Neuromuscular Disorders* **9**, 399-402.

Hamada, K., Kato, M., Shimizu, T., Ihara, K., Mizuno, T. & Hakoshima, T. (2005). Crystal structure of the protein histidine phosphatase SixA in the multistep His-Asp phosphorelay. *Genes to Cells* **10**, 1-11.

Hasnain, G., Frelin, O., Roje, S., Ellens, K. W., Ali, K., Guan, J. C., Garrett, T. J., de Crécy-Lagard, V., Gregory, J. F., McCarty, D. R. & Hanson, A. D. (2013). Identification and characterization of the missing pyrimidine reductase in the plant riboflavin biosynthesis pathway. *Plant Physiology* **161**, 48-56.

Hiller, K., Grote, A., Scheer, M., Münch, R. & Jahn, D. (2004). PrediSi: prediction of signal peptides and their cleavage positions. *Nucleic Acids Research* **32**, 375-379.

Hiltunen, H. M., Illarionov, B., Hedtke, B., Fischer, M. & Grimm, B. (2012). Arabidopsis RIBA Proteins: Two out of Three Isoforms Have Lost Their Bifunctional Activity in Riboflavin Biosynthesis. *International Journal of Molecular Sciences* **13**, 14086-14105.

Hooper, L. V., Midtvedt, T. & Gordon, J. I. (2002): How host-microbial interactions shape the nutrient environment of the mammalian intestine. *Annual Review of Nutrition* **22**, 283-307.

Hothorn, M., Neumann, H., Lenherr, E. D., Wehner, M., Rybin, V., Hassa, P. O., Uttenweiler, A., Reinhardt, M., Schmidt, A., Seiler, J., Ladurner, A. G., Herrmann, C., Scheffzek, K. & Mayer, A. (2009). Catalytic core of a membrane-associated eukaryotic polyphosphate polymerase. *Science* **324**, 513-516.

Hottiger, T., De Virgilio, C., Bell, W., Boller, T. & Wiemken, A. (1992). The 70-kilodalton heat-shock proteins of the SSA subfamily negatively modulate heat-shock-induced accumulation of trehalose and promote recovery from heat stress in the yeast, Saccharomyces cerevisiae. *European Journal of Biochemistry* **210**, 125-132.

Hottiger, T., De Virgilio, C., Hall, M. N., Boller, T. & Wiemken, A. (1994). The role of trehalose synthesis for the acquisition of thermotolerance in yeast. II. Physiological concentrations of trehalose increase the thermal stability of proteins in vitro. *European Journal of Biochemistry* **219**, 187-193.

Huang, F., Li, Y., Spencer, J. B., Haydock, S. F., Mironenko, T. & Spiteller, D. (2005). The neomycin biosynthetic gene cluster of Streptomyces fradiae NCIMB 8233: Characterisation of an aminotransferase involved in the formation of 2-deoxystreptamine. *Organic and Biomolecular Chemistry* **8**, 1410-1418.

Imai, K., Asakawa, N., Tsuji, T., Akazawa, F., Ino, A., Sonoyama, M. & Mitaku, S. (2008). SOSUI-GramN: high performance prediction for sub-cellular localization of proteins in Gram-negative bacteria. *Bioinformation* **2**, 417-421.

Käll, L., Krogh, A. & Sonnhammer, E. L. L. (2004). A Combined Transmembrane Topology and Signal Peptide Prediction Method. *Journal of Molecular Biology* **338**, 1027-1036.

Käll, L., Krogh, A. & Sonnhammer, E. L. L. (2007). Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. *Nucleic Acids Research* **35**, 429-432.

Kanehisa, M. & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids* Research **28**, 27-30.

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Research* **42**, 199-205.

Kim, T., Porres, J. M., Roneker, K. R., Crowe, S., Rice, S., Ko, T., Lei, X. G., Mullaney, E. J., Ullah, A. H. J., Daly, C. B. & Welch, R. (2006). Shifting the pH profile of Aspergillus niger PhyA phytase to match the stomach pH enhances its effectiveness as an animal feed additive. *Applied and Environmental Microbiology* **72**, 4397-4403.

Kornberg, A. (1999). Inorganic polyphosphate: a molecule of many functions. *Progress in Molecular and Subcellular Biology* **23**, 1-18.

Koropatkin, N. M., Cameron, E. A. & Martens, E. C. (2012). How glycan metabolism shapes the human gut microbiota. *Nature Reviews Microbiology* **10**, 323-335.

Kotting, O., Pusch, K., Tiessen, A., Geigenberger, P., Steup, M. & Ritte, G. (2005). Identification of a novel enzyme required for starch metabolism in Arabidopsis leaves. The phosphoglucan, water dikinase. *Plant Physiology* **137**, 242-252.

Kotting, O., Santelia, D., Edner, C., Eicke, S., Marthaler, T., Gentry, M. S., Comparot-Moss, S., Chen, J., Smith, A. M., Steup, M., Ritte, G. & Zeeman, S. C. (2009). STARCH-EXCESS4 is a laforin-like Phosphoglucan phosphatase required for starch degradation in Arabidopsis thaliana. *Plant Cell* **21**, 334-346.

Kotting, O., Kossmann, J., Zeeman, S. C. & Lloyd, J. R. (2010). Regulation of starch metabolism: the age of enlightenment? *Current Opinion in Plant Biology* **13**, 321-329.

Landau, M., Mayrose, I., Rosenberg, Y., Glaser, F., Martz, E., Pupko, T. & Ben-Tal, N. (2005). ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Research* **33**, 299-302.

