Forschungszentrum Karlsruhe
in der Helmholtz-Gemeinschaft

Fachhochschule
Weihenstephan
University of Applied Sciences

A comparative analysis of protein-protein interactions involved in bacterial motility

Johannes Goll

Diplomarbeit zur Erlangung des akademischen Grades Dipl.-Ing. (FH)

Institut für Toxikologie und Genetik
Forschungszentrum Karlsruhe

Fachhochschule Weihenstephan
University of Applied Sciences Weihenstephan

Fachbereich Biotechnologie und Bioinformatik
Studiengang Bioinformatik

Betreuer: Prof. Dr. Bernhard Haubold
Prof. Dr. habil. Diethard Baron

Eingereicht am 28. Juni 2006

Erklärung zur Urheberschaft

Gemäß § 31 Abs. 7 der Rahmenprüfungsordnung für die Fachhochschulen (RaPO):

Ich erkläre hiermit, dass die vorliegende Arbeit von mir selbst und ohne fremde Hilfe verfasst und noch nicht anderweitig für Prüfungszwecke vorgelegt wurde.

Es wurden keine anderen als die angegebenen Quellen oder Hilfsmittel benutzt. Wörtliche und sinngemäße Zitate sind als solche gekennzeichnet.

Freising, den 28.06.2006

.....
Johannes Goll

Acknowledgments

I would like to say thanks to all the people from the Uetz Lab, especially to the 'flagellum guys', Björn Titz and Dr. Rajagopala S. V. for their relentless support, fruitful discussions and proofreading. My special thanks to Wolf-Gerolf Thies for helping me out with server problems. I express my sincere thanks to Prof. Dr. Bernhard Haubold for extensive comments on an earlier version of this thesis. Also to Christian Zinsmeister for suggestions and corrections during the final phase of writing. Finally, I would like to thank Dr. Peter Uetz for his confidence and support.

Summary

Analysis of protein-protein interactions promise to reveal new insights into bacterial locomotion. Although elementary components of bacterial motility, such as the flagellum and the chemotaxis signaling pathway, are well characterized, it is not clear if all components and functional links have been identified. Recently, high-throughput yeast two-hybrid and complex purification studies identified protein-protein interactions in four pathogenic Gram-negative bacteria: the syphilis spirochete *Treponema pallidum*, *Campylobacter pylori*, the gastritis causing *Helicobacter pylori* and the well-studied model bacteria *Escherichia coli*. Motility-centered subsets (170–580 interactions) were pairwise compared, graph theoretically characterized, and functionally classified. To measure reliability, biological relevance as well as to identify potential biases between the two experimental methods, networks were compared with each other, with a reference set of previously reported motility interactions, and with predicted associations derived from three-dimensional structures and genomic context. In the course of this thesis, I identified an evolutionary conserved core comprising 94 interactions among 65 orthologous protein families. Integration of genomic context predictions, genes with motility phenotype, and evolutionary conservation among 68 flagellated bacteria, revealed that core protein families and interactions are of high biological relevance. High confidence interactions were used to predict $\sim$18,000 motility interactions for 64 other flagellated bacteria. For identifying potential new motility candidates, conserved hypothetical proteins of each species were ranked based on interactions with known motility proteins, genes with motility phenotype, motility regulated gene expression, genomic context association and co-evolution among flagellated bacteria. To estimate how the four species and their conserved core network evolved, I conducted a phylogenetic analysis of 32 species based on 35 flagellar protein families.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The ability of bacteria to actively interact with their environment depends on membrane-embedded complexes, chemoreceptors, and flagellar-motors which are linked by a signal transduction pathway, the chemotaxis system (Figure 1.1). Thus, bacteria are able to direct their movement toward regions with higher concentrations of beneficial chemicals, mostly nutrients or lower concentrations of detrimental chemicals, i. e. toxins [1].

1.1 Patterns of bacterial motility

Escherichia coli and *Salmonella typhimurium* 'run' by rotating their helical flagella counterclockwise (CCW) which causes them to rotate in a bundle that propels the cell steadily forward. Switching motors to clockwise rotation (CW) disrupts this bundle, and causes the cell to 'tumble'. When motors are switched back to running mode, bacteria reorientate toward a new direction. In homogeneous environments, these bacteria 'run' and 'tumble' changing their direction once a second, which leads to random movement. In inhomogeneous environments, intervals are controlled by positive or negative stimuli, which produces directed movement (taxis) [3].

1.2 Chemotaxis

In *E. coli* stimulants are detected by transmembrane chemoreceptors, mostly methyl- accepting chemotaxis proteins (MCP) at the poles of the cell (Figure 1.1). The adapter protein CheW links the MCPs to the cytoplasmic histidine protein

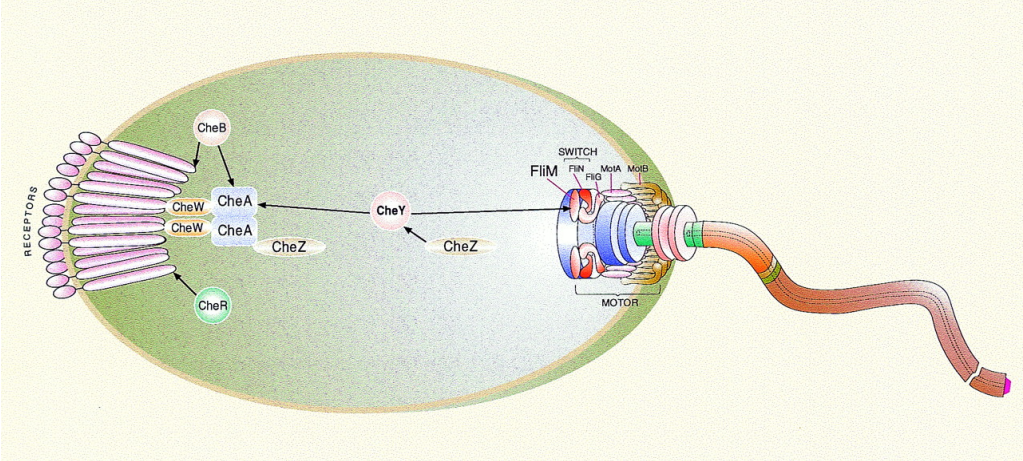

Figure 1.1 | Bacterial chemotaxis. A signal is detected by transmembrane receptors and transmitted by chemotaxis proteins to the flagellar-motor. Taken from Bren et al. [2].

kinases CheA. Two response regulators, CheY and CheB, compete for binding to CheA. A decrease of beneficial signals mediated by the MCPs stimulates autophosphorylation of CheA (CheA-P) and CheB (CheB-P). CheA-P transfers its phosphate group to CheY. Subsequently, CheY-P interacts via FliM, which is part of the flagellar motor, switching motor rotation from CCW to CW which causes 'tumbling' and a direction change. Pre-stimuli movement is restored by decreasing the concentration of CheY-P via CheZ and CheB-P. While CheZ directly dephosphorylates CheY-P, an increasing concentration of CheB-P indirectly reduces CheY's phosphorylation by an increased demythylation of the MCPs.

An increase of beneficial signals detected by the MCPs inhibits autophosphorylation of CheA, which in turn reduces the number of direction changes caused by CheY-P. Thus, the beneficial direction is retained longer. Along with decreased activity of CheB-P comes an increased activity of CheR, which increases methylation of the MCPs and restores pre-stimuli movement [2, 4].

1.3 Flagellar structure

The flagellum is one of the most complex molecular machines known. It comprises more than 50 distinct proteins (Figure 1.2) [7]. Several of its subunits are assembled from proteins in multiple copies from a few to tens of thousands.

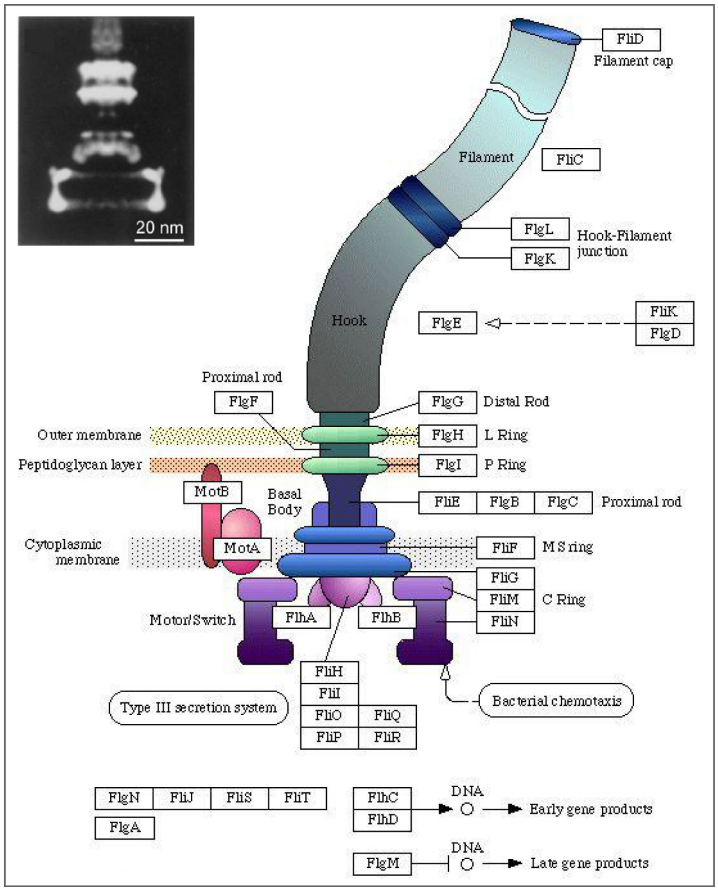

Figure 1.2 | Bacterial flagellum. Upper left panel depicts a *S. typhimurium* hook-basal body complex derived from electron microscopic images [5]. Bottom right panel represents a schematic drawing of a bacterial flagellum (taken from the KEGG database [6]; pathway K02412).

Three main components can be distinguished [8]:

- the basal body with components in the cytoplasm, across the cytoplasmic membrane, the periplasmic space, and the outer membrane
- a spiral protein filament in the external space
- an external flexible joint, the so-called hook, connecting the two components

Energy is provided by a cell-wall anchored proton channel consisting of multiple MotA and MotB proteins. At the same time these proteins constitute the stator

part of the motor. Torque is thought to be generated by inflowing protons which change the conformation of the cytoplasmic part of MotA. Energized MotA is thought to exert a force on a ring of FliG proteins associated with the rotational element of the motor.

FliG, together with FliM and FliN proteins make up the rotor basis, the cytoplasmic C-ring, also referred to as 'motor switch complex'. While FliN is thought to play a role in protein export, FliM is responsible for transmitting chemotaxis signals to the motor (Figure 1.2). The interaction between CheY-P and FliM results in a conformational change of FliM. This leads to a FliG triggered reversal of motor rotation [9]. The motor switch complex is attached to the MS ring consisting of multiple copies of FliF proteins anchored in the cytoplasmic membrane.

The rod or drive shaft is formed from FlgF, FliE, FlgB, FlgC and FlgG proteins (Figure 1.2). The drive shaft is guided through the outer layers of the cell wall by FlgH and FlgI (L and P rings). The flagellar hook, a short, highly curved cylindrical tube, functions as a joint between the basal body and the filament. Hook-associated proteins FlgK and FlgL act as a structural adapter between the flexible hook and the more rigid filament. Consisting of tens of thousands FliC proteins, the filament forms a long helical shaped structure. It therefore functions like a propeller, when rotated (Archimedes screw principle).

Recent systematic protein-protein interaction screenings have identified motility-related interactions in four pathogenic Gram-negative bacteria, the syphilis spirochete *Treponema pallidum* [10], *Campylobacter pylori* (personal communication with Finley RL Jr), the gastritis causing *Helicobacter pylori* [11] and the well-studied model bacteria *E. coli* [12]. Although elementary components of bacterial motility are well characterized, it is not clear if all components have been identified. Have all interactions among and between bacterial chemotaxis and flagellar proteins been found? Which of these interactions have been maintained throughout evolution? To answer these questions, I conducted the first comparative analysis of four systematic motility-centered interaction studies in bacteria.

1.4 Analyzing protein-protein interactions

Protein-protein interactions (PPIs) are essential for all cellular processes [13]. Being fundamental elements of cellular complexes and pathways, PPIs are key determinants of protein function. Thus, PPIs not only provide clues about new functional associations among cellular processes but most importantly about the function of hypothetical proteins in the context of their interacting partners. For instance, PPIs played a crucial role in the elucidation of the bacterial chemotaxis-signaling cascade (Figure 1.1) [2].

Although protein-protein interactions have been studied for decades, only recent advances have made them accessible to systematic computational analysis. First, more and more comprehensive interaction studies (interactomes) are conducted. Second, recently established databases of PPIs make interaction data easily accessible [14–18]. Third, increasing number of solved 3-dimensional structures of proteins and protein complexes enable us to study such assemblies in atomic detail [19]. However, in contrast to hundreds of completely sequenced genomes only a handful of comprehensive PPI studies have been carried out and there is no organism for which all PPIs are known. The field has exploded since the first interaction maps were published in 1997 for a subset of yeast proteins and in 2000 for a genome-wide dataset [20, 21]. With an increasing number of high-throughput experiments, more and more computational studies are carried out that analyze and compare these interaction datasets.

1.4.1 Experimental methods

Although there are many experimental methods for detecting PPIs, the bulk of data has been produced with just a handful of them. The two most popular are the yeast two-hybrid (Y2H [22–24]) and the complex purification method (CP [25]). Their popularity mostly stems from the fact that both can be carried out in a high-throughput fashion to produce large datasets of fairly consistent quality. Both experiments are conducted in an asymmetric way, i. e. the methods distinguish between bait and prey proteins. While baits are systematically screened against a whole or a subset of a proteome, preys are not. Successfully or positively tested baits are those which identified at least one prey.

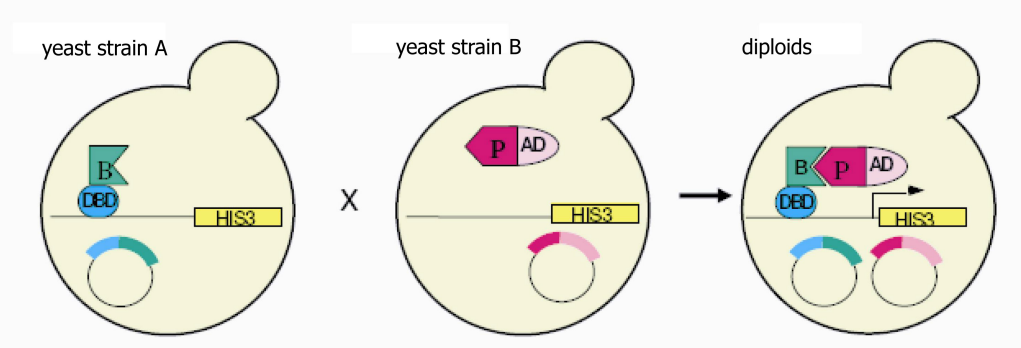

Figure 1.3 | Yeast two-hybrid principle. Protein B (bait) is expressed with a DNA-binding domain (DBD) in yeast (strain A). Protein P (prey) is fused and expressed with a transcriptional activation domain (AD) in yeast (strain B). A and B strains are mated to express the two fusion proteins in one diploid cell. If both fusions interact they reconstitute a transcription factor which activates a reporter gene (here HIS3) which in turn allows the cell to grow on selective media (here media lacking histidine). Taken from Rajagopala et al. [10].

Yeast two-hybrid

The Y2H method is a genetic screening system for PPI detection carried out in yeast, which was the first to be used for several large-scale studies (reviewed in [26]). It is based on the observation that protein domains, especially those of transcription factors, can be separated, recombined, and still retain their properties. It uses two fusion proteins ('hybrids') whose interactions reconstitutes a transcription factor which in turn activates one or more reporter genes or enzymes. Transcriptional activation can be detected (e. g. the activation of a HIS3 reporter gene allows a cell to grow in the absence of histidine) or measured quantitatively (Figure 1.3 [22–24]).

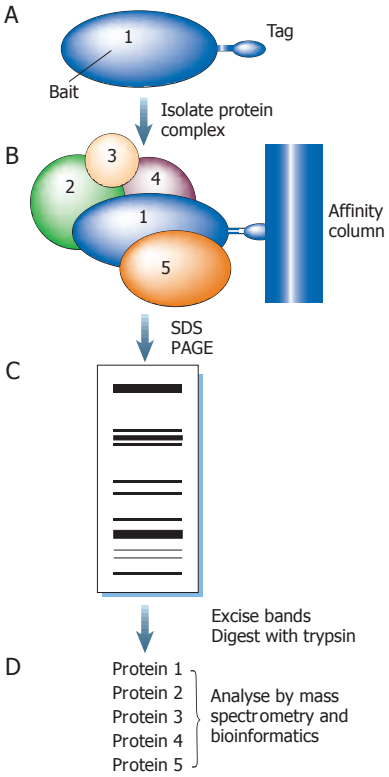

Figure 1.4 | Complex purification principle. (A) Target proteins (baits) are affinity tagged. (B) Complexes with associated proteins (preys) are systematically purified using an affinity column. (C) Purified proteins are separated according to their mass by one-dimensional 'sodium dodecyl sulfate' gel electrophoresis (SDS PAGE). (D) Trypsin-digested proteins are analyzed by mass spectrometry and identified by their unique mass spectra (modified after Kumar et al. [27]).

Complex purification

Protein complex purification in conjunction with mass spectrometry (MS) is the other major method for detecting PPIs (Figure 1.4). First, a piece of DNA, encoding a tagged protein (bait), is inserted into the target organism. After cells have expressed the bait fusion, cells are broken up. Via its tag, the bait is pulled out with all its attached proteins (preys) using techniques such as co-immunoprecipitation or tandem affinity purification (TAP). Finally, purified proteins are identified by mass spectrometry.

	Yeast two-hybrid	**Complex purification**
Advantages		
	- *in vivo* technique	- *in vivo* technique
	- detection of transient and unstable interactions	- detection of stable interactions
	- independent of endogenous protein expression	- reduction of steric interference (only one fusion protein)
	- fine resolution, enabling epitope mapping within proteins	- detection of interactions that depend on higher-order complexes
	- suitable for large-scale applications	- suitable for large-scale applications
Disadvantages		
	- does not identify cooperative binding	- binding relationships among purified proteins are unknown
false-positives	- proteins are artificially brought together in the nucleus although they might be differentially localized or expressed	- over-expression of bait proteins as well as unspecifically binding preys might lead to the detection of false-positives
false-negatives	- non-yeast proteins might not interact due to missing post-translational modifications	- low-abundance proteins might be missed
	- interactions of transcription factors cannot be detected (self-activators)	- weakly associated proteins might be washed off
	- sterical effects between proteins and their fused domains might prevent proteins from interacting	- tagging may disturb complex formation

Table 1.1 | Advantages and disadvantages of Y2H and CP

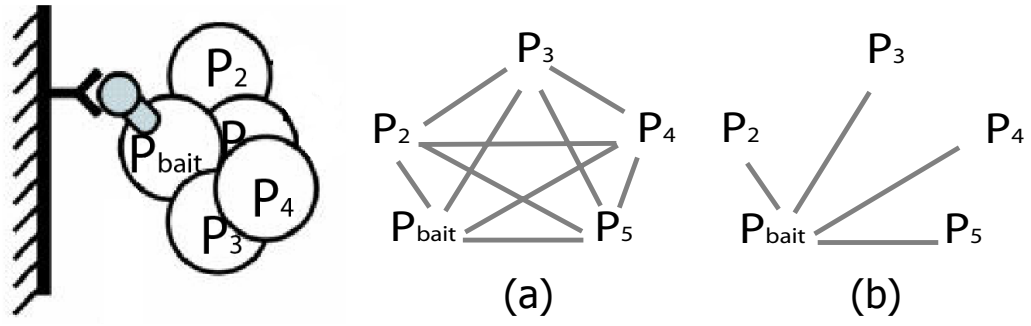

Figure 1.5 | Models to predict interactions from complex purification data. A purified complex may consist of 5 subunits ($P_{bait} - P_5$) whose precise topology is not known. (a) The matrix model predicts pairwise interactions among all subunits whereas the spoke model (b) predicts only interactions between the bait and its co-purified proteins (modified after [28, 29]).