Ledesma-Amaro, R., Kerkhoven, E. J., Revuelta, J. L. & Nielsen, J. (2014). Genome scale metabolic modeling of the riboflavin overproducer Ashbya gossypii. *Biotechnology and Bioengineering* **111**, 1191-1199.

Lee, D., Redfern, O. & Orengo, C. (2007). Predicting protein function from sequence and structure. *Nature Reviews Molecular Cell Biology* **8**, 995-1005.

Lee, J. M. & Sonnhammer, E. L. (2003). Genomic gene clustering analysis of pathways in eukaryotes. *Genome research* **13**, 875-882.

Lemarchandel, V., Joulin, V., Valentin, C., Rosa, R., Galacteros, F., Rosa, J. & Cohen- Solal, M. (1992). Compound heterozygosity in a complete erythrocyte bisphosphoglycerate mutase deficiency. *Blood* **80**, 2643-2649.

Li, W. & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659.

Li, W., McWilliam, H., Goujon, M., Cowley, A., Lopez, R. & Pearson, W. R. (2012). PSI-Search: iterative HOE-reduced profile SSEARCH searching. *Bioinformatics* **28**, 1650-1651.

Loewenstein, Y., Raimondo, D., Redfern, O. C., Watson, J., Frishman, D., Linial, M., Orengo, C., Thornton, J. & Tramontano, A. (2009). Protein function annotation by homology-based inference. *Genome Biology* **10**, 207-215.

Lukk, T., Sakai, A., Imker, H. J., Song, L., Nair, S. K., Almo, S. C., Gerlt, J. A., Kalyanaraman, C., Brown, S. D., Babbitt, P. C., Jacobson, M. P., Fedorov, A. A., Fedorov, E. V., Toro, R., Hillerich, B., Seidel, R., Patskovsky, Y. & Vetting, M. W. (2012). Homology models guide discovery of diverse enzyme specificities among dipeptide epimerases in the enolase superfamily. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 4122-4127.

Lyskov, S., Chou, F. C., Conchúir, S. Ó., Der, B, S., Drew, K., Kuroda, D., Xu, J., Weitzner, B. D., Renfrew, P. D., Sripakdeevong, P., Borgo, B., Havranek, J. J., Kuhlman, B., Kortemme, T., Bonneau, R., Gray, J. J. & Das, R. (2013). Serverification of Molecular Modeling Applications: The Rosetta Online Server That Includes Everyone (ROSIE). *PLoS One* **8**, 1-11.

Marchler-Bauer, A., Derbyshire, M. K., Gonzales, N. R., Lu, S., Chitsaz, F., Geer, L. Y., Geer, R. C., He, J., Gwadz, M., Hurwitz, D. I., Lanczycki, C. J., Lu, F., Marchler, G. H., Song, J. S., Thanki, N., Wang, Z., Yamashita, R. A., Zhang, D., Zheng, C. & Bryant, S. H. (2015). CDD: NCBI's conserved domain database. *Nucleic Acids Research* **43**, 222-226.

Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O. & Eisenberg, D. (1999a). Detecting Protein Function and Protein-Protein Interactions from Genome Sequences. *Science* **285**, 751-753.

Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O. & Eisenberg, D. (1999b). A combined algorithm for genome-wide prediction of protein function. *Nature* **402**, 83-86.

Martens, E. C., Koropatkin, N. M., Smith, T. J. & Gordon, J. I. (2009). Complex glycan catabolism by the human gut microbiota: the Bacteroidetes Sus-like paradigm. *Journal of Biological Chemistry* **284**, 24673-24677.

Martens, E. C., Lowe, E. C., Chiang, H., Pudlo, N. A., Wu, M., McNulty, N. P., Abbott, D. W., Henrissat, B., Gilbert, H. J., Bolam, D. N. & Gordon, J. I. (2011). Recognition and degradation of plant cell wall polysaccharides by two human gut symbionts. *PLoS-Biology* **9**, 1-16.

Matsuda, S., Vert, J. P., Saigo, H., Ueda, N., Toh, H. & Akutsu, T. (2005). A novel representation of protein sequences for prediction of subcellular location using support vector machines. *Protein Science* **14**, 2804-2813.

McNeil, N. I. (1984). The contribution of the large intestine to energy supplies in man. *The American Journal of Clinical Nutrition* **39**, 338-342.

Meekins, D. A., Raththagala, M., Auger, K. D., Turner, B. D., Santelia, D., Kötting, O., Gentry, M. S. & Vander Kooi, C. W. (2015). Mechanistic Insights into Glucan Phosphatase Activity Against Polyglucan Substrates. *Journal of Biological Chemistry* [31 July 2015 Epub ahead of print]

Mello, L. V. & Rigden, D. J. (2012). A new family of bacterial DNA repair proteins annotated by the integration of non-homology, distant homology and structural bioinformatic methods. *FEBS letters* **586**, 3908-3913.

Mello, L. V., Chen, X. & Rigden, D. J. (2010). Mining metagenomic data for novel domains: BACON, a new carbohydrate-binding module. *FEBS letters* **584**, 2421-2426.

Miteva, M. A., Guyon, F. & Tufféry, P. (2010). Frog2: Efficient 3D conformation ensemble generator for small compounds. *Nucleic Acids Research* **38**, 622-627.

Moreno-Hagelsieb, G., Trevino, V., Perez-Rueda, E., Smith, T. F. & Collado-Vides, J. (2001). Transcription unit conservation in the three domains of life: a perspective from *Escherichia coli*. *TRENDS in Genetics* **17**, 175-177.