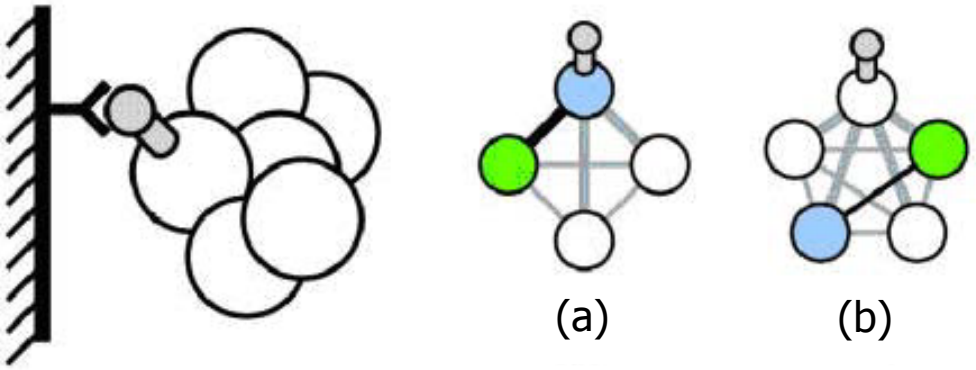

Figure 1.6 | Socio affinity index. To estimate if two proteins interact, Gavin et al. [30] have derived a formula to process raw purification data (see also section 2.1). In brief, their socio affinity index (SAI) quantifies the tendency for a protein pair (depicted in blue and green) to identify each other when tagged (a) and to co-purify when other proteins are tagged (b) relative to what would be expected from their frequency in the data set. High affinities are measured if both proteins purify each other when tagged (without purifying many other proteins) and if both are always seen together in purifications of other baits (modified after [29]).

Yeast two-hybrid versus complex purification

In contrast to Y2H, CP does not detect direct interactions (except in cases where only two proteins are co-purified). Instead, purified protein assemblies are held together by protein-protein interactions whose precise topology is usually not known (Figure 1.4 B). In order to predict direct interactions either the matrix or spoke model is applied (Figure 1.5). Bader et al. demonstrated that the spoke model is three times more accurate than the matrix model [31]. To quantify such interactions, Gavin et al. have introduced the socio-affinity index (SAI [30]) (see Figure 1.6 and Section 2.1). The outcome is equivalent to the matrix model (Figure 1.5 b) but with affinity-weighted interactions. Y2H and CP both suffer from false-negatives, i. e. protein-protein interactions that occur *in vivo* but could not be identified. For example, in both cases sterical effects between proteins and their fused domain might prevent interaction or complex formation. Both also have the reputation of generating false-positives, i. e. interactions which do not take place *in vivo* and thus have falsely been identified. Advantages and drawbacks of each method are summarized in Table 1.1.

In addition to a certain error margin, results are fairly reproducible. Only about half of all Y2H screens yield reproducible interactions [21]. Gavin et al. repeatedly pulled out 139 baits and their associated proteins. On average, 69% of purified proteins were common to both purifications.

Besides the experimental limitations of each method, Aloy et al. demonstrated that Y2H and CP tend to detect different kinds of interactions [32] and thus are highly complementary. While Y2H leads primarily to the identification of transient interactions, CP results more often in the discovery of stable interactions.

Y2H and CP studies have generated large PPI datasets (Table 1.2). Compared to eukaryotes, only a few systematic PPI studies have been carried out in bacteria [11, 12, 33].

1.4.2 Validation of protein-protein interactions

Experimental methods suffer from a certain number of false-positives and false-negatives. However, high-throughput methods are more prone to such artifacts as they generate them as systematically as they generate valid data. Several methods have been proposed to evaluate the quality of PPI data.

Organism	Purifications	Interactions	Method	Reference
Helicobacter pylori	-	1,465	Y2H	[11]
Escherichia coli	648	5,254	CP	[33]
	2,667	11,202	CP	[12]
Plasmodium falciparum	-	2,846	Y2H	[34]
Saccharomyces cerevisiae	-	1,511	Y2H	[21]
	-	4,549	Y2H	[35]
	589	3,757	CP	[36]
	741	2,583	CP	[37]
	1,993	21,107	CP	[30]
	2,357	NA	CP	[38]
Caenorhabditis elegans	-	4,624	Y2H	[39]
Drosophila melanogaster	-	20,405	Y2H	[40]
	-	2,300	Y2H	[41]
Homo sapiens	-	2,800	Y2H	[42]
	-	3,186	Y2H	[43]

Table 1.2 | High-throughput protein-protein interaction studies. Interactions given for complex purification studies are according to the spoke model.

Crystal structures

The best benchmarking data (gold standard) for evaluating protein-protein interactions are crystal structures of protein complexes. Unfortunately, there are not many such structures available. One of the most well studied crystal structure is the structure of the yeast RNA polymerase II. It consists of 10 subunits which are connected by 18 interactions [44]. While Y2H studies of a similar complex (RNA polymerase III and several associated proteins) found only 12 interactions among the 19 proteins [45], the crystal structure of RNA polymerase II shows a number of weak interactions where subunits barely touch each other. It is unlikely that such weak interactions will be detected by any method except by structure analysis.

Validation by network intersection

Edwards et al. [46] analyzed the overlap of various interaction datasets with interactions predicted from the MIPS complex catalog [47], a set of hand-curated

protein complexes. Based on the number of overlapping PPIs, these authors estimated the rate of false-negatives to be between 51% and 85% for various high-throughput Y2H datasets and to be 50% for CP studies. Von Mering et al. demonstrated that the number of false positives can be reduced by focusing on the intersection of PPIs generated by different kinds of experimental technologies, such as Y2H and CP [48]. In principle, PPIs detected by low-throughput experiments are considered to be more reliable than those identified by high-troughput techniques.

Interactions among paralogous proteins

According to the paralogous verification method (PVM) [49] proposed by Deane et al. an interaction is more reliable if the putatively interacting pair has paralogs that also interact. On a test set, this method identified correctly 40% of true interactions with an estimated false-positive rate of about 1% [49].

Genomic context

Several algorithms predict protein associations on the basis of sequence data from completely sequenced genomes and are inspired by comparative genomics techniques [29, 50]. The main methods are as follows:

Gene fusion: Functional associations predicted by the gene fusion or Rosetta Stone method, is based on the fact that multi-domain proteins found in one organism may be split in another. Therefore, it is likely that the fused domains interact within the multi-domain protein and thus the separate domains may interact as well [51].

Gene neighborhood: The gene neighborhood approach rests on the fact that many functionally related genes in bacteria are organized in operons, that is, they are gene neighbors. Furthermore, often proteins encoded in one operon are part of a complex, for example a multi protein enzyme complex. Neighboring genes in bacteria are in theory much more likely to interact than proteins encoded in other regions of the chromosome [52].

Gene co-occurence: The phylogenetic profile method predicts functional associations between genes according to the similarity of their co-occurrence patterns

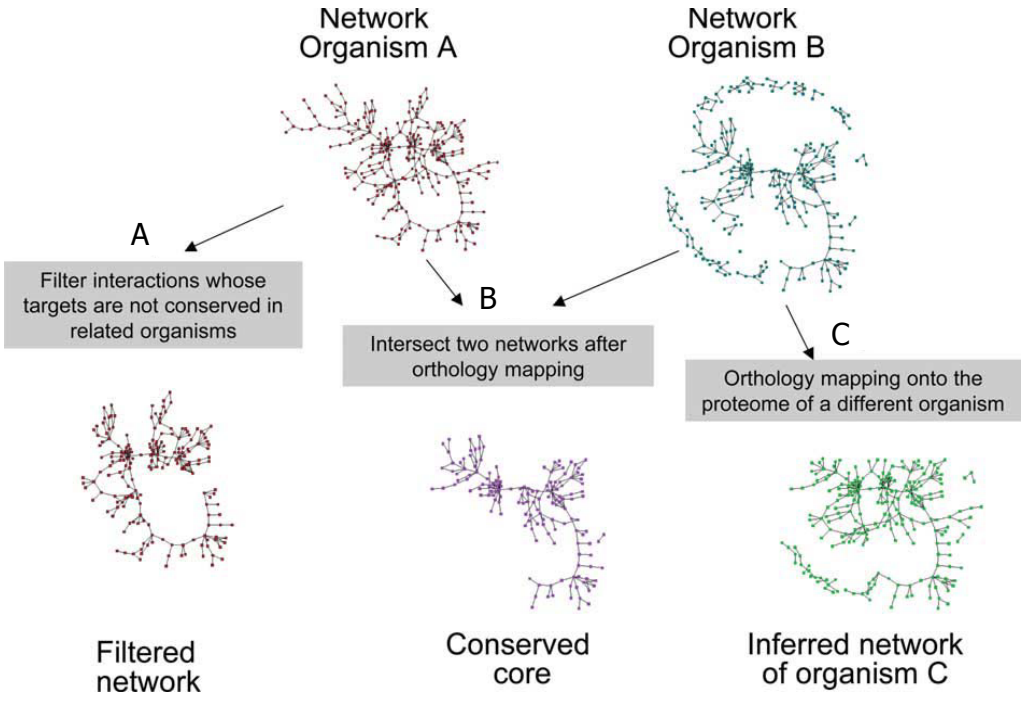

Figure 1.7 | Three interactomics approaches. Network information from two species may be used to: (A) extract PPIs of one species which are not conserved in the other or (B) to cross-validate experimental PPIs and to identify conserved modules (pathways or complexes) or (C) to predict PPIs *in silico*. Modified after Cesareni et al. [54].

(phylogenetic profiles) among orthologous genes in a set of reference genomes. Pellegrini et al. showed that proteins with a correlated evolution throughout many different genomes strongly tend to be functionally or physically linked [53].

1.4.3 Comparative interactomics

Proteins and their functions are usually well conserved throughout evolution. It is also known that PPIs are conserved. For example, in hemoglobins whose heterotetrameric structure is found in all vertebrates. Similar to the ongoing sequencing projects [56], comparative genomics techniques are used to exchange PPI information between organisms on the basis of homologous DNA or protein sequences [57]. In the context of PPIs the focus is not primarily on assigning a function to unknown proteins but rather to transfer the network information obtained experimentally to different organisms. This information can be used to

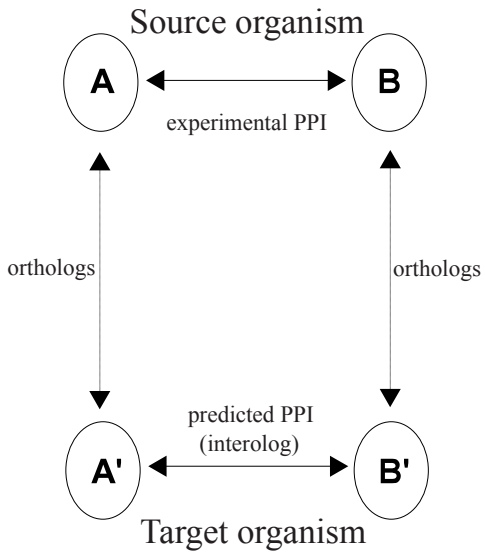

Figure 1.8 | Protein-protein interaction transfer via interologs. A-A' and B-B' are orthologs between the source and the target organism. Interolog mapping can be generalized when whole families of orthologous proteins are considered (as opposed to single orthologs). Modified after Matthews et al. [55].

cross-validate experimental PPIs, to predict PPIs *in silico* (which may be verified experimentally) [55, 58], or to isolate species specific PPIs (Figure 1.7). Most importantly, it can be used to detect conserved modules (pathways or complexes) within networks [54, 59–61] (Figure 1.7 B).

Orthologs vs. paralogs

Two types of evolutionary relationships among genes and proteins from different species can be distinguished: orthologs and paralogs. Homologous proteins in two species that have evolved from a common ancestral protein are called orthologs. Paralogous proteins are encoded by homologous genes that have diverged after gene duplication in the same species. Typically, orthologs occupy the same functional niche in different species, whereas paralogs tend to evolve toward functional diversification. Thus, like for gene function prediction [56], the identification of orthologous genes/proteins is crucial to transfer network information between organisms (Figure 1.8). Several attempts have been made to predict orthologs. Notably, Tatusev et al. used proteome-wide sequence similarity searches to extract reciprocal best hits which were supported by at least

three lineages to generate clusters of orthologs or orthologous groups of paralogs termed COGs [56, 62].

1.4.4 Graph theoretic aspects of protein-protein interaction networks

Protein interactions can be represented by graphs and be mathematically analyzed using graph theory[1]. Most PPI networks represent unweighted and undirected graphs, i. e. an unquantified mutual binding relationship between proteins (if protein A interacts with B than B interacts with A). In few cases, edges are weighted, e. g. when expression correlation of the respective genes is integrated, or directed to reflect asymmetrical experiments, e.g. to differentiate between bait and prey interactions.

A *graph* G is a set of *vertices* (or nodes or points, here: proteins) and *edges* (or lines, here: interactions) denoted by $G = (V, E)$, where the elements of V are vertices and the elements of E are edges. The usual way to picture a graph is by drawing a dot for each vertex and joining two of these dots by a line if the corresponding two vertices form an edge. It is the job of graph drawing algorithms to layout and display this information optimally. A *neighbor* of a vertex v is a node adjacent to v. The *neighborhood* is the set of neighbors of vertex v denoted by $N(v)$. The *degree* of a vertex v is the number of edges incident with v; this is equal to the number of neighbors of v. A *path* in a graph is a unique sequence of vertices and edges starting and ending with a node. The *path length* is the number of vertices in that sequence. The *distance* $d(u, v)$ in G between two vertices u, v, is the path length of the *shortest path* connecting u, v found in G. The *diameter* of G is the greatest distance, i. e. the longest shortest path, between any two vertices in G. The *clustering coefficient* of a vertex v is defined as

$$C(v) = \frac{2|E_v|}{k_v(k_v - 1)} \tag{1.1}$$

It describes the ratio between $|E_v|$, the number of edges between the neighbors of v and the largest possible number of edges between v and its neighbors. It is 1 if every neighbor connected to v is also connected to every other vertex within its neighborhood, and 0 if no vertex connected to v connects to any other vertex that is connected to v.

[1] http://mathworld.wolfram.com/topics/GraphTheory.html

The most elementary global networks features are:

- the *connectivity distribution* (or *node degree distribution*) $P(k)$ which reflects the probability of a node to have k neighbors.
- the *average shortest path length* (or average distance) defined by the average of the distances between any two vertices u and v, within the network.

$$\langle l \rangle = \frac{2}{n(n-1)} \sum_{u<v} d(u,v) \tag{1.2}$$

- the *average clustering coefficient* defined by the average of the clustering coefficients of all vertices within the network.

$$\langle C \rangle = \frac{1}{n} \sum_{i=1}^{n} C_i \tag{1.3}$$

1.5 Statement of objectives

- to analyze motility networks mathematically using graph theory.
- to evaluate motility interactions based on neighboring genes, gene-fusion events, gene co-occurrence, and other biological data.
- to transfer motility interactions between the four bacteria.
- to identify a conserved core of motility interactions.
- to integrate motility interactions with phenotypic, expression and other biological data.
- functional predictions for conserved proteins of yet unknown function based on their interaction with motility proteins.
- to analyze motility interactions in an evolutionary context.
- to predict motility interactions for other flagellated bacteria.

Chapter 2

Materials and methods

Genes and gene-related features of *T. pallidum*, *C. jejuni*, *H. pylori*, *E. coli*, and *Bacillus subtilis* including DNA and protein sequences were gathered from RefSeq [63] and KEGG [6]. Basic features were supplemented by predicted protein localizations (PSORTb 2.0 [64]), predicted domain-domain interactions derived from three-dimensional structures (3DID [65]) and by known and predicted PPIs (STRING 6.3 [29, 50]). Orthologous protein relationships were taken from the COG [56, 62] and the STRING database including non-supervised orthologous groups (NOGs [29]). For each organism, a set of conserved hypothetical proteins was identified by manually inspecting conserved proteins whose KEGG description contained the word 'hypothetical'[1]. A set of known motility proteins was compiled from KEGG using its pathway-based classification of orthologs (KO [66]). The following KO classes were selected:

- Bacterial chemotaxis (ko02030)
- Flagellar assembly (ko02040; Figure 1.2)
- Bacterial motility proteins (ko02035)

T. pallidum, *C. jejuni*, *H. pylori* and *E. coli* PPIs were filtered for motility interactions by retaining only PPIs which contain at least one known motility protein. In addition to motility interactions, experimental data comprises manually curated and extracted lists of interactions retrieved from PubMed, phenotypic data from *E. coli* [10], *B. subtilis* [67], *C. jejuni*, *H. pylori* [68, 69], and motility-related *E. coli* expression data [70]. For flexible and fast analysis, biological data was

[1] downloaded February 2006

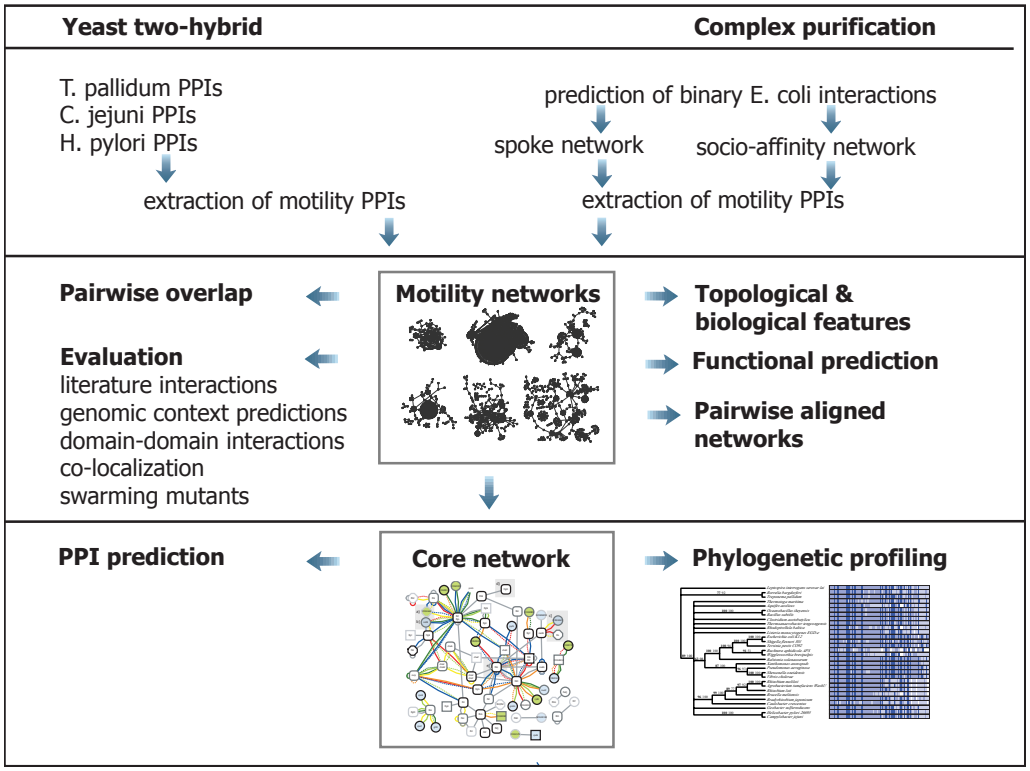

Figure 2.1 | Analysis pipeline

stored in a MySQL 5.0.18 relational database management system (RDBMS). STRING data was stored in a PostgreSQL 8.0.4 RDBMS. Data was retrieved via database connectivity interfaces and processed by Java programs developed using the integrated environment Eclipse. A graphical summary of the comparative analysis is given in Figure 2.1.

2.1 Socio affinity index

To predict binary interaction from *E. coli* raw purification data, the socio affinity index was applied as described in [30] (see also Figure 1.6):

$$A(i,j) = S_{i,j|i=bait} + S_{i,j|j=bait} + M_{i,j} \tag{2.1}$$

$$S_{i,j|i=bait} = \log\left(\frac{n_{i,j|i=bait}}{f_i^{bait} n_{bait} f_j^{prey} n_{i=bait}^{prey}}\right) \tag{2.2}$$

$$M_{i,j} = \log\left(\frac{n_{i,j}^{prey}}{f_i^{prey} f_j^{prey} \sum_{all-baits} n_{prey}\left(n_{prey}-1\right)/2}\right) \tag{2.3}$$

$n_{i,j|i=bait}$ is the number of times that protein i retrieves j when i is tagged; f_i^{bait} is the fraction of purifications where protein i was bait; f_j^{prey} is the fraction of all retrieved preys that were protein j; n_{bait} is the total number of all purifications, i. e. all successfully tested baits; $n_{i=bait}^{prey}$ is number of preys retrieved with protein i as bait; $n_{i,j}^{prey}$ is the number of times that proteins i and j are seen in purifications with baits other than i or j; n_{prey} is the number of preys observed with a particular bait (excluding itself).

2.2 Detection of homologous proteins

Sequence similarity analysis of *T. pallidum*, *C. jejuni*, *H. pylori* and *E. coli* proteins were performed using the blastall 2.2.8 program of the stand-alone local alignment search tool (BLAST) software [71].[2] Blastall searches were separately performed against each proteome using the subprogram blastp (default parameters). Results were extracted from BLAST XML outputs using a Java XML parser of the biojava package 1.4 and were stored in the MySQL database. BLAST *E*- values of two reciprocal hits were combined using the geometric mean.

2.3 Pairwise alignments of motility networks

Pairwise alignments of the PPI networks were performed using the Network Comparison Toolkit (NCT)[3], a Java implementation of the PathBLAST algorithm, as proposed by Kelley et al. [59, 72]. Briefly, the algorithm integrates PPIs from two species with protein sequence homology to generate an 'aligned network'. Proteins (one from each species) are merged into single nodes if their BLAST *E*-value is lower than a certain cut-off. The rule for creating an edge between two such nodes is that proteins of one species must directly be linked. In addition, proteins of the other species have to be in one of three states:

[2] ftp://ftp.ncbi.nih.gov/blast/

[3] http://chianti.ucsd.edu/nct/

1. the two proteins are the same protein
2. the two proteins are directly linked
3. the two proteins do not directly interact with each other, but interact with a common neighbor, also referred to as gap

Based on manual inspection of conserved motility interactions among orthologous proteins (supplementary Table A.4), a cut-off of BLAST *E*-value $\leq 10^{-5}$ was defined. Networks were generated by NCT and drawn using Cytoscape [73].

2.4 Construction of the core network

Nodes of the pairwise aligned networks were merged into nodes of orthologous proteins if both homologous proteins are members of the same orthologous group. The remaining nodes were discarded. Edges were transferred from the pairwise aligned networks if the connected nodes were both either part of the same or a different orthologous group. Based on manual inspection, nodes were labeled according to the common names of the merged proteins in combination with names of the motility COG set (supplementary Table A.1). In addition to motility interactions, I integrated small-scale interactions (literature set), phenotoypic data and FliC co-occurrence. Flagellin proteins (FliCs) are conserved throughout flagellated bacteria. FliC co-occurrence reflects a COG's conservation ratio among 68 species with FliC (COG1344 Flagellin and related hook-associated proteins) as reported by STRING [29]. The network was drawn using Cytoscape [73].

2.5 Flagellum supertree construction

FASTA-formatted sequences of proteins involved in the 'Flagellar assembly' pathway (ko02040; Figure 1.2) were downloaded from KEGG [6]. In total, this set comprises 48 families of orthologous proteins (pathway-based classification of orthologs [66]) conserved in up to 32 species (taxa).

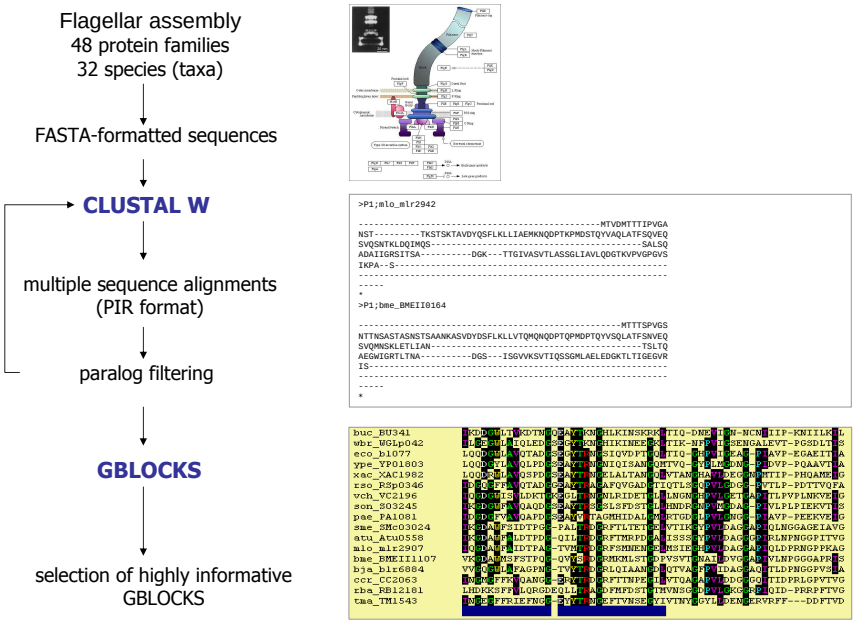

Figure 2.2 | Preprocessing of flagellar protein sequences

2.5.1 Multiple sequence alignments and processing

The FASTA formatted protein sequences were aligned using CLUSTAL W 1.83 [74] (default parameters). Multiple alignments were submitted to GBLOCKS [75] (default parameters). GBLOCKS searches for highly informative phylogenetic blocks. A block contains sites which are conserved in at least 50% of the taxa and is flanked by highly conserved sites (conserved in at least 85% of the taxa) (Figure 2.2). If a family contained recent paralogs (paralogs which are more similar to each other than to proteins of other species), one protein was randomly chosen and removed. If there were early paralogs (paralogs which are more similar to proteins from other species than to its own), only the most similar compared to the majority of proteins was retained. Protein families with less than 4 taxa or no conserved GBLOCKS sites were excluded from further analysis (in total 13 families). Subsequently, GBLOCKS of the remaining 35 protein families:

FlbD FlgA FlgB FlgC FlgD FlgE FlgF FlgG FlgH FlgI FlgK FlgL
FlgM FlgN FlhA FlhB FlhC FlhF FlhG FliA FliC FliD FliE FliF

FliG FliI FliK FliN/FliY FliP FliQ FliR FliS FliT motA motB

were used to construct the elementary protein family trees.

2.5.2 Phylogenetic analysis using maximum parsimony and neighbor-joining

Maximum parsimony (MP) is a character-based method that infers a phylogenetic tree by minimizing the total number of evolutionary steps (character changes) required to explain the observed sequence alignment. Neighbor-joining (NJ) infers a phylogenetic tree based on a distance matrix (converted from the observed sequence alignment) that represents the evolutionary distances between all pairs of species. Phylogenetic inference is a computationally demanding task. There are $2.9 \cdot 10^{40}$ possible unrooted trees for 32 taxa. Therefore, I have used a heuristic search which made computation feasible but does not guarantee to find the best solution. In both cases, statistical confidence estimates (bootstrap values) were calculated (standard bootstrapping procedure as proposed by Felsenstein [76]).

Construction of maximum parsimony consensus trees

The PIR formatted GBLOCKS were converted into the NEXUS format by the READSEQ program [4]. The NEXUS files were subjected to phylogenetic analysis using PAUP* win-4b10 [77]. For each family, a bootstrap analysis [76] with 100 bootstrap replicates was performed using a heuristic search based on the MP method. In total, 35 bootstrap consensus (50% majority-rule) trees were constructed. These trees were compiled into a single tree file and gene names were translated into species names.

Construction of neighbor-joining consensus trees

The PIR formatted GBLOCKS were converted into PHYLIP format by the READSEQ program. The PHYLIP files were bootstrapped with SEQBOOT [78] with 100 bootstrap replicates [76]. Maximum likelihood (ML) distance matrices were computed by TREE-PUZZLE 5.2 [79] using the Dayhoff amino acid substitution model incorporating among-site rate variation (gamma law based model, alpha parameter estimated by TREE-PUZZLE, eight gamma rate categories) in

[4] http://iubio.bio.indiana.edu/soft/molbio/readseq/java/

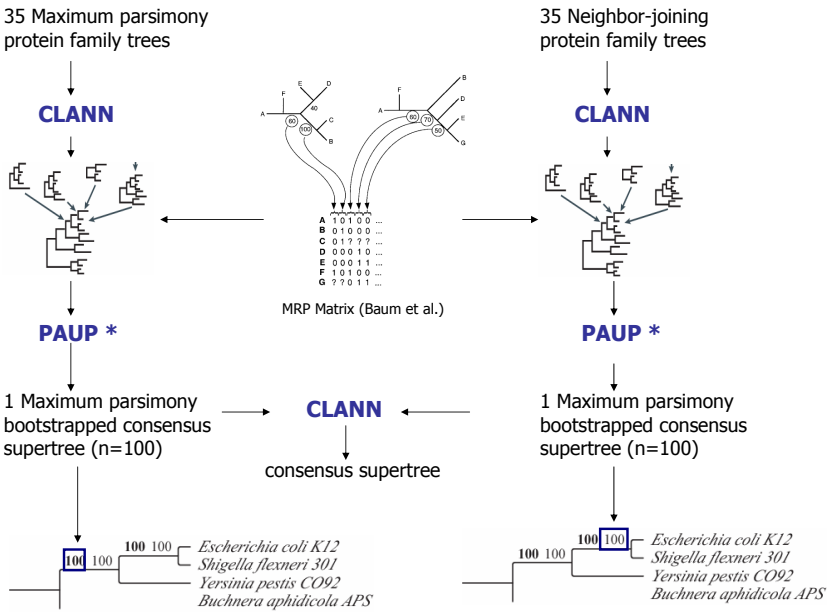

Figure 2.3 | Supertree construction

combination with PUZZLEBOOT 1.03[5]. Trees were generated from these ML distance matrices using NEIGHBOR [78] and summarized into consensus trees (50% majority-rule) using CONSENSE [78]. Consensus trees were compiled into a single tree file and protein names were translated into species names.

2.5.3 Supertree construction

Elementary protein family trees were merged into a single tree using the supertree approach [80,81]. Co-occurrence matrices of taxa among the MP and NJ trees were computed using the CLANN supertree software [82] indicating that *Streptomyces coelicolor* and *Chlamydia trachomatis serovar* had significantly lower co-occurrence values than the majority of taxa. Thus, those two taxa were removed from further analysis and the supertrees were constructed with the remaining 30 taxa. A matrix representation using parsimony (MRP) approach (Figure 2.3) [83] was used to represent the bootstrapped consensus trees as a single binary matrix (only branches with a bootstrap support higher than 50%

[5]distributed by A. J. Roger and M. E. Holder; http://members.tripod.de/korbi/puzzle/

were considered). The MRP matrices of the MP and NJ bootstrapped consensus trees were constructed with CLANN [82]. For each matrix a bootstrapped (100 bootstrap replicates) consensus tree (50% majority-rule) was generated by PAUP* [77] using a heuristic search based on the MP method. The resulting two trees were merged using CLANN [82] (50% majority-rule) and drawn using TreeGraph[6] [84].

2.6 Ranking of conserved hypothetical proteins

Conserved hypothetical proteins (CHPs) were extracted from each of the PPI sets and scored based on evidence supporting their role in motility.

Experimental evidence

Each interaction between a known motility protein and a CHP is evaluated based on

- its 2-way score i_1. It equals one, if the interaction has been reported in both directions, i. e. the bait protein interacted with the prey and vice versa.
- its pvm score i_2 according to the paralogous verification method (PVM) proposed by Dean et al. [49]. It is equal to the number of reproduced interactions among paralogs.
- its 3did score i_3 based on domain-domain interactions derived from three-dimensional structures (3DID [65]). It equals one if the interaction is supported by at least one predicted domain-domain interaction.
- its interolog score i_4. It equals one if the interactions is reported among orthologous proteins of at least another species.

The overall interaction score I is defined as

$$I = \sum_{k=1}^{4} i_k \tag{2.4}$$

If CHPs interact with more than one motility protein, the maximum interaction score is selected.

[6] http://www.nees.uni-bonn.de/downloads/TreeGraph/

Furthermore, a CHP is evaluated based on

- its eco mutant score m_1. It is one, if the CHP has a swarming mutant ortholog in *E. coli* [10].
- its bsu mutant score m_2 . It is one, if the CHP has a swarming mutant ortholog in *B. subtilis* [67].
- its cje mutant score m_3. It is one, if the CHP has a swarming mutant ortholog in *C. jejuni*.
- its hpy mutant score m_4. It is one, if the CHP has a swarming mutant ortholog in *H. pylori* [68, 69].
- its expression score x. It is one, if the CHP has an ortholog which has shown to be regulated by FhlD [70].

The overall mutant score M is defined as

$$M = \sum_{k=1}^{4} m_k \tag{2.5}$$

The overall experimental score E is defined as

$$E = I + M + x \tag{2.6}$$

Predicted motility links

COGs involved in motility (see supplementary A.1 on page 68) were collected. For each CHP, the top associated STRING [29] motility COG was extracted and its score was used to assign a string score S (predictions included genomic context, expression, literature mining and experimental evidence).

Associated orthologs

Orthologous CHPs found to be motility-associated in multiple species are more valuable than single CHPs. This evidence is scored by the motility association score A. It is equivalent to the number of species it has been found in.

FliC co-evolution

The ortholgous group COG1344 comprises flagellin and related hook-associated proteins (*FliC*). *FliC* proteins are conserved throughout flagellated bacteria. The *FliC* conservation score F of a CHP is defined as the conservation ratio of its orthologous group among 68 flagellated species as reported by STRING [29].

Combined Score

The combined CHP score C is defined as

$$C = E \cdot S \cdot F \cdot A \tag{2.7}$$

Chapter 3

Results

Recently, Rajagopala et al. tested known *T. pallidum* motility proteins as baits (fused to a Gal4-DNA binding domain) against a whole genome prey library (fusions with a Gal4 activation domain) using a systematic array-based Y2H approach [10]. This PPI set will be termed TPA in the remainder of this thesis. Similarly, known motility proteins (fused to a *lexA* DNA-binding domain) were systematically tested for their protein-protein interactions (preys were fused with a *B42* protein) in *C. jejuni* (personal communication with Finley RL Jr). This set will henceforth be referred to as CJE ALL. Both Y2H screenings used an array-based approach which is known to reduce the number of potential false-positives by allowing for stringent background control, assessment of reproducibility, and filtering of unspecifically interacting prey proteins [26]. In addition, CJE ALL interactions were assigned confidence scores using a logistic regression procedure incorporating several parameters relevant for the system (personal communication with Finley RL Jr). Based on these scores, a high confidence set (termed CJE HCF) was compiled. These three sets were complemented by PPIs identified by a partial Y2H screening in *H. pylori* [11]. As this Y2H study did not focus on motility proteins a subset of *H. pylori* motility interactions (HPY) was extracted.

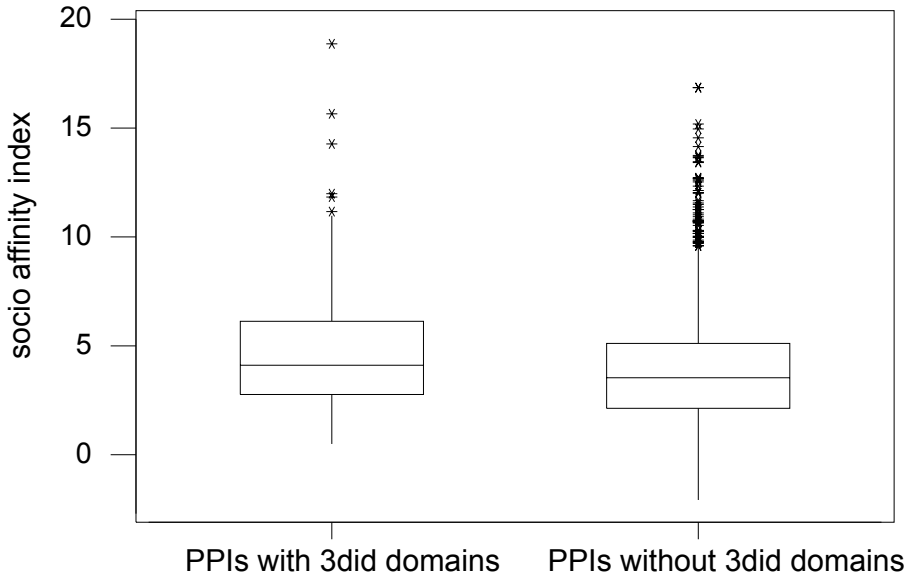

Figure 3.1 | Boxplots of socio affinities of PPIs with and without 3DID evidence

3.1 Interactions predicted from complex purifications

Arifuzzaman et al. conducted a comprehensive complex purification study in *E. coli*. Using a His-tagged *E. coli* ORF clone library (4,339 proteins), they were able to purify 2,667 proteins successfully and identified the co-purified proteins by MS [12]. Other than Y2H, CP does not directly yield PPI data, but protein complexes (baits and their co-purified proteins Figure 1.4 B). Usually, PPIs are predicted by applying either the spoke or the matrix model (Figure 1.5). Arifuzzaman et al. provided their results according to the spoke model. From this genome-wide set, a motility-centered subset (ECO SPK) was extracted. The spoke model may miss potential true interactions (true-positives) among preys whereas the matrix model contains all true interactions but unavoidably predicts false interactions (false-positives). Hence, I applied the socio affinity (SAI) method invented by Gavin et al. [30]. Similar to the matrix model it predicts PPIs among all proteins. The difference is that PPIs are weighted according to the pair's propensity

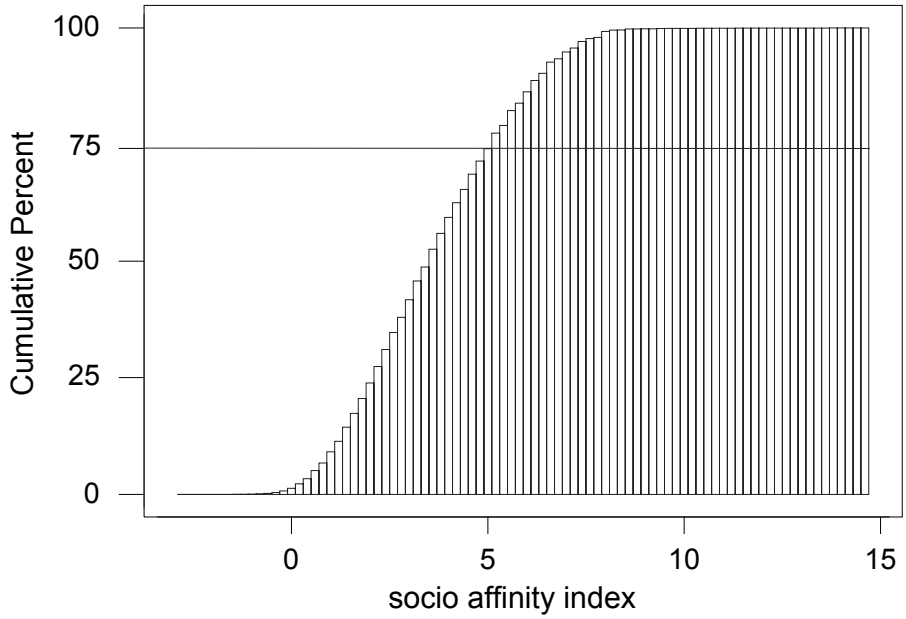

Figure 3.2 | Cumulative percentage distribution of socio affinity indexes

to associate which each other relative to what would be expected from their frequency in the data set (see Section 2.1 and Figure 1.6). I compared the affinity indexes of protein-protein interactions mediated by domain-domain interactions derived from three-dimensional structures (3DID [65]) with indexes of interactions without 3DID evidence (Figure 3.1). A non-parametric one sided two-sample rank (Mann-Whitney) test of the two population medians was performed.

$$\mathrm{H}_0 : \eta_1 = \eta_2$$

$$\mathrm{H}_1 : \eta_1 > \eta_2$$

Equality of population medians H_0 could be rejected with $\mathrm{p} < 10^{-4}$ in favor of the alternative hypothesis H_1 that the median of socio affinities of PPIs with 3DID evidence (η_1) is greater than those without (η_2). The test underscores the biological relevance of the socio affinity approach in the context of *E. coli* complex purifications. Based on the cumulative percentage distribution of socio affinities, I defined the top 25% of PPIs to be highly associated (Figure 3.2). Interactions with affinities > 5 were selected and a motility subset (ECO SAI) was extracted.