Mouillon, J. M. & Persson, B. L. (2006). New aspects on phosphate sensing and signalling in Saccharomyces cerevisiae. *FEMS Yeast Research* 6, 171-176.

Müller, O., Bayer, M. J., Peters, C., Andersen, J. S., Mann, M. & Mayer, A. (2002). The Vtc proteins in vacuole fusion: coupling NSF activity to V0 trans-complex formation. *The EMBO Journal* **21**, 259-269.

Müller, O., Neumann, H., Bayer, M. J. & Mayer, A. (2003). Role of the Vtc proteins in V-ATPase stability and membrane trafficking. *Journal of Cell Science* **116**, 1107-1115.

Nakai, K. & Horton, P. (1999). PSORT: a program for detecting the sorting signals of proteins and predicting their subcellular localization. *Trends in Biochemical Science* **24**, 34-35.

Neidhart, D. J., Kenyon, G. L., Gerlt, J. A. & Petsko, G. A. (1990). Mandelate racemase and muconate lactonizing enzyme are mechanistically distinct and structurally homologous. *Nature* **347**, 692-694.

Néron, B., Ménager, H., Maufrais, C., Joly, N., Maupetit, J., Letort, S., Carrere, S., Tuffery, P. & Letondal, C. (2009). Mobyle: a new full web bioinformatics framework. *Bioinformatics* **25**, 3005-3011.

Neves, M. J. & François, J. (1992). On the mechanism by which a heat shock induces trehalose accumulation in Saccharomyces cerevisiae. *Biochemical Journal* **288**, 859-864.

Ogawa, N., DeRisi, J. & Brown, P. O. (2000). New components of a system for phosphate accumulation and polyphosphate metabolism in Saccharomyces cerevisiae revealed by genomic expression analysis. *Molecular Biology of the Cell* **11**, 4309-4321.

Ortlund, E., LaCount, M. W. & Lebioda, L. (2003). Crystal structures of human prostatic acid phosphatase in complex with a phosphate ion and α-benzylaminobenzylphosphonic acid update the mechanistic picture and offer new insights into inhibitor design. *Biochemistry* **42**, 383-389.

Oshima, Y. (1997). The phosphatase system in Saccharomyces cerevisiae. *Genes & Genetic Systems* **72**, 323-334.

Osiewacz, H. D. (2002). The Mycota: A Comprehensive Treatise on Fungi as Experimental Systems for Basic and Applied Research. Volume X. Industrial Applications. (Esser, K. & Bennett, J. W., Eds.), pp. 235-237. Springer-Verlag, Berlin Heidelberg, New York.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D. & Maltsev, N. (1998). Use of contiguity on the chromosome to predict functional coupling. *In Silico Biology* **1**, 93-108.

Palanichelvam, K., Oger, P., Clough, S. J., Cha, C., Bent, A. F. & Farrand, S. K. (2000). A second T-region of the soybean-supervirulent chrysopine-type Ti plasmid pTiChry5, and construction of a fully disarmed vir helper plasmid. *Molecular Plant-Microbe Interactions* **13**, 1081-1091.

Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences* **96**, 4285-4288.

Petersen, T. N., Brunak, S., von Heijne, G. & Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods* **8**, 785-786.

Powers, H. J. (2003). Riboflavin (vitamin B-2) and health. *The American Journal of Clinical Nutrition* **77**, 1352-1360.

Pugsley, A. P., Francetic, O., Possot, O. M., Sauvonnet, N. & Hardie, K. R. (1997). Recent progress and future directions in studies of the main terminal branch of the general secretory pathway in Gram-negative bacteria–a review. *Gene* **192**, 13-19.

Radivojac, P., Clark, W. T., Oron, T. R., Wittkop, T., Mooney, S. D., Schnoes, A. M., Babbitt, P. C., Sokolov, A., Graim, K., Ben-Hur, A., Funk, C., Verspoor, K., Pandey, G., Repo, S., Yunes, J. M., Talwalkar, A. S., Souza, M. L., Piovesan, D., Casadio, R., Wang, Z., Cheng, J., Fang, H., Gough, J., Koskinen, P., Törönen, P., Nokso-Koivisto, J., Holm, L., Cozzetto, D., Buchan, D. W. A., Bryson, K., Jones, D. T., Limaye, B., Inamdar, H., Datta, A., Manjari, S. K., Joshi, R., Chitale, M., Kihara, D., Lisewski, A. M., Erdin, S., Venner, E., Lichtarge, O., Rentzsch, R., Orengo, C., Yang, H., Romero, A. E., Bhat, P., Paccanaro, A., Hamp, T., Kaßner, R., Seemayer, S., Vicedo, E., Schaefer, C., Achten, D., Auer, F., Boehm, A., Braun, T., Hecht, M., Heron, M., Hönigschmid, P., Hopf, T. A., Kaufmann, S., Kiening, M., Krompass, D., Landerer, C., Mahlich, Y., Roos, M., Rost, B., Björne, J., Salakoski, T., Wong, A., Shatkay, H., Gatzmann, F., Sommer, I., Wass, M. N., Sternberg, M. J. E., Škunca, N., Supek, F., Bošnjak, M., Šmuc, T., Panov, P., Džeroski, S., Kourmpetis, Y. A. I., van Dijk, A. D. J., Braak, C. J. F., Zhou, Y., Gong, Q., Dong, X., Tian, W., Falda, M., Lavezzo, E., Toppo, S., Fontana, P., Di Camillo, B., Lan, L., Djuric, N., Guo, Y., Vucetic, S., Bairoch, A., Linial, M., Brenner, S. E. & Friedberg, I. (2013). A large-scale evaluation of computational protein function prediction. *Nature Methods* **10**, 221-227.