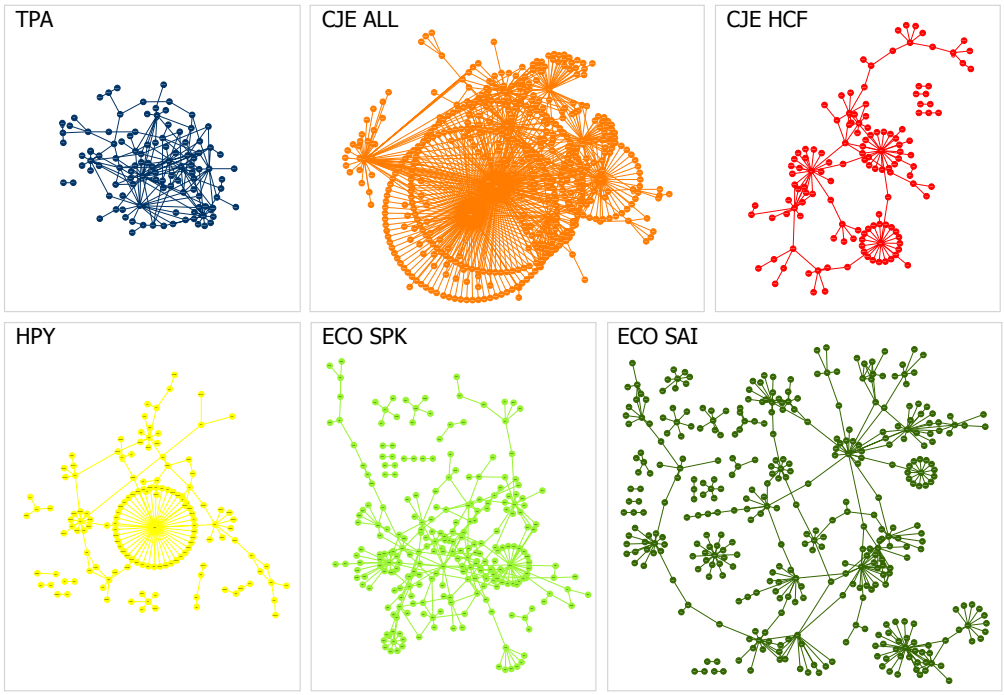

Figure 3.3 | Bird's eye view of motility networks. While the motility network of *T. pallidum* (TPA) looks tightly clustered, the comprehensive network of *C. jejuni* (CJE ALL) seems to contain more highly connected (unspecific) interactions than its high-confidence subset (CJE HCF). Being wide spread and less interconnected, *E. coli*'s ECO SAI appears to be the network with the greatest diameter. Networks were drawn with Cytoscape [73].

3.2 Topological features of motility networks

Motility networks vary considerably in their size, structure and protein composition (Figure 3.3 and Table 3.1). The number of distinct proteins ranges from 110 in TPA to 525 in CJE ALL. While the Y2H studies identified 176 PPIs in *T. pallidum* (TPA) and 140 high confidence interactions in *C. jejuni* (CJE HCF), a similar number of 139 motility interactions has been identified in *H. pylori* (HPY). More interactions have been found in CJE ALL and in *E. coli* (ECO SPK and ECO SAI). On average a protein was connected with two to three other proteins (Table 3.1).

Feature	TPA	CJE ALL	CJE HCF	HPY	ECO SPK	ECO SAI
Topological features						
Nodes	110	525	133	141	257	374
Edges	176	690	140	139	289	407
Avg. degree	3.182	2.621	2.09	1.965	2.249	2.177
Degree exponent	1.291	1.031	1.202	1.224	1.516	1.514
R-Sq	0.79	0.728	0.645	0.731	0.835	0.8306
Diameter	8	9	14	13	16	18
Avg. clustering coefficient	0.008	0.047	0.042	0	0.002	0.091
Avg. shortest path	3.7	3.591	5.121	4.357	4.907	6.749
Biological features						
Inter-motility PPIs	32	19	12	10	5	10
	18%	3%	9%	7%	2%	2%
Percentage of known motility proteins	69%	76%	63%	69%	72%	71%
Motility proteins	34	35	29	31	49	48
	31%	7%	22%	22%	19%	13%
Non-motility proteins	33	278	53	53	153	230
	30%	53%	40%	38%	60%	61%
Conserved hypotheticals	31	174	37	41	55	94
	28%	33%	28%	29%	21%	25%
Hypotheticals	12	38	14	16	0	2
	11%	7%	11%	11%	0%	1%

Table 3.1 | Topological and biological features of motility networks

TPA		**CJE ALL**		**CJE HCF**	
Protein	Degree	Protein	Degree	Protein	Degree
flaB3	20	fliM	160	fliM	25
fliY	19	flgG2	103	flgG2	22
flgG-2	19	fliY	58	fliL	21
HPY		**ECO SPK**		**ECO SAI**	
flgB	47	cheW	28	fliC	30
fliS	14	cheA	23	tsr	24
flgH	10	cheZ	17	cheZ	23

Table 3.2 | Top three highly connected proteins

Degree distribution analysis (Figure 3.4) indicated that the motility centered networks are not scale-free with $P(k) \not\sim k^{-y}$, i.e their degree distributions $P(k)$, which reflect the probability of a node to have k neighbors, could not well be approximated by a power law relationship (R-Sq 0.65 - 0.83). Nevertheless, degree distributions indicate few highly connected proteins. For example, FliM and FligG2 interacted with more than 100 proteins in CJE ALL (see circles in Figure 3.3). In ECO SPK the three most highly connected proteins are the chemotaxis proteins CheA, CheW and CheZ (Table 3.2). The network diameter, i. e. the longest shortest path between any two proteins reveals that TPA and CJE ALL are the most compact networks (8–9 proteins) while the ECO sets are the most wide spread (16–18 proteins). This is partially confirmed by the average shortest path length, which measures the average distance between any two proteins. The average clustering coefficient characterizes the overall tendency of nodes to form clusters with their neighbors (see Section 1.4.4). On average, ECO SAI has with 0.091 the highest clustering coefficient. This is not surprising given that edges among its nodes have been predicted by the matrix model which by definition links all co-purifying neighbors (without socio affinity filtering the clustering coefficient would be 1). In contrast, ECO SPK has a very low clustering coefficient (0.002) as its underlying model by definition predicts no links among prey proteins. Among the Y2H sets CJE ALL and CJE HCF showed the highest clustering tendency (0.047 and 0.042 respectively) while HPY did not show any clustering at all. Overall, the low clustering coefficient is probably an artifact of the filtering procedure used.

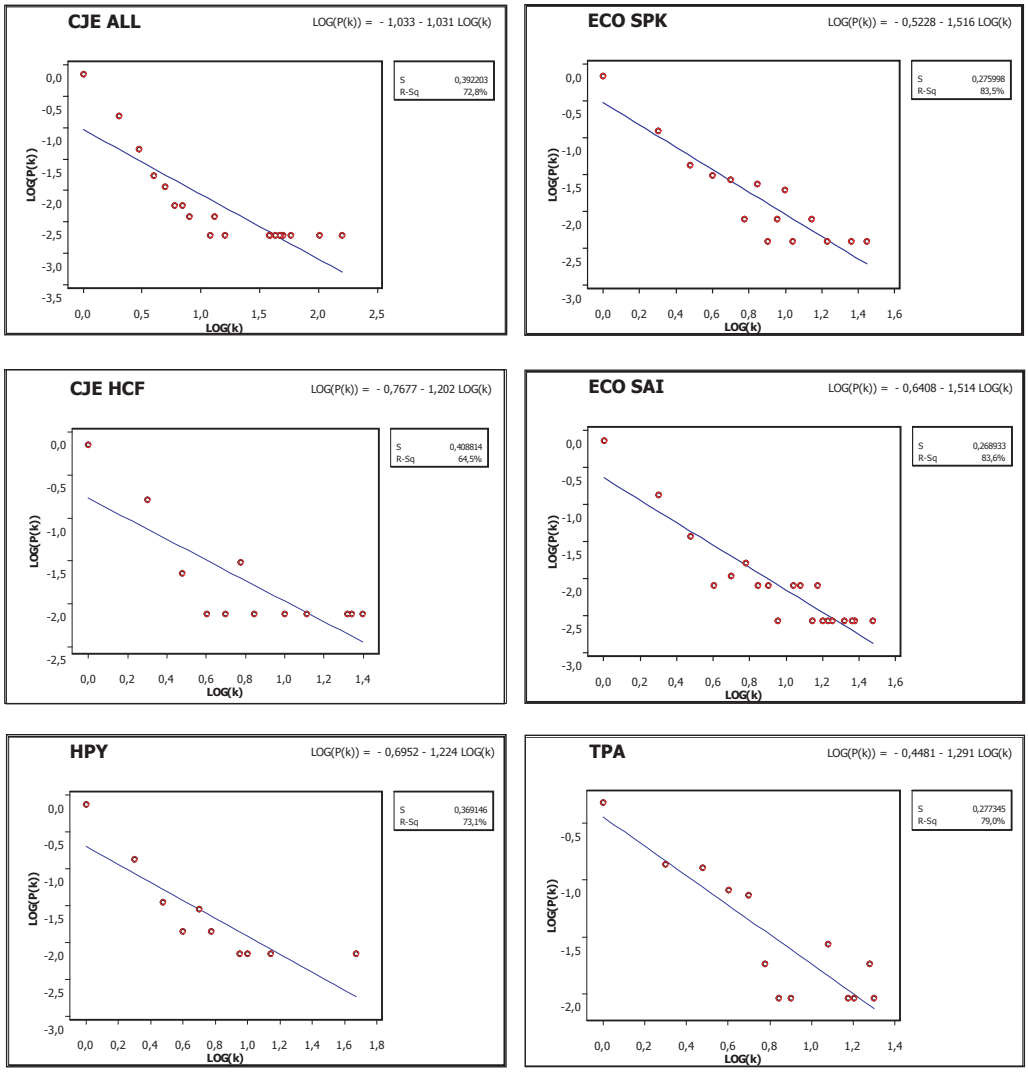

Figure 3.4 | Node degree distribution analysis of motility networks

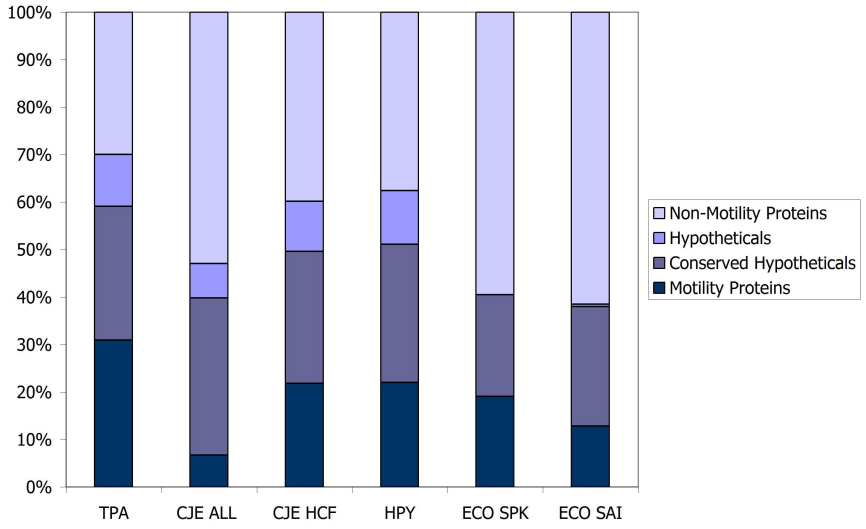

Figure 3.5 | Proportion of protein classes among motility interactions

3.3 Biological features of motility networks

It is well known that chemotaxis signals and proteins of the flagellum apparatus are transmitted/assembled via protein-protein interactions [2]. Notably, 32 inter-motility PPIs were found in TPA, supporting its higher quality compared to the others (Table 3.1). While ECO SAI predicted 10, ECO SPK only reported 5 interactions among motility proteins. Although the Y2H study in *H. pylori* was not comprehensive and not centered around motility it identified interactions linking 69% of its known motility proteins (Table 3.1). A similar fraction was identified by the others suggesting a good overall coverage. The remaining 30% have either not been shown to interact (including potential false-positives) or have not been tested (will be discussed in the next Section). In addition, other proteins (either non-motility proteins or proteins with unknown function) were identified to be directly linked with motility (henceforth referred to as associated proteins). Proteins with unknown functions are either conserved in other species (conserved hypotheticals) or or species-specific (hypotheticals). An overview is given in Figure 3.5. The percentage of proteins of other functional classes varied between 30% in TPA and 60% in the ECO sets. More importantly, on average 27% were conserved hypotheticals—potential new motility candidates. While the Y2H sets also comprised around 10% of hypothetical proteins the CP sets contained none or just a few species specific hypothetical proteins which is ex-

Set	Functional class	Percentage
TPA	Translation, ribosomal structure and biogenesis	15%
	Cell wall/membrane/envelope biogenesis	13%
	Function unknown	12%
CJE ALL	General function prediction only	14%
	Amino acid transport and metabolism	10%
	Cell wall/membrane/envelope biogenesis	9%
CJE HCF	Energy production and conversion	13%
	General function prediction only	12%
	Function unknown	12%
HPY	Replication, recombination and repair	18%
	General function prediction only	11%
	Posttranslational modification, protein turnover, chaperones	8%
ECO SPK	Translation, ribosomal structure and biogenesis	11%
	Transcription	10%
	Energy production and conversion	9%
ECO SAI	Transcription	12%
	General function prediction only	9%
	Energy production and conversion	8%

Table 3.3 | Top three functional classes among associated proteins

pected by a proteome-wide fraction of around 1% of hypothetical *E. coli* proteins.

Associated proteins were classified according to 25 functional classes defined by the COG database [62]. Strikingly, a strong link between motility and 'Energy production and conversion' was found in both *E. coli* sets (8%–9% of all classified associated proteins) and in CJE HCF (13%) (Table 3.3). Except for ECO SPK, associated proteins with 'General function prediction only' and 'Function unknown' were among the most frequent classes. Numerous 'Cell wall/membrane/envelope biogenesis', 'Replication, recombination and repair', and 'Translation, ribosomal structure and biogenesis' proteins were identified as well indicating that motility proteins are embedded in a broader functional context (Figure 3.6). Figure 3.7 depicts functional class compositions restricted to associated proteins which are conserved in a certain number of species. For example, the third bar represents functional classes of conserved proteins found to be associated in three species. Interestingly, two protein families were asscoci-

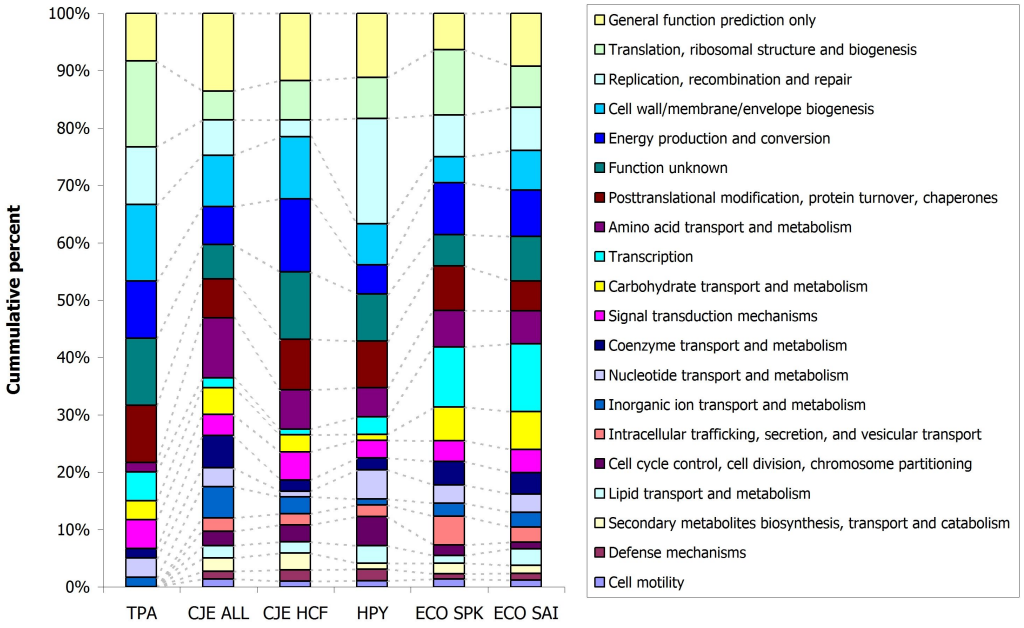

Figure 3.6 | Functional classification of associated proteins found in motility networks

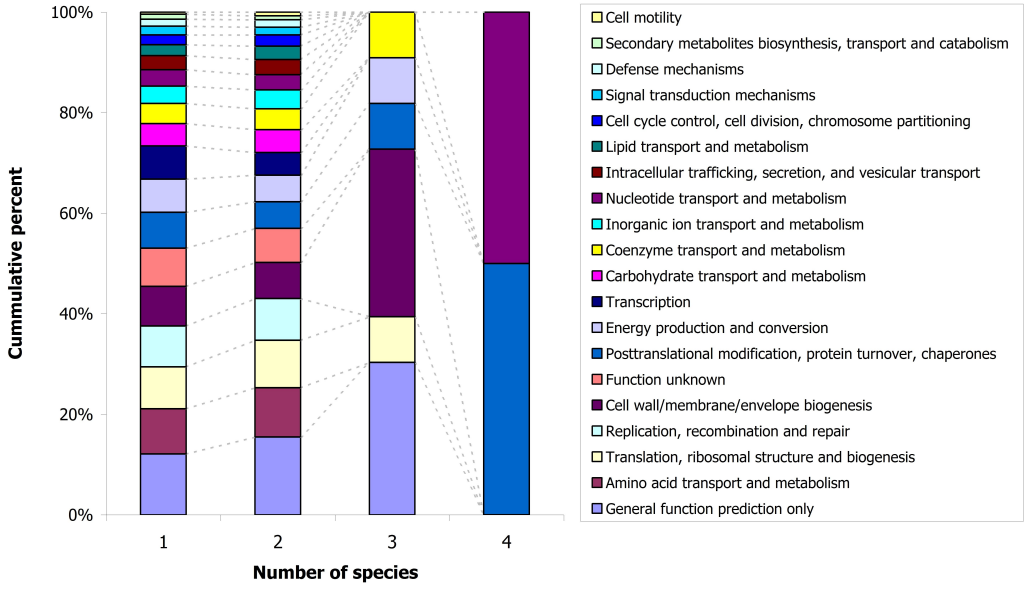

Figure 3.7 | Functional classification of associated proteins which are conserved in a certain number of species

	T. pallidum	*C. jejuni*	*H. pylori*	*E. coli*
Motility proteins	49	46	45	68
Positively tested	29	18	10	34
% positively tested	59%	39%	22%	50%

Table 3.4 | Fraction of positively tested motility proteins. Motility proteins as defined by the KEGG database [6].

number of organisms	number of families	percentage of families
1	9	29%
2	11	35%
3	10	32%
4	1	3%

Table 3.5 | Bait overlap. Overlap between positively tested protein families which are conserved in all four organisms.

ated and conserved in four species. One family belongs to 'Nucleotide transport while the other is involved in 'Posttranslational modification, protein turnover, chaperones'. While the fomer seems to influence bacterial metabolism on the DNA level, the latter has an metabolic effect on the proteome level. Thus, bacterial motility appears to be interweaved with basic metabolic processes.

3.4 How comprehensive are these studies?

Between 63% and 76% of known motility proteins have been identified by the Y2H and CP studies either as bait or as prey (Figure 3.1). Bait proteins are of special interest since those proteins were systematically screened against the whole proteome or a subset of proteins. The more baits are tested successfully the more comprehensive a study gets. When looking at the protein level, the fraction of positively tested known motility proteins varies between 22% for *H. pylori* to 59% for *T. pallidum* (Table 3.4). This shows that a high fraction of baits either could not detect any binding partner (as in *T. pallidum*, *C. jejuni* and *E. coli*) or has not been tested at all (as in *H. pylori*).

On the protein family level, 52 out of 80 (65%) known motility orthologous groups (supplementary Table A.1) contained proteins positively tested for at least one organism. Altough 31 of those families (60%) were conserved in all four or-

Pairwise Comparison		Set 1			Set 2			Similarity		
Set 1	**Set 2**	$\mathbf{a}_1$	$\mathbf{b}_1$	$\mathbf{c}_1$	$\mathbf{a}_2$	$\mathbf{b}_2$	$\mathbf{c}_2$	$\mathbf{c}_1 + \mathbf{c}_2$	$\mathbf{b}_1 + \mathbf{b}_2$	$\mathbf{d}_{1,2}$.
TPA	CJE ALL	17	59	4	21	100	4	8	159	5.0%
TPA	CJE HCF	15	50	3	15	22	3	6	72	8.3%
TPA	HPY	10	28	6	9	21	3	9	49	18.4%
TPA	ECO SPK	17	64	2	17	47	1	3	111	2.7%
CJE ALL	HPY	14	22	4	12	18	2	6	40	15.0%
CJE ALL	ECO SPK	20	199	4	19	54	4	8	253	3.2%
CJE HCF	HPY	7	9	4	10	15	2	6	24	25.0%
CJE HCF	ECO SPK	11	40	1	11	43	1	2	83	2.4%
HPY	ECO SPK	11	57	1	13	61	2	3	118	2.5%

Table 3.6 | Pairwise similarities based on PPIs indentified by conserved baits. $\mathbf{a}_1$ = Set 1 baits which have an orthologous bait in Species 2. $\mathbf{b}_1 = \mathbf{a}_1$ interactions of which the prey has an ortholog in Species 2. $\mathbf{c}_1$ = the subset of PPIs of $\mathbf{b}_1$ that are conserved with PPIs in Set 2. $\mathbf{a}_2$ = Set 2 baits which have an orthologous bait in Species 1. $\mathbf{b}_2 = \mathbf{a}_2$ interactions of which the prey has an ortholog in Set 1. $\mathbf{c}_2$ = the subset of PPIs of $\mathbf{b}_2$ that are conserved with PPIs in Species 1. Pairwise similarity $\mathbf{d}_{1,2} = (c_1 + c_2)/(b_1 + b_2)$.

ganisms, only 1 was positively tested for all organisms (COG1344 FLGL/FLIC) (see Table 3.5 and supplementary Table A.2). This implies that not only a partial fraction of motility proteins were tested successfully but also that the majority of these proteins belong to different protein families (65% were tested for one or two organisms). Overall, this supports an integrative approach to reduce the number of potential false negatives.

3.5 How similar are these studies?

Subsets of protein-protein interactions identified by orthologous baits were pairwise compared using the interologs approach (Figure 1.8). For example, in TPA 10 baits were screened which have an orthologous bait in HPY (a_1) (Table 3.6). Vice versa, in HPY 9 baits were screened which have an orthologous bait in TPA (a_2). Due to paralogs, a_1 and a_2 might differ. Based on interactions identified by conserved HPY baits (a_2), 28 interologs (b_1) were predicted for TPA. Conversely, 21 interologs were identified for HPY (b_2). While 6 out of 28 predicted TPA interactions were confirmed experimentally (c_1), 3 confirmed interologs were iden-

Pairwise Comparison		Set 1			Set 2					Similarity
Set 1	**Set 2**	$\mathbf{a}_1$	$\mathbf{b}_1$	$\mathbf{c}_1$	$\mathbf{a}_2$	$\mathbf{b}_2$	$\mathbf{c}_2$	$\mathbf{c}_1 + \mathbf{c}_2$	$\mathbf{b}_1 + \mathbf{b}_2$	$\mathbf{d}_{1,2}$
TPA	CJE ALL	176	90	4	690	226	4	8	316	2.5%
TPA	CJE HCF	176	90	3	140	51	3	6	141	4.3%
TPA	HPY	176	81	8	139	61	6	14	142	9.9%
TPA	ECO SPK	176	99	2	289	95	1	3	194	1.5%
TPA	ECO SAI	176	99	5	407	66	4	9	165	5.5%
CJE ALL	HPY	690	462	4	139	100	3	7	562	1.2%
CJE ALL	ECO SPK	690	459	5	289	129	5	10	588	1.7%
CJE ALL	ECO SAI	690	459	2	407	126	2	4	585	0.7%
CJE HCF	HPY	140	79	4	139	100	3	7	179	3.9%
CJE HCF	ECO SPK	140	92	1	289	129	1	2	221	0.9%
CJE HCF	ECO SAI	140	92	0	407	126	0	0	218	0.0%
HPY	ECO SPK	139	86	3	289	118	4	7	204	3.4%
HPY	ECO SAI	139	86	3	407	111	3	6	197	3.0%

Table 3.7 | Pairwise similarities based on orthology. $\mathbf{a}_1$ = number of PPIs in Set 1. $\mathbf{b}_1$ = Set 1 PPIs whose proteins have orthologs in Species 2. $\mathbf{c}_1$ = the subset of PPIs of $\mathbf{b}_1$ that are conserved with PPIs in Set 2. $\mathbf{a}_2$ = number of PPIs in set 2. $\mathbf{b}_2$ = Set 2 PPIs whose proteins have orthologs in Species 1. $\mathbf{c}_2$ = the subset of PPIs of $\mathbf{b}_2$ that are conserved with PPIs in Set 1. $\mathbf{d}_{1,2} = (c_1 + c_2)/(b_1 + b_2)$.

tified for HPY (c_2). Pairwise similarity of TPA/HPY was then defined as the sum of confirmed interologs divided by the sum of interologs:

$$d_{1,2} = (c_1 + c_2)/(b_1 + b_2) = \frac{9}{49} = 18.4\%$$

Pairwise similarities ranged from 2.4% for HPY/ECO SPK to 25.0% for CJE HCF/HPY. TPA and both CJE sets show the best overlap with HPY (Table 3.6). Principally, CJE HCF retrieved higher similiarities than CJE ALL supporting its higher quality. Overall, ECO SPK obtained the weakest pairwise similarities. If one takes evolutionary variation into account the comparatively high similarity between CJE HCF and HPY is not surprising given the fact that the two ϵ proteobacteria are closly related. Although 10 protein families were positively tested for three organisms (Table 3.5), only one interaction (FliS-FliC) was conserved in three sets. None could be found to be conserved in all. In particular, baits tested by CP and Y2H have identified vastly different kinds of interactions. This discrepancy might be due to their tendency to identify different interactions [32]

as well as due to the limitation of the spoke model.

When comparing all interactions (including ECO SAI), similarities decreased more or less two-fold (Table 3.7). This is expected and reflects the asymmetrical approach of the experimental methods, i. e. only baits were systematically tested against the proteome. When comparing Table 3.7 with Table 3.6 one can see that mostly all conserved interactions were identified among common baits. An overview of all conserved interactions is given in Table 3.8.

COG A	COG B	TPA	CJE ALL	CJE HCF	HPY	ECO SPK	ECO SAI
COG0008	COG1191	-	-	-	gltX-fliA	-	gltX-fliA
COG0085	COG1191	-	-	-	rpoBC-fliA	rpoB-fliA	-
COG0086	COG1191	-	-	-		rpoC-fliA	-
COG0090	COG1868	-	rplB-fliM	-	-	rplB-fliM	-
COG0208	COG1344	nrdB-flaB3	-	-	-	-	nrdF-flgL
COG0442	COG1344	proS-flaB2	-	-	-	proS-fliC	proS-fliC
COG0442	COG1344	proS-flgL	-	-	-		-
COG0459	COG1317	-	groEL-fliH	-	-	mopA-fliH	-
COG0526	COG1843	TP0100-flgD	Cj0864-flgD	-	-	-	-
COG0582	COG4786	-	xerD-flgG2	-	-	intC-flgG	intC-flgG
COG0674	COG1815	TP0939-flgB	-	-	HP0589-flgB	-	-
COG0835	COG0840	-	-	-	cheW-tlpA	cheW-tsr	cheW-tsr
COG0835	COG0840	-	-	-		cheW-tar	
COG0840	COG1344	mcp2-3-flaB3	-	-	cag26-flaA	-	aer-fliC
COG0852	COG1868	-	nuoC-fliM	-	-	nuoC-fliM	nuoC-fliM
COG1344	COG1516	flaB1-fliS	flaA-fliS	flaA-fliS	flaA-fliS	-	-
COG1344	COG1516	flaB2-fliS	flaB-fliS	flaB-fliS	flaB-fliS	-	-
COG1344	COG1516	flaB3-fliS	flaC-fliS	flaC-fliS		-	-
COG1344	COG1699	flaB1-TP0658	-	-	flaA-HP1377	-	-
COG1344	COG1699	flaB2-TP0658	-	-	flaA-HP1154	-	-
COG1344	COG1699	flaB3-TP0658	-	-		-	-
COG1344	COG2199	flaB3-TP0981	-	-	-	-	fliC-b1490
COG1360	COG1463	-	motB-Cj1648	motB-Cj1648	motB-HP1464	-	-
COG1580	COG1826	-	fliL-Cj0579c	fliL-Cj0579c	-	fliL-ybeC	-

Table 3.8 | Conserved interactions

Set	$\mathbf{a}_1$	$\mathbf{a}_2$	$\mathbf{b}_1$	$\mathbf{b}_2$	$\mathbf{c}_1$	$\mathbf{c}_2$
TPA	39	35	8	11	20.5%	22.9%
CJE ALL	38	22	5	8	13.2%	22.7%
CJE HCF	38	22	4	7	10.5%	18.2%
HPY	38	14	4	5	10.5%	28.6%
ECO SPK	37	33	3	3	8.1%	9.1%
ECO SAI	37	33	3	3	8.1%	9.1%

Table 3.9 | Confirmed literature interactions. $\mathbf{a}_1$ = number of predicted literature interologs (i-COGs). $\mathbf{a}_2$ = number of predicted literature interologs (i-COGs) containing a bait protein. $\mathbf{b}_1$ = number of interologs (i-COGs) confirmed experimentally. $\mathbf{b}_2$ = number of interologs (PPIs) confirmed experimentally. $\mathbf{c}_1$ = percentage of confirmed literature interologs (i-COGs). $\mathbf{c}_2$ = percentage of confirmed literature interologs (i-COGs) containing a bait protein.

3.6 How reliable are these studies?

3.6.1 Overlap with small-scale interactions

Several efforts were made to identify PPIs among chemotaxis as well as flagellar proteins. For benchmarking, we thus conducted a comprehensive literature mining of PubMed abstracts resulting in 51 interactions among 39 orthologus groups (i-COGs) known to be involved in motility (supplementary Table A.3). i-COGs were identified by various, mainly small-scale methods ranging from affinity chromatography, immunoblot, Co-IP, genetic suppressor mutant screens, and Y2H to crystallography.

To compare the overlap between our gold standard (henceforth referred to as literature set) and the six interactions sets, i-COGs were predicted (Figure 1.8), i. e. i-COGs of which both orthologous groups are conserved in the respective species (Table 3.9 a_1) and the fraction of experimentally verified i-COGs was determined (b_1). As results would be biased positively towards more comprehensive studies, I predicted a second set which only contained interologs of which at least one orthologous group contained a protein which was positively tested (a_2; b_2 respectively). In both cases homodimers, i. e. interactions among the same proteins were excluded for *E. coli* predictions as both per definition of their underlying models do not contain homodimers.

Taking positively tested baits into account, the fraction of experimentally

confirmed interologs ranged from 18.2% to 28.6% for the Y2H sets. CP sets identified an overlap of 9.1%. CJE HCF missed one true interaction reported in CJE ALL. A sceening of all *E. coli* matrix interactions (38450 PPIs) revealed that both ECO sets identified all possible overlapping interactions. The outcome of this benchmarking is a false negative rate of 71.4%–81.8% for Y2H and 90.9% for CP. Confirmed literature interactions are shown in Table 3.10.

3.6.2 Overlap with predicted domain-domain interactions

I used a collection of 3034 predicted pfam [85] domain-domain interactions derived from three-dimensional structures (3DID [65]). Screening of pfam domains among interacting proteins revealed 17 distinct PPIs which contained at least one pair of interacting 3DID domains (Table 3.11). Notably, 8 out of 17 interactions were also overlapping with the literature set supporting the quality and usefulness of both approaches (Table 3.10 marked in bold). Notably, all four 3DID interactions found in *C. jejuni* were of high confidence. Here, CP performed much better than Y2H constituting 47% of all supported interactions.

3.6.3 Overlap with predicted genomic context links

Genomic context provides an evolutionary framework to predict functional realtionships (functional associations as well as physical interactions) between genes and proteins. It comprises predictions based on gene fusion, gene neighborhood and gene co-occurrence (see Section 1.4.2). STRING [29, 50] is a database of known and predicted PPIs derived from genomic context, high-throughput experiments, co-expression and literature mining based on COG orthology [56,62]. i-COGs were extracted from the PPI sets and scored based on STRING's genomic context scores. For each set a percentage distribution of i-COGs which scored greater than a specific STRING-score S was calculated (Figure 3.8). Such a distribution was also generated for 1000 randomized networks. Observed (signal) and random (noise) percentages were used to compute a signal-to-noise ratio SNR (Figure 3.9) with

$$SNR(\,S) = \log_{10} \frac{observed\,percentage\,(S)}{avg\,(random\,percentage\,(S))} \tag{3.1}$$

Overall, Y2H outperformed CP. Among the Y2H sets, TPA and HPY were mostly supported by genomic context predictions. While CJE HCF scored worse

PubMed ID	Name A	Name B	TPA	CJE ALL	CJE HCF	HPY	ECO SPK	ECO SAI
7578071	CheA	CheY	-	-	-	**HP0392-HP1067**	-	-
10998179	FliI	FliH	-	-	-	HP1420-HP0353	-	-
9095196, 9765212	FliA	FlgM	-	-	-	-	**b1922-b1071**	**b1922-b1071**
10320579	FlgK	FlgN	-	-	-	-	b1082-b1070	b1082-b1070
10783392	PomA	PomA	TP0725-TP0725		-	-	-	-
8757288	MotA	FliM	-	Cj0337c-Cj0060c	-	-	-	-
10998179	FliH	FliH	-	-	-	HP0353-HP0353	-	-
11327763, 12958592	FliC	FliS	TP0792-TP0943	Cj0720c-Cj0549	Cj0720c-Cj0549	HP0601-HP0753	-	-
11327763, 12958593	FliC	FliS	TP0868-TP0943	Cj1339c-Cj0549	Cj1339c-Cj0549	HP0115-HP0753	-	-
11327763, 12958594	FliC	FliS	TP0870-TP0943	Cj1338c-Cj0549	Cj1338c-Cj0549	-	-	-
10320579	FlgL	FlgN	-	-	-	-	b1083-b1070	b1083-b1070
11204784	FlhF	FlhF	-	**Cj0064c-Cj0064c**	**Cj0064c-Cj0064c**	-	-	-
11327763	FliS	FliS	**TP0943-TP0943**	-	-	-	-	-
8757288	FliG	FliG	**TP0026-TP0400**	-	-	-	-	-
10809678, 15126479	FliG	FliF	TP0026-TP0399	-	-	-	-	-
10809678, 15126480	FliG	FliF	TP0400-TP0399	-	-	-	-	-
8631704	FliG	FliM	TP0026-TP0721	-	-	-	-	-
8757288	FliG	FliN	TP0026-TP0720	-	-	-	-	-
10809679	FliE	FlgB	TP0398-TP0396	-	-	-	-	-
9791106	FliM	FliN	-	**Cj0060c-Cj0059c**	**Cj0060c-Cj0059c**	-	-	-
8757288	FliN	FliN	-	**Cj0351-Cj0059c**	**Cj0351-Cj0059c**	-	-	-
8757289	FliN	FliN	-	**Cj0059c-Cj0059c**	**Cj0059c-Cj0059c**	-	-	-

Table 3.10 | Confirmed literature interactions. Interactions supported by 3DID domain-domain interactions are highlighted in bold.

Set	Locus A	Locus B	Gene A	Gene B	Pfam A	PfamB
TPA	TP0400	TP0026	fliG-2	fliG-1	FliG-C	FliG-C
TPA	TP0943	TP0943	fliS	fliS	FliS	FliS
CJE ALL	Cj0059c	Cj0060c	fliY	fliM	SpoA	SpoA
CJE ALL	Cj0059c	Cj0351	fliY	fliN	SpoA	SpoA
CJE ALL	Cj0064c	Cj0064c	flhF	flhF	MobB	MobB
CJE ALL	Cj0064c	Cj0064c	flhF	flhF	SRP54	SRP54
CJE ALL	Cj0059c	Cj0059c	fliY	fliY	SpoA	SpoA
CJE HCF	Cj0059c	Cj0060c	fliY	fliM	SpoA	SpoA
CJE HCF	Cj0059c	Cj0351	fliY	fliN	SpoA	SpoA
CJE HCF	Cj0064c	Cj0064c	flhF	flhF	MobB	MobB
CJE HCF	Cj0064c	Cj0064c	flhF	flhF	SRP54	SRP54
CJE HCF	Cj0059c	Cj0059c	fliY	fliY	SpoA	SpoA
HPY	HP0391	HP0392	cheW	cheA	CheW	H-kinase-dim
HPY	HP0391	HP0392	cheW	cheA	CheW	HATPAse-c
HPY	HP1067	HP0392	cheY	cheA	Hpt	Response-reg
HPY	HP1067	HP0392	cheY	cheA	Response-reg	Response-reg
HPY	HP1198	HP1032	rpoBC	fliA	RNA-pol-Rpb1-1	Sigma70-r2
HPY	HP1198	HP1032	rpoBC	fliA	RNA-pol-Rpb1-1	Sigma70-r4
ECO SPK	b1071	b1922	flgM	fliA	FlgM	Sigma70-r2
ECO SPK	b1071	b1922	flgM	fliA	FlgM	Sigma70-r3
ECO SPK	b1071	b1922	flgM	fliA	FlgM	Sigma70-r4
ECO SPK	b1883	b1914	cheB	uvrY	CheB-methylest	Response-reg
ECO SPK	b1883	b1914	cheB	uvrY	GerE	Response-reg
ECO SPK	b1883	b1914	cheB	uvrY	Response-reg	Response-reg
ECO SPK	b1922	b3988	fliA	rpoC	RNA-pol-Rpb1-1	Sigma70-r2
ECO SPK	b1922	b3988	fliA	rpoC	RNA-pol-Rpb1-1	Sigma70-r3
ECO SPK	b1922	b3988	fliA	rpoC	RNA-pol-Rpb1-1	Sigma70-r4
ECO SPK	b1922	b1040	fliA	csgD	Sigma70-r2	Sigma70-r4
ECO SPK	b1922	b1040	fliA	csgD	Sigma70-r3	Sigma70-r4
ECO SPK	b1922	b1040	fliA	csgD	Sigma70-r4	Sigma70-r4
ECO SAI	b1888	b4113	cheA	basR	Hpt	Response-reg
ECO SAI	b1888	b4170	cheA	mutL	CheW	HATPAse-c
ECO SAI	b1888	b4170	cheA	mutL	DNA-mis-repair	HATPAse-c
ECO SAI	b1888	b4170	cheA	mutL	HATPAse-c	HATPAse-c
ECO SAI	b1071	b1922	flgM	fliA	FlgM	Sigma70-r2
ECO SAI	b1071	b1922	flgM	fliA	FlgM	Sigma70-r3
ECO SAI	b1071	b1922	flgM	fliA	FlgM	Sigma70-r4
ECO SAI	b1886	b4355	tar	tsr	MCPsignal	MCPsignal
ECO SAI	b1886	b4355	tar	tsr	TarH	TarH
ECO SAI	b1040	b1922	csgD	fliA	Sigma70-r2	Sigma70-r4
ECO SAI	b1040	b1922	csgD	fliA	Sigma70-r3	Sigma70-r4
ECO SAI	b1040	b1922	csgD	fliA	Sigma70-r4	Sigma70-r4

Table 3.11 | Interactions supported by 3DID domains

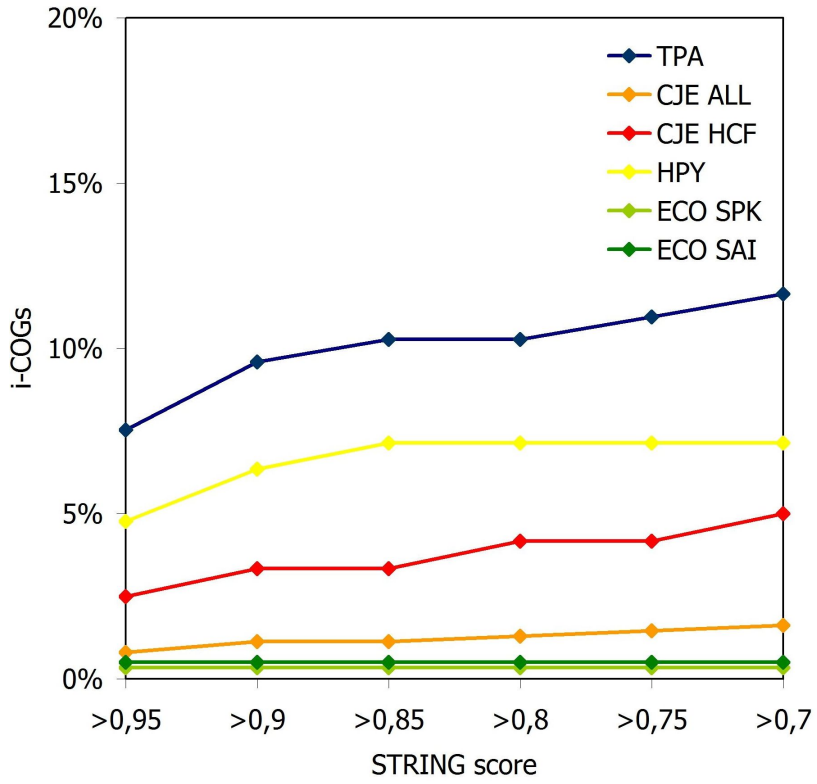

Figure 3.8 | Percentage of high-confidence genomic-context links found in motility networks. Percentage of i-COGs (y-axis) which scored greater than a specific STRING-score (x-axis). Only high (> 0.7) and highest confidence links (> 0.9) are shown (confidence as defined by [29]).

it surpassed CJE ALL whose signal is hardly distinguishable from its noise.