Ravcheev, D. A., Godzik, A., Osterman, A. L. & Dmitry, A. R. (2013). Polysaccharides utilization in human gut bacterium Bacteroides thetaiotaomicron: comparative genomics reconstruction of metabolic and regulatory networks. *BMC Genomics* **14**, 873-890.

Rébora, K., Laloo, B. & Daignan-Fornier, B. (2005). Revisiting Purine-Histidine Cross-Pathway Regulation in Saccharomyces cerevisiae A Central Role for a Small Molecule. *Genetics* **170**, 61-70.

Reeves, A. R., Wang, G. R. & Salyers, A. A. (1997). Characterization of four outer membrane proteins that play a role in utilization of starch by Bacteroides thetaiotaomicron. *Journal of Bacteriology* **179**, 643-649.

Reeves, G. A., Dallman, T. J., Redfern, O. C., Akpor, A. & Orengo, C. A. (2006). Structural diversity of domain superfamilies in the CATH database. *Journal of Molecular Biology* **360**, 725-741.

Rigden, D. J., Littlejohn, J. E., Henderson, K. & Jedrzejas, M. J. (2003). Structures of phosphate and trivanadate complexes of Bacillus stearothermophilus phosphatase PhoE: structural and functional analysis in the cofactor-dependent phosphoglycerate mutase superfamily. *Journal of Molecular Biology* **325**, 411-420.

Rigden, D. J. (2008). The histidine phosphatase superfamily: structure and function. *Biochemical Journal* **409**, 333-348.

Roje, S. (2007). Vitamin B biosynthesis in plants. *Phytochemistry* **68**, 1904-1921.

San Luis, B., Nassar, N. & Carpino, N. (2013). New insights into the catalytic mechanism of histidine phosphatases revealed by a functionally essential arginine residue within the active site of the Sts phosphatases. *Biochemical Journal* **453**, 27-35.

Santelia, D., Kotting, O., Seung, D., Schubert, M., Thalmann, M., Bischof, S., Meekins, D. A., Lutz, A., Patron, N., Gentry, M. S., Allain, F. H. & Zeeman, S. C. (2011). The phosphoglucan phosphatase like sex four2 dephosphorylates starch at the c3-position in Arabidopsis. *Plant Cell* **23**, 4096-4111.

Secco, D., Wang, C., Shou, H. & Whelan, J. (2012). Phosphate homeostasis in the yeast Saccharomyces cerevisiae, the key role of the SPX domain-containing proteins. *FEBS letters* **586**, 289-295.

Shakarian, A. M., Joshi, M. B., Yamage, M., Ellis, S. L., Debrabant, A. & Dwyer, D. M. (2003). Members of a unique histidine acid phosphatase family are conserved amongst a group of primitive eukaryotic human pathogens. *Molecular and cellular biochemistry* **245**, 31-41.

Shen, H. B. & Chou, K. C. (2007). Signal-3L: A 3-layer approach for predicting signal peptides. *Biochemical and Biophysical Research Communications* **363**, 297-303.

Shen, H. B, & Chou, K. C. (2010). Gneg-mPLoc: a top-down strategy to enhance the quality of predicting subcellular localization of Gram-negative bacterial proteins. *Journal of Theoretical Biology* **264**, 326-333.

Shi, S., Chen, T., Zhang, Z., Chen, X. & Zhao, X. (2009). Transcriptome analysis guided metabolic engineering of Bacillus subtilis for riboflavin production. *Metabolic Engineering* **11**, 243-252.

Shipman, J. A., Berleman, J. E. & Salyers, A. A. (2000). Characterization of four outer membrane proteins involved in binding starch to the cell surface of Bacteroides thetaiotaomicron. *Journal of Bacteriology* **182**, 5365-5372.

Silver, D. M., Kotting, O. & Moorhead, G. B. G. (2014). Phosphoglucan phosphatase function sheds light on starch degradation. *Trends in Plant Science* **19**, 471-478.

Slavin, I., Saura, A., Carranza, P. G., Touz, M. C., Nores, M. J. & Luján, H. D. (2002). Dephosphorylation of cyst wall proteins by a secreted lysosomal acid phosphatase is essential for excystation of Giardia lamblia. *Molecular and biochemical parasitology* **122**, 95-98.

Snel, B., Lehmann, G., Bork, P. & Huynen, M. A, (2000). STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic acids research* **28**, 3442-3444.

Söding, J., Biegert, A. & Lupas, A. N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research* **33**, 244-248.

Sonnenburg, E. D., Zheng, H., Joglekar, P., Higginbottom, S. K., Firbank, S. J., Bolam, D. N. & Sonnenburg, J. L. (2010). Specificity of polysaccharide use in intestinal bacteroides species determines diet-induced microbiota alterations. *Cell* **141**, 1241-1252.

Stahmann, K. P., Revuelta, J. L. & Seulberger, H. (2000). Three biotechnical processes using Ashbya gossypii, Candida famata or Bacillus subtilis compete with chemical riboflavin production. *Applied Microbiology and Biotechnology* **53**, 509-516.

Stentz, R., Osborne, S., Horn, N., Li, A. W., Hautefort, I., Bongaerts, R., Rouyer, M., Bailey, P., Shears, S. B., Hemmings, A. M., Brearley, C. A. & Carding, S. R. (2014). A bacterial homolog of a eukaryotic inositol phosphate signaling enzyme mediates cross-kingdom dialog in the mammalian gut. *Cell Reports* **6**, 646-656.