3.6.4 Co-localization of interacting proteins

PSORTb 2.0 [64] is a database of protein locations that were predicted computationally using tools such as SubCellularLocalisationBlast (SCL - BLAST & SCL BLASTe), Support Vector Machines (SVMs), Motif and Profile Analysis, Outer Membrane Motif Analysis, HMMTOP , and Signal Peptide. Each tool focuses on a specific biological feature and predicts one or more localization sites. Each single result is weighted and multiple results are combined to generate the final prediction. PSORTb differentiates between 5 localization sites for Gram-negative bacteria cytoplasm, cytoplasmic membrane, periplasm, outer membrane and extracellular space. For each set the percentage of PPIs whose

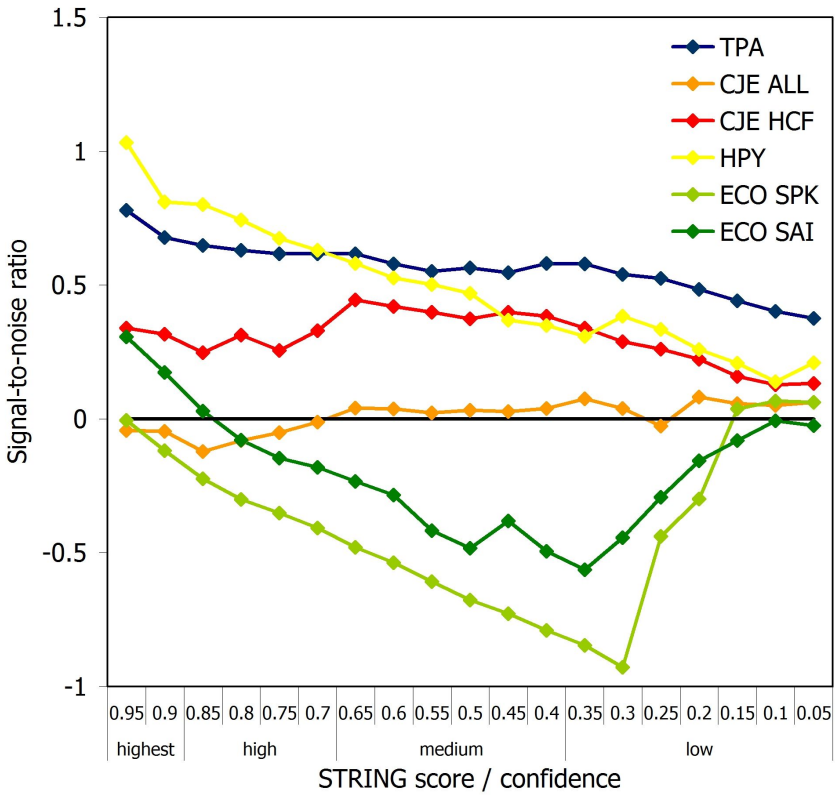

Figure 3.9 | Genomic context signal-to-noise ratio. Percentage of i-COGs (signal) which scored greater than a specific STRING-score (x-axis) compared to the percentage expected from the randomized networks (noise). A signal-to-noise ratio (y-axis) above zero indicates that the signal was stronger than the noise, i. e. the observed percentage was higher than the average percentage found in the randomised networks. Confidence as defined by [29].

interacting proteins share the same predicted localization (except those with 'unknown' localization predictions) was calculated. To estimate the significance of co-localization, observed percentages were compared with those of the randomized networks. Except for ECO SAI, observed co-localization was higher than the mean of the random networks (Figure 3.10). While co-localization in TPA, CJE ALL and ECO SPK was significantly higher with $p<0.05$, that in CJE HCF and HPY was only significant with a probability value of $\sim$10% (Table 3.12). Socio-affinity linked proteins of ECO SAI seem to have vastly different localizations, even more different than the means of interacting proteins in the randomized networks.

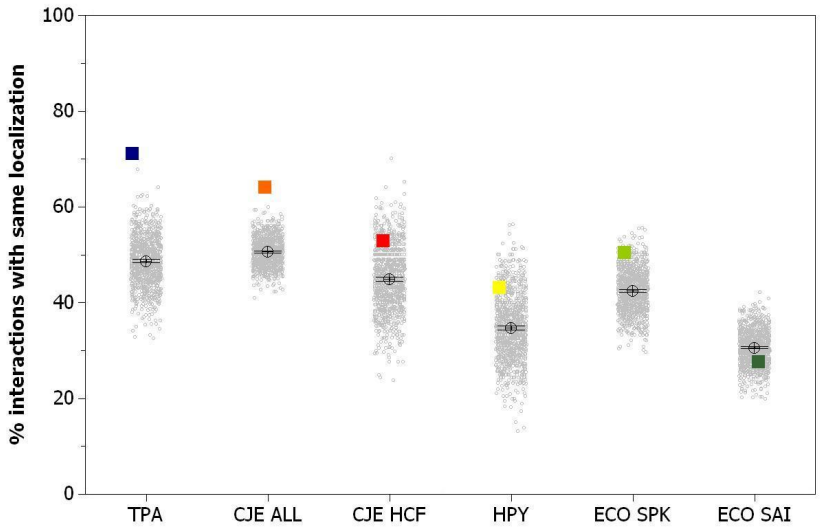

Figure 3.10 | Observed versus random co-localization. Illustrates the percentage of interactions found in the indiviual interaction sets whose proteins share the same localization (colored squares) compared to what would be expected from their randomized networks (1000 randomisation (grey circles), mean (black crosses), 95% CI of mean (black bars)). In both cases, PPIs were excluded if the localization of one or both proteins was 'unknown'.

3.6.5 Overlap with swarming mutants

Although physically linked to known motility proteins, functional relevance of associated proteins remains unclear. Therefore, genes whose deletion affected motility were integrated. Systematic gene mutants were tested for their impact on motility in *B. subtilis* [67]. This set was complemented by swarming mutants identified by a comprehensive screening of 3985 *E. coli* mutant strains [10]. Both datasets contain a similar number of mutants: 146 for *B. subtilis* and 159 for *E. coli*. About 4% of genes in both species show an effect on motility under the conditions tested. Among these are 45 (30%) and 43 (27%) known motility related genes, respectively.

57% (53%) of *E. coli* (*B. subtilis*) were found among interacting proteins implying that either half of the mutants has not been identified to be directly linked to motility, e. g. house-keeping proteins, or are not conserved in the respective species. The percentage of orthologs/proteins which have shown to be essential for motility ranged from 38% in TPA to 14% in ECO SAI (Figure 3.11 and Table 3.13). Among those were known motility and motility associated proteins

	TPA	**CJE ALL**	**CJE HCF**	**HPY**	**ECO SPK**	**ECO SAI**
Random mean	0.487	0.506	0.449	0.347	0.424	0.306
Random stdv.	0.050	0.029	0.066	0.067	0.044	0.036
Observed value	0.712	0.641	0.529	0.432	0.505	0.277
Z-score	4.543	4.590	1.235	1.261	1.813	-0.803
p value	$< 10^{-3}$	$< 10^{-3}$	0.109	0.104	0.035	0.496

Table 3.12 | Significance of co-localization over random networks

Set	**Proteins**	**Number of proteins**	**Percentage of proteins**
TPA	all	42	38.2%
	motility	31	91.2%
	associated	11	14.5%
CJE ALL	all	103	19.6%
	motility	34	97.1%
	associated	69	14.1%
CJE HCF	all	44	33.1%
	motility	28	96.6%
	associated	16	15.4%
HPY	total	44	31.2%
	motility	30	96.8%
	associated	14	12.7%
ECO SPK	all	50	19.5%
	motility	37	75.5%
	associated	13	6.3%
ECO SAI	all	53	14.2%
	motility	36	75.0%
	associated	17	5.2%

Table 3.13 | Overlap with swarming mutants. Numbers are derived from orthologous proteins which have either shown to be essential for motility in *E. coli* or in *B. subtilis*.

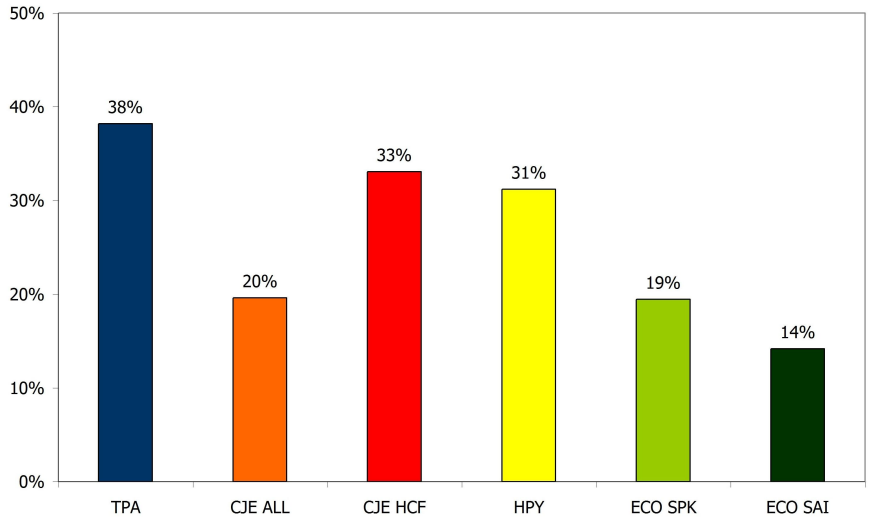

Figure 3.11 | Percentage of interacting proteins with motility phenotype. Percentage is derived from orthologous proteins which have either shown to be essential for motility in *E. coli* or in *B. subtilis*.

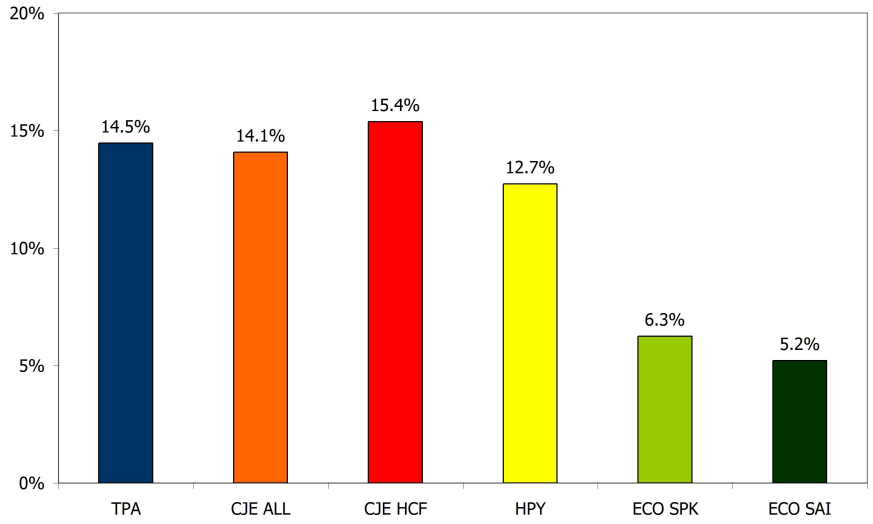

Figure 3.12 | Percentage of associated proteins with motility phenotype. Percentage is derived from orthologous proteins which have either shown to be essential for motility in *E. coli* or in *B. subtilis*.

(either proteins with different or unknown function). As expected, the overlap with known motility proteins (Table 3.13) was with 75% (ECO SPK and ECO SAI) and 97% (CJE ALL) high whereas the overlap with associated proteins was much smaller ranging from 5% in ECO SAI to 15% in CJE HCF (Table 3.12). An overview of interacting proteins with motility phenotype is given in supplementary Table A.5.

In both cases, CJE HCF contained a higher fraction of mutant orthologs than CJE ALL supporting its higher reliability. Overall, the Y2H sets, except for CHE ALL contained $\sim$10% more essential motility proteins when compared to the CP sets. Furthermore, the fraction among motility associated proteins was approximately two-fold greater. Thus, proteins indentified by Y2H seem to have a higher functional relevance than those identifed by CP. While Y2H identifies direct links (distance 1), links among CP proteins may have a greater distance (mediated by a subcomplex).

An integrated view of bacterial motility

To account for experimental errors and evolutionary variations, I performed pairwise alignments of the individual networks using the PathBLAST method proposed by Kelley et al. [59,72]. Homologous proteins and their interactions were identified based on their sequence similarity using BLAST (E-value $\leq 10^{-5}$). Notably, such an aligned network is not restricted to conserved proteins which are interacting in both sets. A gap is included if conserved proteins do not directly interact but are indirectly linked via a common protein [59].

Aligned protein networks are given in Figure 3.13 and Figure 3.14. In addition to the conservation of protein pairs according to their BLAST E-value , swarming mutants as well as links among 3DID domains were integrated. A complete list of the PathBLAST results including BLAST E-value is given in supplementary Table A.6.

Although these networks provide insights about conserved proteins and their interactions, it is difficult to get an overall picture, i. e. to relate all observations to each other. Furthermore, many interactions were identified among paralogs, e. g. interactions among FliC and FliS orthologs. Most importantly, as proteins were solely aligned based on homology they do not necessarily represent orthologs.

To solve this issues, I integrated all six aligned protein networks into a single network, henceforth referred to as core network (Figure 3.15). Nodes represent

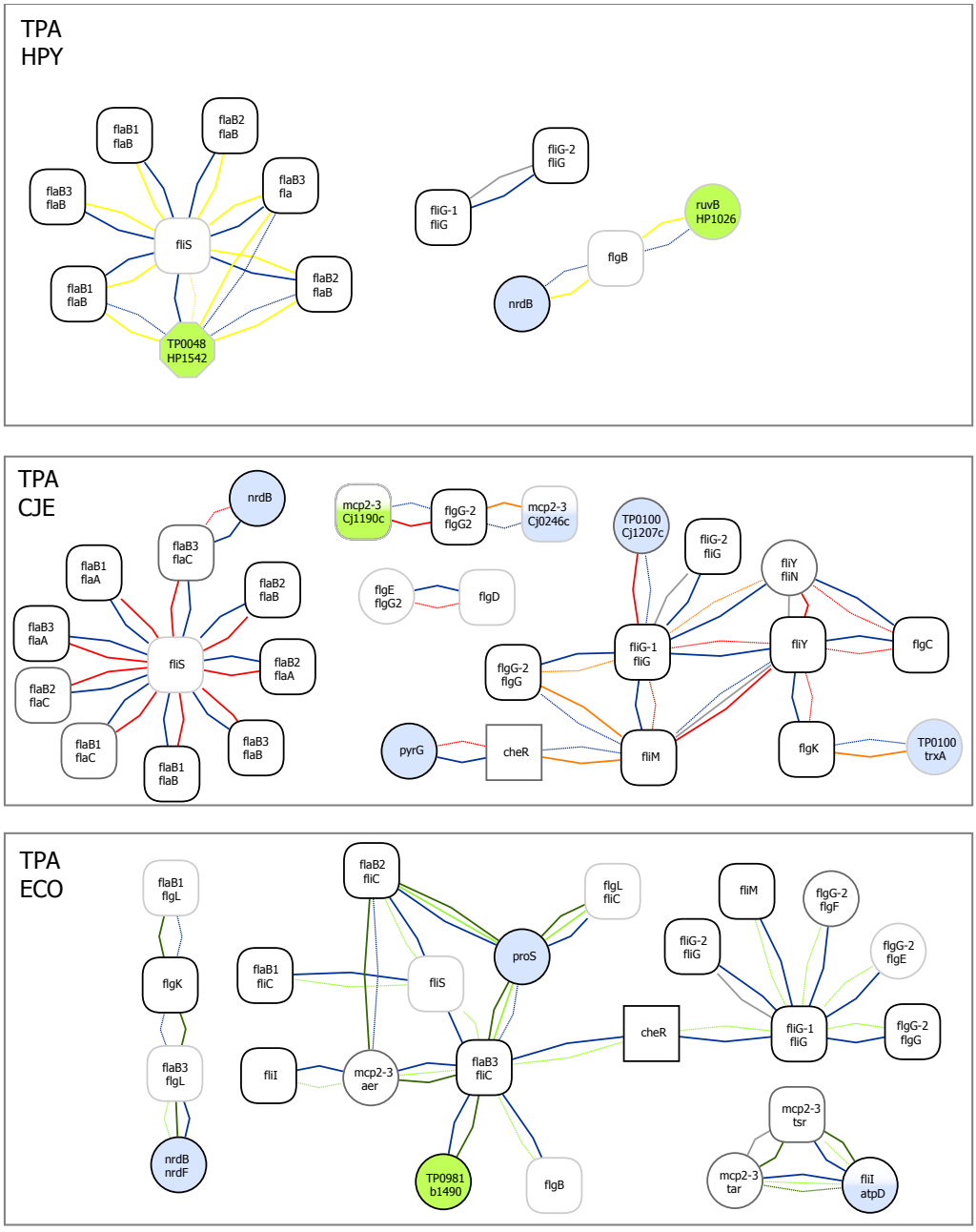

Figure 3.13 | Aligned protein-protein interaction networks part I

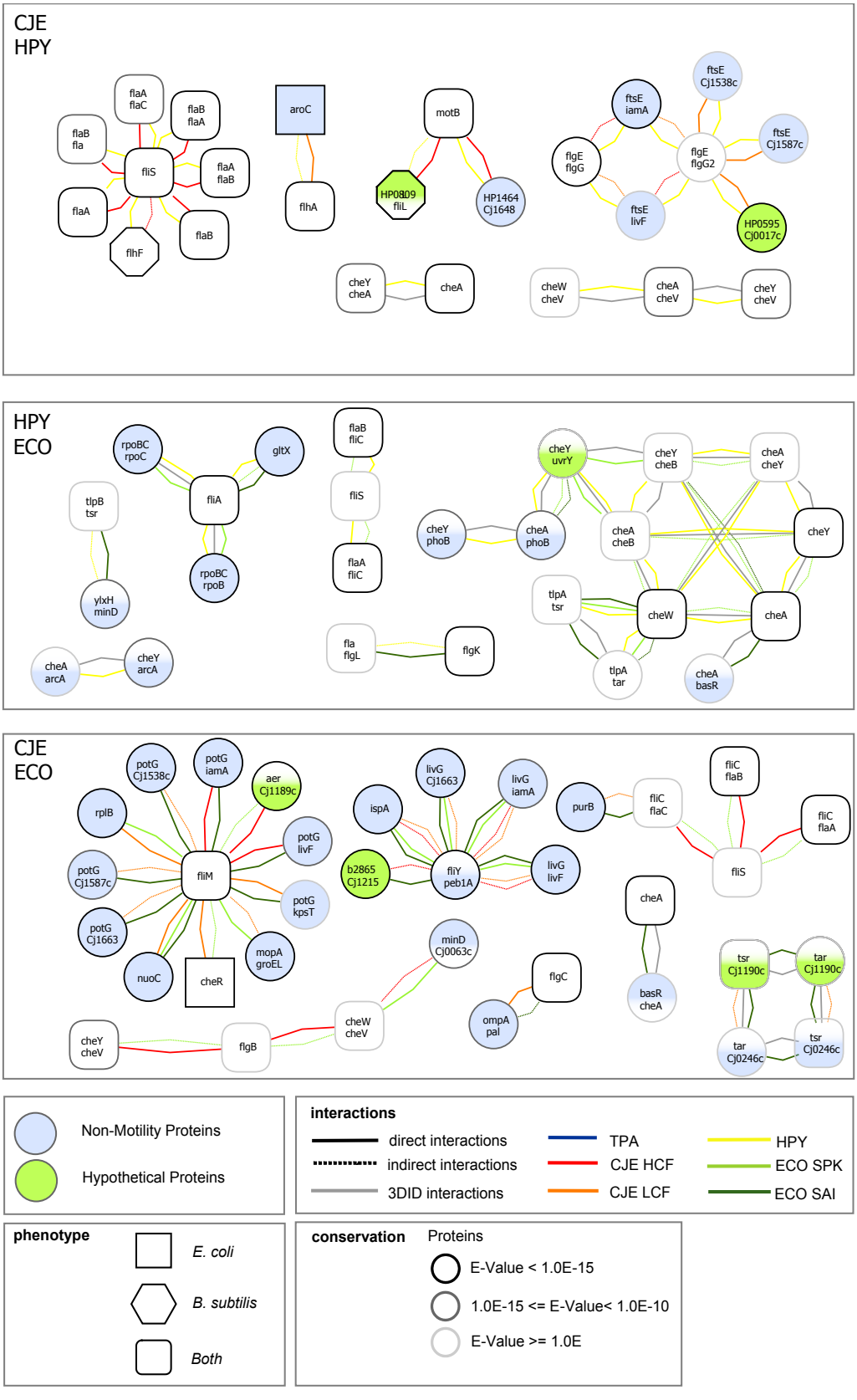

Figure 3.14 | Aligned protein-protein interaction networks part II

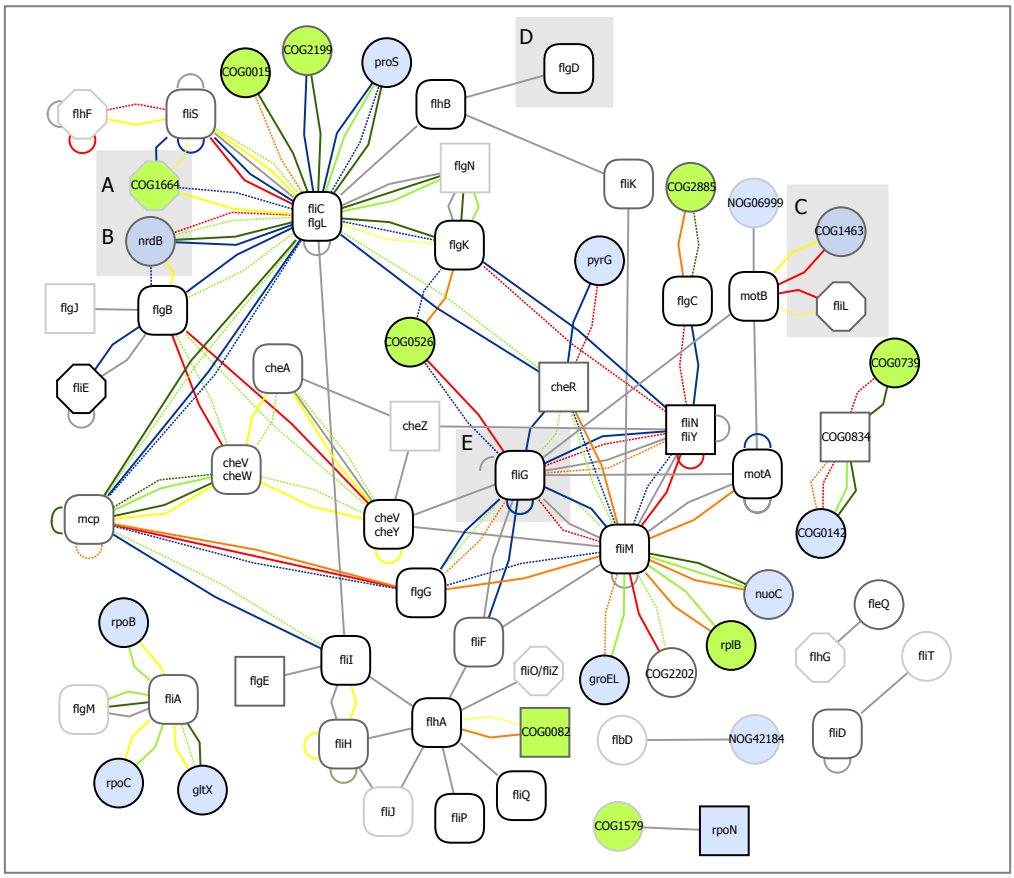

Figure 3.15 | Core motility network

orthologous protein families rather than individual proteins which reduces complexity and improves quality. In addition to motility PPIs, I integrated literature interactions, motility phenotypes and evolutionary conservation. Here, conservation reflects the ratio of species in which a certain orthologous group is conserved restricted to flagellated species (with the filament protein FliC).

The tightly clustered (average clustering coefficient= 0.14) core comprises 96 interactions between 65 orthologous groups. The most highly connected orthologous groups were FliC/FlgL, FliM and FliG. Among interactions, 73% connected known motility groups, 30% were identified in more than one species and 45% were predicted to be strongly associated by STRING (highest confidence: $S > 0.9$).

Among orthologous families, 68% contained orthologs known to be essential for motility. Clearly, these numbers indicate that this conserved core is much

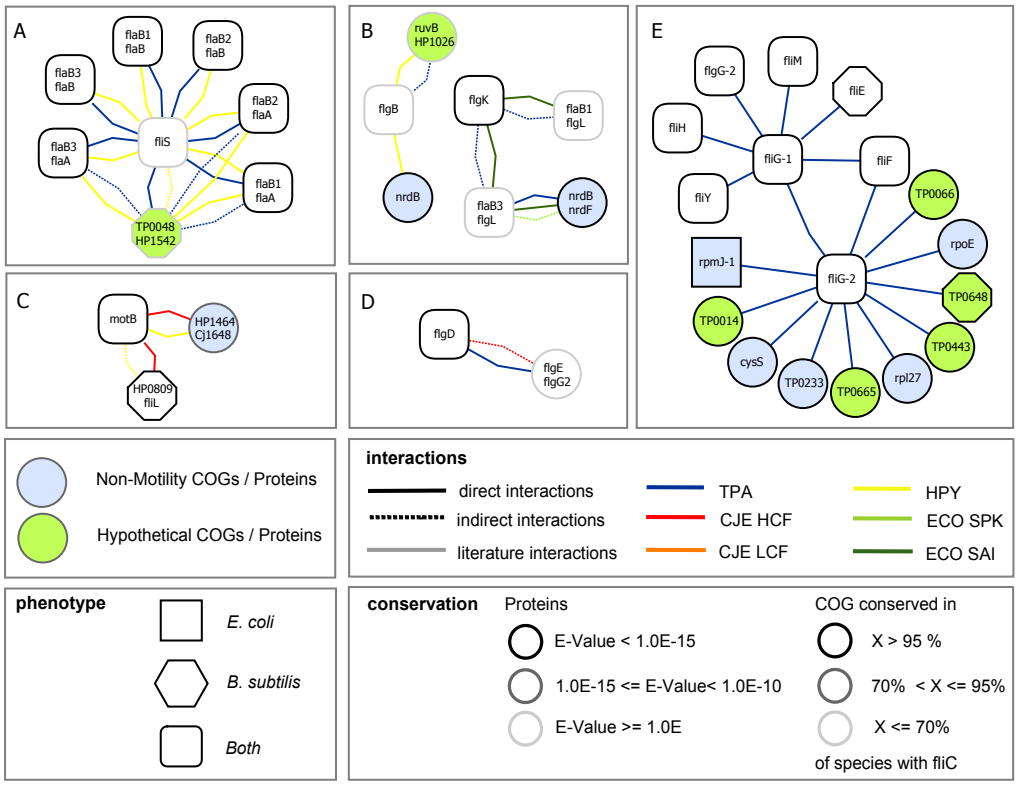

Figure 3.16 | Legend and selected parts of the core motility network

more reliable and biologically relevant than any of the individual networks.

The core network incorporates and connects many known components (white nodes) which are crucial for bacterial taxis. For instance, the interaction of FliC with its chaperon FliS is seen in all species except *E. coli*. FlgL is connected to the second hook-associated protein FlgK and both are stabilized by their export chaperon, FlgN. The basal body complex with FliN/FliY, FliG, FliM, and the export system component, FliF forms another functional module. It is connected to the motor proteins motA and motB as well as to rod proteins like FlgC and FlgG. Interestingly, evidence for a direct interaction of FliM with motA is provided. Orthologous groups involved in chemotaxis signaling are only connected to the basal body via literature interactions (grey lines). In addition to previously known inter-motility interactions, conserved links between chemotaxis proteins and rod proteins like FlgB and FlgG can be found. Another interesting connection is the conserved motB-FliL interaction in TPA and HPY (Figure 3.16 C). For *Proteus mirabilis* FliL is thought to be involved in sensing of the actual

flagellum status [86]. TPA and HPY interactions proved evidence that this sensing is mediated by a direct connection to the motor apparatus. Besides connections between known motility components, interesting links of flagellar proteins and proteins of other functional classes (blue nodes), and proteins of unknown function (green nodes) can be observed. NrdB (ribonucleoside-diphosphate reductase), the key enzyme for the conversion of ribonucleosides into desoxyribonucleosides, would not usually be assumed to be directly involved motility. Strikingly, conserved interactions of this enzyme to two flagellar proteins, FliC and FlgB can be found (Figure 3.16 B). An orthologous group of an ABC-type transport system (COG1463) is found to directly interact with motB in *C. jejuni* and *H. pylori* (Figure 3.16 C). Such a link between an ABC-transporter and a motor protein is also observed in the TonB-dependent Fe-uptake [87]. Furthermore, one can also gain insights into species-specific modules. Here, the spirochetes' flagellum (*T. pallidum*) is of special interest. One unique feature is its periplasmatic localization of two asymmetrically rotating flagellum bundles fixed to the cell poles. The molecular basis of this asymmetry is unknown. FliG is thought to play a role in this behavior since it is the only duplicated basal body complex protein in spirochetes. In *T. pallidum* these paralogs are called FliG-1 (TP0026) and FliG-2 (TP0400). Despite high sequence similarity both proteins show a differential interaction pattern (Figure 3.15 E). Although these patterns do not clearly explain the asymmetric behavior, they provide evidence that these two proteins are functional distinct.

3.7 Conserved hypotheticals involved in motility

Interaction sets contain a huge portion of motility-associated proteins (Table 3.1). Among those, proteins which are conserved but have no (or only a vague) functional annotation are of special interest. As the number of conserved hypotheticals (CHPs) among the interaction sets ranged from a few in TPA to hundreds of proteins in CJE ALL (Figure 3.1) ranking of potential new motility candidates became essential. CHPs were ranked based on the reliability of their motility interaction(s), swarming mutant overlap, FhlD regulation, STRING motility association, FliC co-occurrence (see Section 2.6 for more details). A list of top ten ranked CHPs is given in Table 3.14 and Table 3.15.

	COG	CHP	Neighbor(s)	Score
TPA	COG0457	TP0648	fliG-2	5.529
	COG1699	TP0658	flaB2,flaB1,flaB3	2.522
	COG2199	TP0981	flaB3	1.778
	COG3391	TP0421	TP0567	0.534
	COG1664	TP0048	fliY,fliS	0.231
	COG1774	TP0046	fliE,cheR,flgD,flaB3,TP0959	0.128
	COG1512	TP0561	fliF,flhB,fliR,fliQ,fliL	0.103
	NOG46983	TP0711	flhF,flaB3	0
	NOG45794,COG1208,COG1207	TP0851	cheR,cheR,cheR	0
	NOG43115	TP0174	flaB3	0
CJE	COG0457,COG0419	Cj0055c	fliM,fliM	13.822
	COG0457	Cj0497	fliM	8.293
	COG0457	Cj1637c	flgG2	8.293
	COG0642,COG2202,COG4191	Cj1492c	flgG2,flgG2,flgG2	7.76
	COG0840,COG0840	Cj1190c	flgG2,flgG2	7.401
	COG3206,COG0642,COG0419	Cj0254	fliG,fliG,fliG	5.82
	COG0642	Cj1222c	fliN	5.82
	COG0840	Cj0092	fliL,fliG	5.55
	COG0840	Cj0202c	fliY	5.55
	COG0419,NOG12190,COG0840	Cj0700	fliR,fliR,fliR	5.55
HPY	COG0457	HP1479	flgB	8.293
	COG0419,COG1196	HP0488	flgE,flgB,flgE,flgB	4.675
	COG0419,COG1196	HP1116	flgB,flgB	4.675
	COG1495,COG1196	HP0595	flgE,flgE	3.456
	COG1196	HP0120	flgB	3.456
	COG0419,NOG13219	HP0406	fliH,flgB,fliH,flgB	2.337
	COG1699	HP1154	flaA	1.892
	COG1699	HP1377	flaA	1.892
	COG0210,COG0443,NOG44676	HP0149	tlpB,ylxH,tlpB,ylxH,tlpB,ylxH	1.762
	COG0791	HP0087	flgB	1.398
ECO	COG2199	b1490	fliC	1.778
	NOG27152,COG0791	b3937	mbhA,mbhA	1.398
	COG0791	b1655	fliJ	1.398
	COG0451,COG0451	b0868	tar,tsr,cheW	1.258
	COG2197,COG2197	b1914	flgN,flgL,cheB,fliG,flgL	0.803
	COG1309,COG1309	b3641	mbhA,mbhA	0.793
	COG0842,COG0842	b0793	flgB,sfmC,flgB	0.743
	COG1396	b3021	motA	0.635
	COG3121	b2110	flgA	0.441
	COG0726,COG0726	b0130	flgG,flgG	0.435

Table 3.14 | Top ten conserved hypotheticals

	CHP	ECO MUT	BSU MUT	CJE MUT	HPY MUT	FHLD EXP
TPA	TP0648	-	X	-	X	-
	TP0658	-	X	-	-	-
	TP0981	-	-	-	-	-
	TP0421	X	-	-	-	-
	TP0048	-	X	-	-	-
	TP0046	-	X	-	-	-
	TP0561	-	X	-	-	-
	TP0711	-	-	-	-	-
	TP0174	-	-	-	-	-
	TP0064	-	-	-	-	-
CJE	Cj0055c	-	X	-	X	-
	Cj0497	-	X	-	X	-
	Cj1637c	-	X	-	X	-
	Cj1492c	X	X	-	X	X
	Cj1190c	X	X	-	-	X
	Cj1222c	X	X	-	X	-
	Cj0254	X	X	-	X	-
	Cj0202c	X	X	-	-	X
	Cj0092	X	X	-	-	X
	Cj0700	X	X	-	-	X
HPY	HP1479	-	X	-	X	-
	HP1116	-	-	X	-	-
	HP0488	-	-	X	-	-
	HP0120	-	-	X	-	-
	HP0595	-	-	X	-	X
	HP0406	-	-	-	-	-
	HP1377	-	X	-	-	-
	HP1154	-	X	-	-	-
	HP0149	X	-	-	-	-
	HP0087	-	X	-	-	-
ECO	b1490	-	-	-	-	-
	b3937	-	X	-	-	-
	b1655	-	X	-	-	-
	b0868	-	-	-	-	-
	b1914	-	-	-	-	-
	b3641	-	-	-	-	-
	b0793	-	X	-	-	-
	b3021	-	X	-	-	-
	b2110	-	-	-	-	-
	b0130	-	X	-	-	-

Table 3.15 | Top ten motility-associated conserved hypotheticals with experimental evidence. X indicates that an ortholog is essential for motility in *E. coli* (ECO MUT), in *B. subtilis* (BSU MUT), in *C. jejuni* (CJE MUT), in *H. pylori* (HPY) or is regulated by FhlD (FHLD EXP).

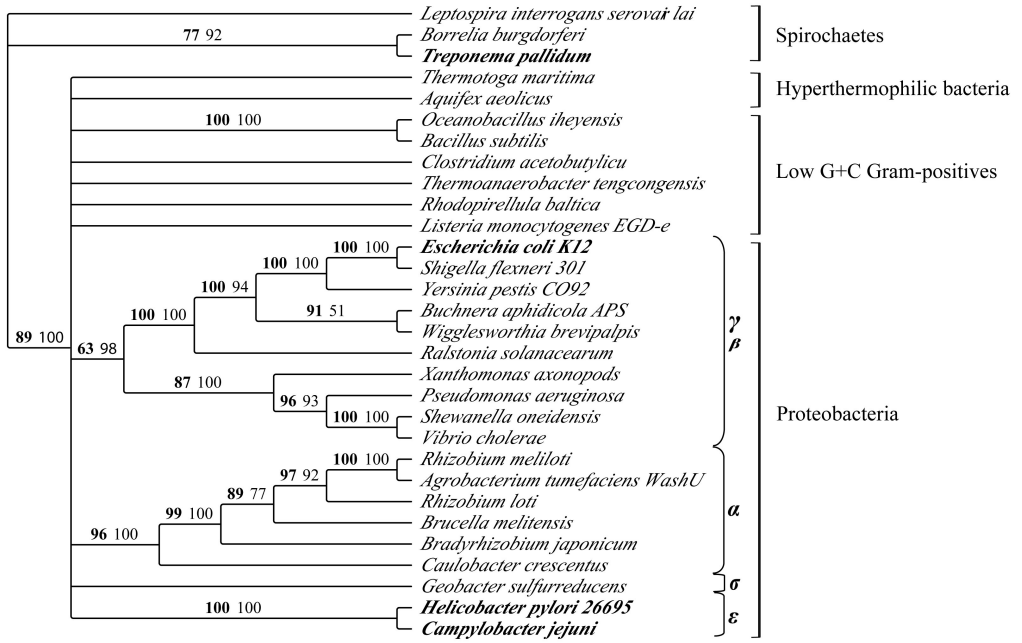

Figure 3.17 | Supertree of the flagellum complex. Bacterial flagellum supertree of 30 species constructed with 35 protein families. Two alternative treeing methods, maximum parsimony (MP) and neighbor-joining (NJ) were used to generate bootstrapped (100 replicates) protein family trees merged into supertrees. The cladogram reflects the consensus of these supertrees generated and merged by the CLANN software [82]. Numbers along the branches are the bootstrap values which inidcate reproducibility of each branch during bootstrap analysis (100 replicates) of MP analyses of the supertrees. Bootstrap values of the MP supertree are marked in bold, values of the NJ supertree are marked in plain.

3.8 Phylogeny of the flagellum

To put the four species and their aligned network into a broader evolutionary context, a phylogenetic analysis of 30 species based on flagellar protein families (Figure 1.2) was conducted. First, protein family trees were inferred from highly conserved regions of 35 protein families using two alternative treeing methods as described in Section 2.5. Next, protein family trees were merged into a single tree using the supertree approach.

Flagellar phylogeny strongly supports monophyly of spirochaetes, γ and β, ϵ, and α proteobacteria while low G+C Gram positives are poorly resolved (Fig-

ure 3.17). Monophylies, suggest that spirochaetes possess the most differential flagellar machinery while those of other groups seem to be more similar. Phylogeny inferred from ribosomal RNA (rRNA) is often considered as the gold standard as it is derived from the most ubiquitous and constrained molecules available. Except for G+C Gram positives, the reported monophylies are in line with universal rRNA trees which have shown that spirochaetes and the subdivisions of proteobacteria are strongly monophyletic [88]. Spirochaetes have the earliest derived flagellum if we combine the flagellum phylogeny with results from Brown et al. who has shown that spirochaetes are the earliest while proteobacteria are the most recently derived bacteria [89].

To examine the evolutionary conservation of the core network (Figure 3.15), its 96 i-COGs were used for phylogenetic profiling (30 species) and results (blue stretches) were mapped onto the supertree (Figure 3.18). I-COGs have been ranked and stretches were drawn according to their conservation ratio. Dark blue stretches reflect conserved i-COGs found in more than one set (including literature interactions). Although expected those are not necessarily conserved among all species. Nevertheless, it is obvious that most interactors of the core network are well conserved among the 30 species indicating that interaction results may easily be transferred to the other 26 species without loosing much information. Strikingly, parts are in line with the phylogeny of the supertree, e. g. the monophyletic group of α proteobacteria. Also *Buchnera aphidicola APS* and *Wigglesworthia brevipalpis* form a group. Although *Rhodopirellula baltica* and *Aquifex aeolicus* were not closely related by the flagellum phylogeny, profiling suggests a similar evolution. Furthermore, phylogenetic profiling revealed that parts of the network are not well conserved in α proteobacteria (Figure 3.19).

3.9 Prediction of motility interactions

61 reliable motility interactions of the core network (literature and conserved interactions) were used to predict protein-protein interactions for 64 other flagellated bacteria. To filter out orthologous proteins which are only partially conserved, predictions were restricted to proteins with a COG family conservation of more than 50% of their size. In total, 18,110 motility interactions were predicted. Predictions for *Listeria monocytogenes*, *Bacillus anthracis*, and *Shigella flexneri* are summarized in supplementary Table A.7.