Suhre, K. (2007). Inference of gene function based on gene fusion events: the Rosetta-stone method. *Methods in Molecular Biology* **396**, 31-41.

Suzek, B. E., Huang, H., McGarvey, P., Mazumder. R. & Wu, C. H. (2007). UniRef: Comprehensive and Non-Redundant UniProt Reference Clusters *Bioinformatics* **23**, 1282-1288.

Szklarczyk, D., Jensen, L. J., Franceschini, A., Simonovic, M., Roth, A., Stark, M., von Mering, C., Kuhn, M., Minguez, P., Doerks, T., Bork, P. & Muller, J. (2011). The STRING database in 2011: Functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research.* **39**, 561-568.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., Kuhn, M., Bork, P., Jensen, L. J. & von Mering, C. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research* **43**, 447-452.

Takeda, Y. & Hizukuri, S. (1981). Re-examination of the action of sweet-potato beta-amylase on phophorylated (1->4)-a-D-glucan. *Carbohydrate Research* **89**, 174-178.

Tancula, E., Feldhaus, M. J., Bedzyk, L. A. & Salyers, A. A. (1992). Location and characterization of genes involved in binding of starch to the surface of Bacteroides thetaiotaomicron. *Journal of Bacteriology* **174**, 5609-5616.

Teichmann, S. A. & Babu, M. M. (2002). Conservation of gene co-regulation in prokaryotes and eukaryotes. *Trends in Biotechnology* **20**, 407-410.

Teichmann, S. A. (2002). The constraints protein-protein interactions place on sequence divergence. *Journal of molecular biology* **324,** 399-407.

Tibbetts, A. S., & Appling, D. R. (2000). Characterization of two 5-aminoimidazole-4-carboxamide ribonucleotide transformylase/inosine monophosphate cyclohydrolase isozymes from Saccharomyces cerevisiae. *Journal of Biological Chemistry* **275**, 20920-20927.

Uttenweiler, A., Schwarz, H., Neumann, H. & Mayer, A. (2007). The vacuolar transporter chaperone (VTC) complex is required for microautophagy. *Molecular Biology of the Cell* **18**, 166-175.

Wang, Y., Liu, L., Wei, Z., Cheng, Z., Lin, Y. & Gong, W. (2006). Seeing the process of histidine phosphorylation in human bisphosphoglycerate mutase. *Journal of Biological Chemistry* **281**, 39642-39648.

Waterhouse, A. M., Procter, J. B., Martin, D. M. A, Clamp, M., Barton, G. J. (2009). Jalview version 2: A Multiple Sequence Alignment and Analysis Workbench. *Bioinformatics* **25**, 1189-1191.

Worby, C. A., Gentry, M. S. & Dixon, J. E. (2006). Laforin, a dual specificity phosphatase that dephosphorylates complex carbohydrates. *Journal of Biological Chemistry* **281**, 30412-30418.

Xiang, T., Liu, Q., Deacon, A. M., Koshy, M., Kriksunov, I. A., Lei, X. G., Hao, Q. & Thiel, D. J. (2004). Crystal structure of a heat-resilient phytase from Aspergillus fumigatus, carrying a phosphorylated histidine. *Journal of Molecular Biology* **339**, 437-445.

Xu, J., Bjursell, M. K., Himrod, J., Deng, S., Carmichael, L. K., Chiang, H. C., Hooper, L. V. & Gordon, J. I. (2003). A genomic view of the human-Bacteroides thetaiotaomicron symbiosis. *Science* **299**, 2074-2076.

Yu, N. Y., Wagner, J. R., Laird, M. R., Melli, G., Rey, S., Lo, R., Dao, P., Sahinalp, S. C., Ester, M., Foster, L. J., Brinkman, F. S. L. (2010). PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* **26**, 1608-1615.

Zhao, S., Kumar, R., Sakai, A., Vetting, M. W., Wood, B. M., Brown, S., Bonanno, J. B., Hillerich, B. S., Seidel, R. D., Babbitt, P. C., Almo, S. C., Sweedler, J. V., Gerlt, J. A., Cronan, J. E. & Jacobson, M. P. (2013). Discovery of new enzymes and metabolic pathways by using structure and genome context *Nature* **502**, 698-702.

Zheng, Q., Jiang, D., Zhang, W., Zhang, Q., Zhao, Q., Jin, J., Li, X., Yang, H., Bartlam, M., Shaw, N., Zhou, W. & Rao, Z. (2014). Mechanism of Dephosphorylation of Glucosyl-3-phosphoglycerate by a Histidine Phosphatase. *Journal of Biological Chemistry* **289**, 21242-21251.

APPENDICES

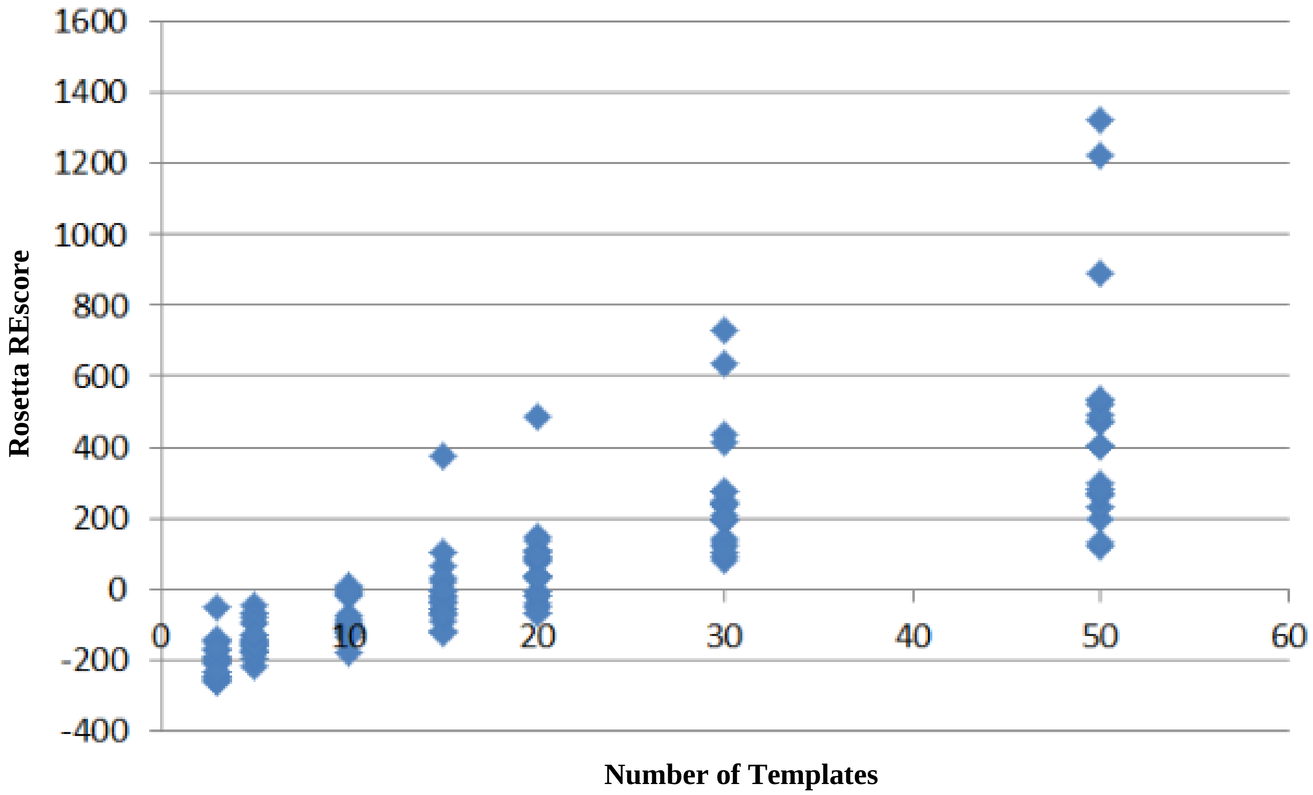
Appendix 1

The HP branch 1 representative sequences used as queries to iteratively search the UniRef90 database (Bateman *et al*., 2015; Suzek *et al*., 2011) using Jackhmmer (Eddy, 1998; Finn *et al*., 2011).

Accession Number	**Entry Name**	**Protein**	**Gene**	**Organism**
P00950	PMG1_YEAST	Phosphoglycerate mutase 1	GPM1	*Saccharomyces cerevisiae*
P07738	PMGE_HUMAN	Bisphosphoglycerate mutase	BPGM	*Homo sapiens*
P36069	PMU1_YEAST	Probable phosphomutase PMU1	PMU1	*Saccharomyces cerevisiae*
P76502	SIXA_ECOLI	Phosphohistidine phosphatase SixA	sixA	*Escherichia coli*
Q7YTB0	Q7YTB0_BOMMO	Ecdysteroid-phosphate phosphatase	N/A	*Bombyx mori*
Q96HS1	PGAM5_HUMAN	Serine/threonine-protein phosphatase PGAM5, mitochondrial	PGAM5	*Homo sapiens*
O43980	M1PAS_EIMTE	Mannitol-1-phosphatase	N/A	*Eimeria tenella*
Q9NQ88	TIGAR_HUMAN	Fructose-2,6-bisphosphatase TIGAR	TIGAR	*Homo sapiens*

Appendix 2

Determining the optimum number of templates to use for homology modelling using Rosetta REscore

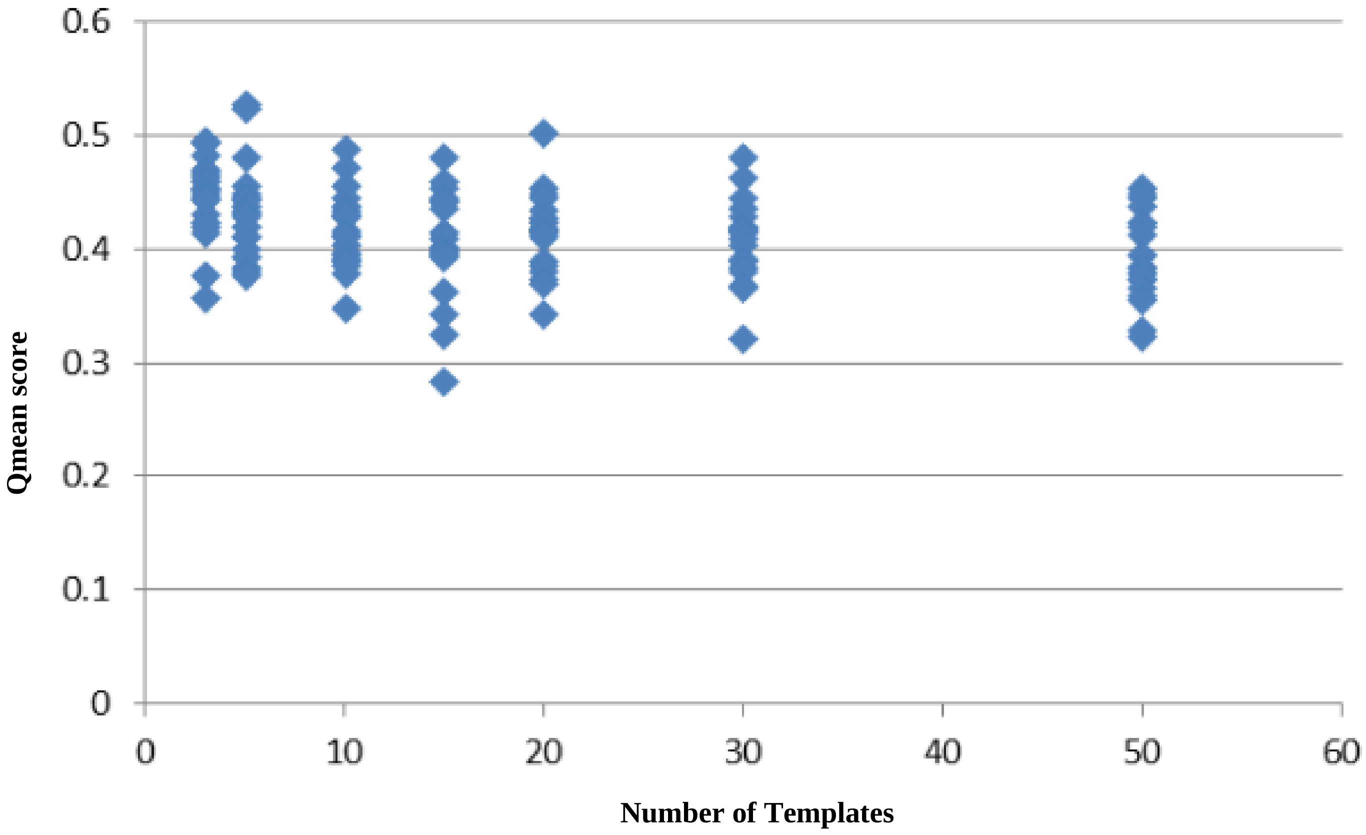


The above graph shows the number of templates (3, 5, 10, 15, 20, 30 and 50) used to create the preliminary homology models of a representative member of cluster 6541 (Accession Number G3AXW8), against Rosetta's REscore for each model produced.

Higher quality models have lower Rosetta REscores (Combs *et al*., 2013). As the number of templates used for modelling increases the quality of the models produced decreases.

Appendix 3

Determine the optimum number of templates to use for homology modelling using Qmean scores

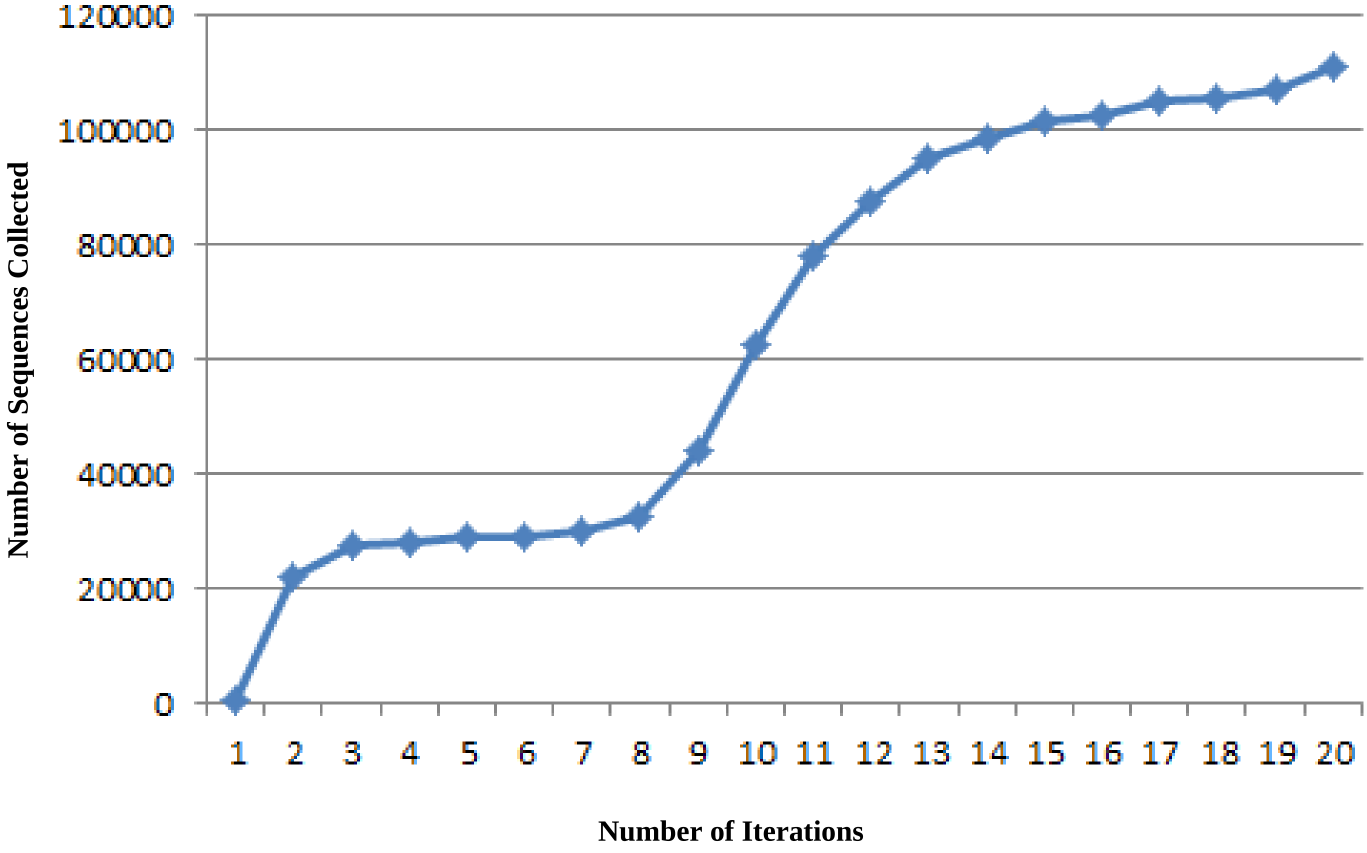


The above graph shows the number of templates (3, 5, 10, 15, 20, 30 and 50) used to create the preliminary homology models of a representative member of cluster 6541 (Accession Number G3AXW8), against Qmean scores for each model produced.

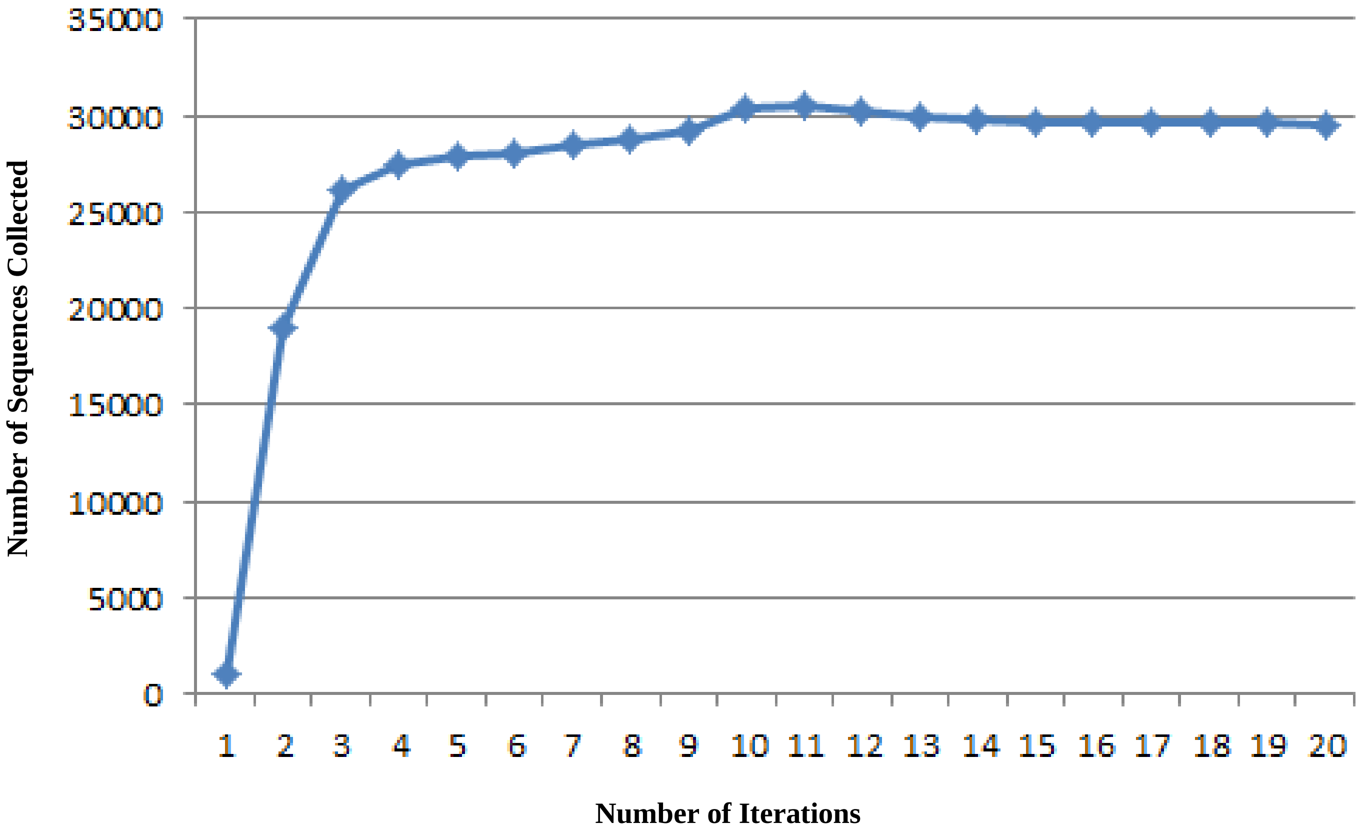
Higher quality models have higher Qmean scores (Benkert *et al*., 2008). An increase from 3 to 5 templates improves the models scores, however further increases in the number of templates has a negative effect on the quality of the models produced.

Appendix 4

Number of iterations performed by Jackhmmer against the number of sequences collected from the UniRef90 database, using *Homo sapiens* PGAM5 (Accession number Q96HS1) as the query.



Number of iterations performed by Jackhmmer against the number of sequences collected from the UniRef90 database, using *Escherichia coli* SixA (Accession number P76502) as the query.


Number of Sequences Collected
Number of Iterations