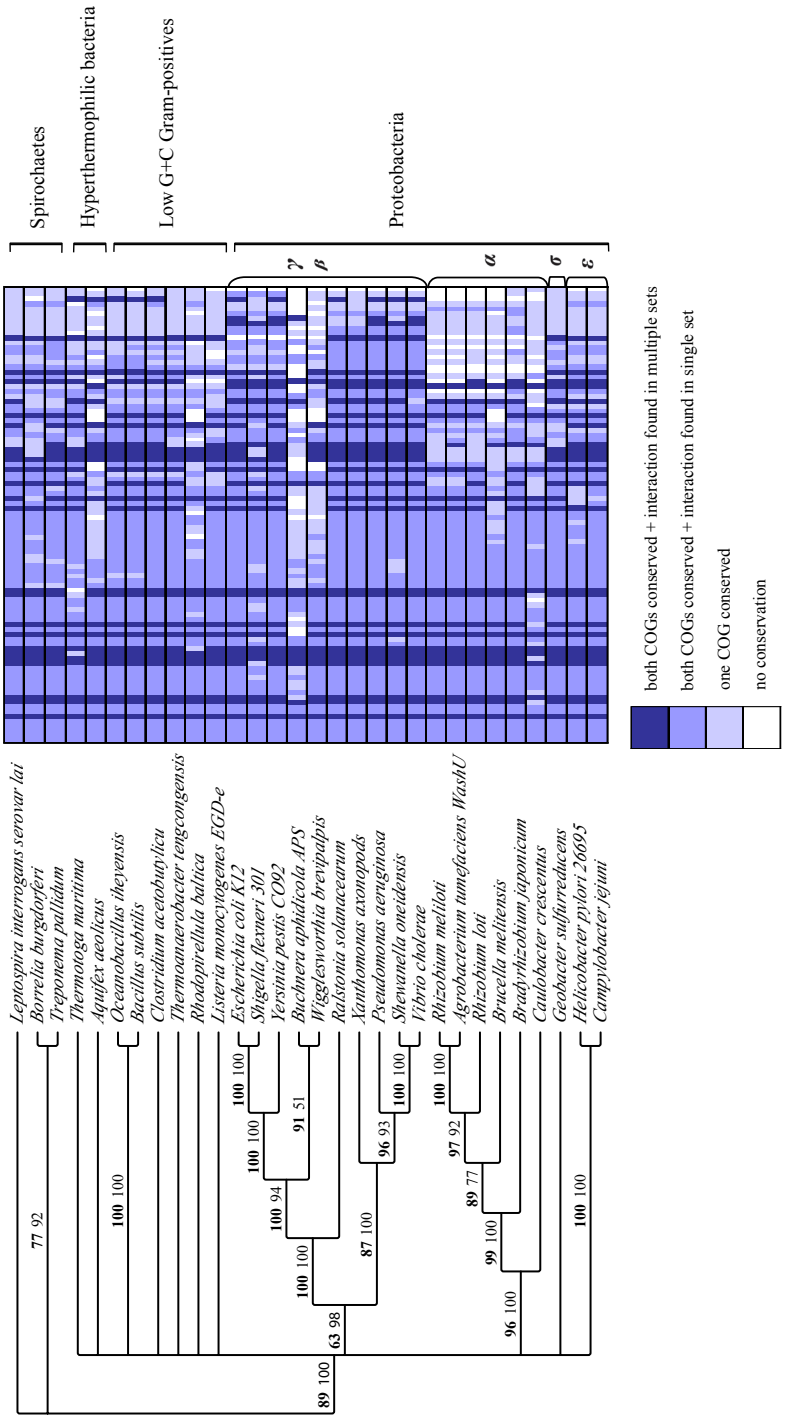

Figure 3.18 | Supertree with phylogenetic profile of the core network

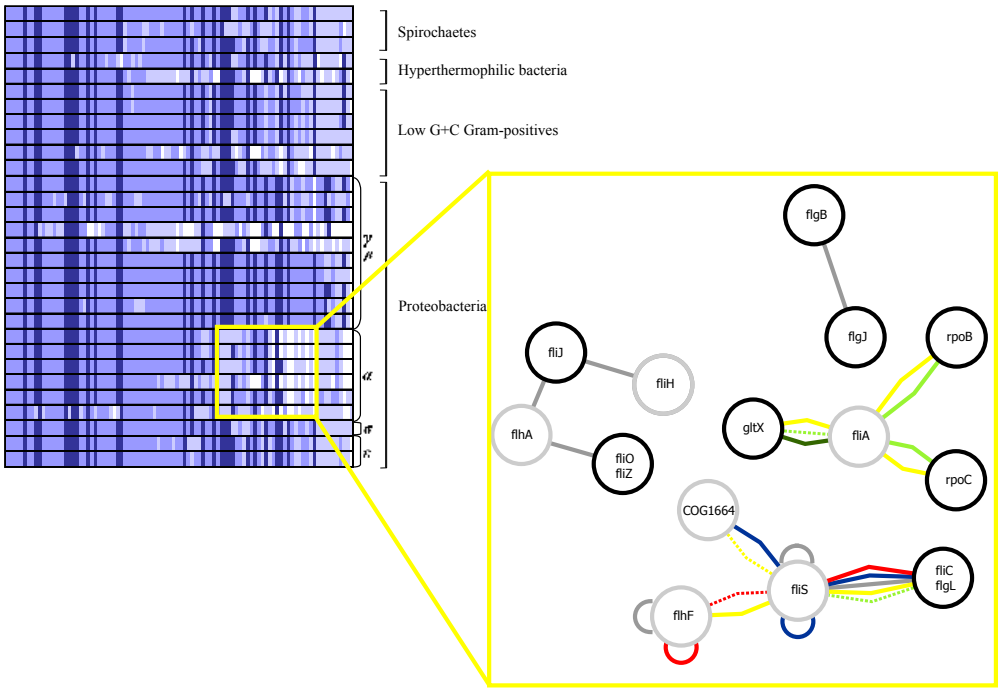

Figure 3.19 | Part of the core network which is not conserved in alpha proteobacteria. Black border colors indicate that a certain orthologus group is conserved while grey border colors indicate that a certain orthologus group is not conserved in alpha proteobacteria.

Chapter 4

Discussion

This study is the first comparative analysis of motility interactions of four bacteria detected by two different high-throughput methods. It mainly aimed to identify a conserved core of protein-protein interactions which are essential for chemotaxis signaling and flagellar complex formation. Unfortunately, an ultimate picture of such a core is hampered by limitations of the experimental methods.

4.1 False-negatives

Results suggest that many physiologically relevant interactions were not detected. For instance, only a partial fraction of motility proteins were tested successfully. One popular way to estimate the percentage of missed interactions is based on comparison with small-scale interactions gathered from the literature or PPI databases. One drawback of such false-negative benchmarking is that systematic differences between PPI detection methods might lead to overestimation of false-negative rate. For instance, Aloy and Russell showed that Y2H tends to detect transient interactions, whereas interactions within protein complexes are more efficiently detected using CP [32]. Structural analysis of protein complexes identifies weak interactions that seem not to be reproducible by any other method [46]. Given that different types of interactions are detected, estimation of false-negative rate is not trivial.

Literature benchmarking of motility interactions implies a false-negative rate of 71%–81.8% for Y2H and 91% for CP. As only positiveley tested baits were considered, percentage of missed interactions might be greater. One explanation

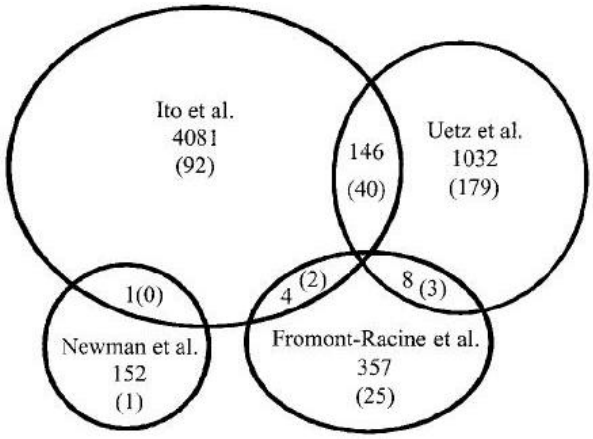

Figure 4.1 | Overlap of high-throughput studies carried out in yeast

for Y2H false negatives are post-translational modification dependent interactions. For example, phosporylated CheY (CheY-P) binds to FliM and its phosphatase CheZ. False-negatives in *E. coli* CP data may be due to a large number of membrane-associated, transmembrane proteins and homodimerizing proteins among the benchmark set. Transmembrane proteins are difficult to purify while homodimers were not predicted by the spoke and the matrix model. Overall, the literature set comprised PPIs detected by various methods that tend to identify different kinds of interactions, e. g. the interaction between MotA and FliG was reported in mutational and structural studies (supplementary Table A.3).

4.2 Overlap between motility networks

Pairwise overlap analysis reveals that conserved baits have detected vastly different preys. Although 10 protein families were positively tested in three organisms, only one interaction could be reproduced (FliC with its chaperon FliS). However, even when the same bait is tested repeatedly within the same organism using the same protocol, only a fraction can be reproduced. Uetz et al. demonstrated that only about half of all Y2H screens yield reproducible interactions [21]. Gavin et al. repeatedly pulled out 139 baits and their associated proteins. On average, 69% of purified proteins were common to both purifications [30]. Furthermore, comparative studies of yeast PPIs indicated that only a small fraction of interactions is supported by more than one study (Figure 4.1) [48, 49]. The same is true for CP studies (Goll et al. unpublished and Cornell et al. [28]). Stelzl et al.

evaluated their human Y2H data by verifying a random sample of 116 PPIs by a co-immunoprecipitation assay. In total, 72 (62%) interactions could be reproduced [43].

This observation could have various reasons. PPI studies may differ in their screening protocols and non-physiological conditions. For example, Finley et al. (*C. jejuni*) and Rajagopala et al. (*T. pallidum*) not only used different binding and activation domains but also different reporter genes resulting in different steric interference and quantitative measurement. In addition to a high fraction of false-negatives, PPI detection studies might also have produced a significant number of false-positives. Here, comparison is complicated by the fact that species boundaries have to be crossed. Therefore, beside experimental limitations, evolutionary variation among proteins and their interactions have to be taken into account [59]. Another issue is that such a comparison depends on accurate algorithms to identify orthologous protein relationships [54]. Here, interolog-predictions are based on orthologous relationships predicted by the COG database [62]. Number of conserved interactions might differ considerably if a different orthology approach would be used. For instance, manually curated clusters of orthologous proteins which are part of the same KEGG pathway (KEGG Orthology (KO) [66]) or predicted clusters from the KEGG Sequence Similarity Database SSDB [6].

Appendix A

Supplementary Tables

Table A.1 | Orthologous groups involved in motility

COG	Common name
COG0455	FLHG, fleN
COG0630	CPAF, FLAI-A, flaI
COG0642	PILS
COG0643	CHEA, CHPA, PILL
COG0784	CHEV, CHEY, CHPA, PILG, PILH, PILL
COG0834	FLIY
COG0835	CHEV, CHEW, PILI
COG0840	AER, HEMAT, MCP, MCPI, tsr, MCPII, tar, MCPIII, trg, MCPIV, tap, PILJ
COG0849	PILM
COG1157	FLII
COG1191	FLIA
COG1256	FLGK
COG1261	FLGA
COG1280	CHPE
COG1291	MOTA
COG1298	FLHA
COG1317	FLIH
COG1334	FLAG
COG1338	FLIP
COG1344	FLGL, FLIC
COG1345	FLID
COG1352	CHER, PILK
COG1360	MOTB
COG1377	FLHB
COG1406	CHEX
COG1419	FLHF
COG1450	CPAC, PILQ
COG1459	PILC
COG1516	FLIS
COG1536	FLIG
COG1558	FLGC
COG1580	FLIL

Table A.1 | continued...

COG	Common name
COG1582	FLBD
COG1677	FLIE
COG1684	FLIR
COG1705	FLGJ
COG1706	FLGI
COG1749	FLGE
COG1766	FLIF
COG1776	CHEC
COG1815	FLGB
COG1843	FLGD
COG1868	FLIM
COG1886	FLINY, fliN
COG1987	FLIQ
COG1989	CPAA, FLAK-A, flaK, PILD
COG2063	FLGH
COG2165	PILA, PILE, PILV
COG2201	CHEB, CHPB
COG2202	AER
COG2204	PILR
COG2207	CHPD
COG2747	FLGM
COG2804	PILB
COG2805	PILT, PILU
COG2882	FLIJ
COG2894	CPAE
COG3143	CHEZ
COG3144	FLIK
COG3166	PILN
COG3190	FLIOZ, fliO
COG3418	FLGN
COG0031	FLA
COG0419	FLIH
COG1196	FLIH
COG1418	
COG3121	SFMC,FIMC
COG3188	SFMD,FIMD
COG3210	FLGE2
COG3539	SFMA,SFMF,FIMA,FIMI,FIMF,FIMG
COG3951	FLGJ
COG4786	FLGG
COG4787	FLGF
COG5295	FLGE
NOG04255	FLHC
NOG06008	FLIZ
NOG07455	FLHD
NOG07456	FLHE
NOG08749	FLIT
NOG14615	

Table A.2 | Bait overlap

COG ID	Name	*T. pallidum*	*C. jejuni*	*H. pylori*	*E. coli*
COG1344	FLGL, FLIC	TP0659	Cj0720c	HP0115	b1083
COG1345	FLGL, FLIC	TP0792	Cj1338c	HP0601	b1923
COG1346	FLGL, FLIC	TP0868	Cj1339c	-	-
COG1347	FLGL, FLIC	TP0870	-	-	-
COG0784	CHEV, CHEY	-	Cj0285c	HP1067	b1882
COG0835	CHEV, CHEW	-	Cj0285c	HP0391	b1887
COG1191	FLIA	TP0709	-	HP1032	b1922
COG1256	FLGK	TP0660	Cj1466	-	b1082
COG1317	FLIH	TP0401	-	HP0353	b1940
COG1516	FLIS	TP0943	-	HP0753	b1925
COG1580	FLIL	TP0722	Cj1408	-	b1944
COG1749	FLGE	TP0727	-	HP0870	b1076
COG1868	FLIM	TP0721	Cj0060c	-	b1945
COG4786	FLGG	TP0961	Cj0697	-	b1078
COG0840	MCP	TP0640	-	-	b1421
COG0841	MCP	-	-	-	b3072
COG0842	MCP	-	-	-	b4355
COG1157	FLII	TP0402	-	-	b1941
COG1196	FLIH	TP0567	-	HP0353	-
COG1352	CHER	TP0630	-	-	b1884
COG1360	MOTB	-	Cj0336c	-	b0230
COG1419	FLHF	TP0713	Cj0064c	-	-
COG1536	FLIG	TP0400	-	-	b1939
COG1677	FLIE	TP0398	Cj0526c	-	-
COG1684	FLIR	TP0716	Cj1179c	-	-
COG1766	FLIF	TP0399	-	-	b1938
COG1815	FLGB	-	-	HP1559	b1073
COG1843	FLGD	TP0728	Cj0042	-	-
COG1886	FLIN, FLIY	TP0720	Cj0059c	-	-
COG0455	FLHG	-	Cj0063c	-	-
COG0643	CHEA	-	-	-	b1888
COG1291	MOTA	TP0725	-	-	-
COG1298	FLHA	TP0714	-	-	-
COG1334	FLAG	-	Cj0547	-	-
COG1345	FLID	-	-	-	b1924
COG1377	FLHB	TP0715	-	-	-
COG1418	FLBB	TP0567	-	-	-
COG1558	FLGC	-	Cj0527c	-	-
COG1706	FLGI	-	Cj1462	-	-
COG1776	CHEC	TP0720	-	-	-
COG1987	FLIQ	TP0717	-	-	-
COG2063	FLGH	-	-	HP0325	-
COG2201	CHEB, CHPB	-	-	-	b1883
COG2202	AER	-	-	-	b3072
COG2747	FLGM	-	-	-	b1071
COG2805	PILT, PILU	-	-	-	b2950
COG2882	FLIJ	-	-	-	b1942
COG3121	SFMC	-	-	-	b0531
COG3143	CHEZ	-	-	-	b1881

Table A.2 | continued...

COG ID	Name	*T. pallidum*	*C. jejuni*	*H. pylori*	*E. coli*
COG3144	FLIK	-	-	-	b1943
COG3418	FLGN	-	-	-	b1070
COG3951	FLGJ	TP0959	-	-	-
COG4787	FLGF	-	-	-	b1077
COG5295	FLGE	-	-	HP0870	-
NOG06008	FLIZ	-	-	-	b1921
NOG07455	FLHD	-	-	-	b1892
NOG14615	-	TP0567	-	-	-
-	-	TP0403	-	-	-

Table A.3 | Literature interactions

PubMed ID	Name A	Name B	COG A	COG B	Method	Species
7578071	CheA	CheY	COG0643	COG0784	Biochemical	*E. coli*
8820640	CheA	CheZ	COG0643	COG3143	Biochemical	*E. coli*
377295	cheC	cheZ	COG1886	COG3143	Genetic screening	*E. coli*
7623663	CheY	CheZ	COG0784	COG3143	Genetic screening	*E. coli*
1400175	CheY	FliG	COG0784	COG1536	Genetic screening	*E. coli*
11135671	CheY	FliM	COG0784	COG1868	Structural	*E. coli*
15491362	FlbD	FliX	COG1582	NOG42184	Y2H	*C. crescentus*
11673434	FleN	FleQ	COG0455	COG2204	Biochemical, Y2H	*P. aeruginosa*
11554792	FlgB	FlgJ	COG1815	COG3951	Biochemical	*S. typhimurium*
10320579	FlgK	FlgN	COG1256	COG3418	Biochemical	*E. coli*
10320579	FlgL	FlgN	COG1344	COG3418	Biochemical	*E. coli*
11160096	FlhA	FliF	COG1298	COG1766	Genetic screening	*S. typhimurium*
15516571	FlhA	FliH	COG1298	COG1317	Biochemical	*S. typhimurium*
12949107, 15516571	FlhA	FliJ	COG1298	COG2882	Biochemical	*S. typhimurium*
15516571	FlhA	FliO	COG1298	COG3190	Biochemical	*S. typhimurium*
15516571	FlhA	FliP	COG1298	COG1338	Biochemical	*S. typhimurium*
15516571	FlhA	FliQ	COG1298	COG1987	Biochemical	*S. typhimurium*
10940035	FlhB	FlgD	COG1377	COG1843	Biochemical	*S. typhimurium*
15757683	FlhB	FliK	COG1377	COG3144	unknown	*S. typhimurium*
11204784	FlhF	FlhF	COG1419	COG1419	Y2H	*X. oryzae*
9095196, 9765212	FliA	FlgM	COG1191	COG2747	Biochemical, Structural	*S. typhimurium*, undefined
10940035	FliC	FlhB	COG1344	COG1377	Biochemical	*S. typhimurium*
8986772	FliC	FliC	COG1344	COG1344	Structural	*S. typhimurium*
11327763, 12958592	FliC	FliS	COG1344	COG1516	Biochemical, Structural	*A. aeolicus*, *S. typhimurium*
11169117	FliD	FliT	COG1345	NOG08749	Biochemical	*S. typhimurium*
10809679	FliE	FlgB	COG1677	COG1815	Biochemical	*S. typhimurium*
1551848	FliE	FliE	COG1677	COG1677	Biochemical	*S. typhimurium*
8206846	FliF	FliM	COG1766	COG1868	Biochemical	*E. coli*
10809678, 15126479	FliG	FliF	COG1536	COG1766	Genetic screening	*E. coli*, *S. typhimurium*
8757288	FliG	FliG	COG1536	COG1536	Biochemical	*E. coli*
8631704	FliG	FliM	COG1536	COG1868	Y2H	*E. coli*
8757288	FliG	FliN	COG1536	COG1886	Biochemical	*E. coli*
10998179	FliH	FliH	COG1317	COG1317	Biochemical	*S. typhimurium*
12949107	FliH	FliJ	COG1317	COG2882	Biochemical	*S. typhimurium*
10350613	FliI	FlgE	COG1157	COG1749	Biochemical	*S. typhimurium*
15516571	FliI	FlhA	COG1157	COG1298	Biochemical	*S. typhimurium*
10350613	FliI	FliC	COG1157	COG1344	Biochemical	*S. typhimurium*
10998179	FliI	FliH	COG1157	COG1317	Biochemical	*S. typhimurium*
8757288	FliM	FliM	COG1868	COG1868	Biochemical	*E. coli*
9791106	FliM	FliN	COG1868	COG1886	Biochemical	*E. coli*
15101977	FliM	MotD	COG1868	COG3144	Biochemical	*S. meliloti*
8757288	FliN	FliN	COG1886	COG1886	Biochemical	*E. coli*
11327763	FliS	FliS	COG1516	COG1516	Biochemical	*S. typhimurium*
9878359	HAP2	HAP2	COG1345	COG1345	unknown	undefined
10440379	MotA	FliG	COG1291	COG1536	Structural	*T. maritima*
8757288	MotA	FliM	COG1291	COG1868	Biochemical	undefined
8627625	MotB	FliG	COG1360	COG1536	Genetic screening	*E. coli*
15101977	MotB	MotC	COG1360	NOG06999	Biochemical	*S. meliloti*
10783392	PomA	PomA	COG1291	COG1291	Biochemical	*V. alginolyticus*
10783392	PomA	PomB	COG1291	COG1360	Biochemical	*V. alginolyticus*
15968056	sigma(54)	HP0958	COG1508	COG1579	Genetic screening, Y2H	*H. pylori*

Table A.4 | Conserved interactions with Blast results

TPA A	TPA B	CJE ALL A	CJE ALL B	CJE HCF A	CJE HCF B	HPY A	HPY B	ECO SPK A	ECO SPK B	ECO SAI A	ECO SAI B	E-Value AA	E-Value BB	Identity AA	Identity BB	BEST HIT AA	BEST HIT BB
-	-	-	-	flaA	fliS	flaA	fliS	-	-	-	-	2.24E-125	2.32E-41	50.31%	59.06%	1	1
-	-	flaA	fliS	-	-	flaA	fliS	-	-	-	-	2.24E-125	2.32E-41	50.31%	59.06%	1	1
-	-	flaA	fliS	-	-	flaA	fliS	-	-	-	-	2.24E-125	2.32E-41	50.31%	59.06%	1	1
-	-	-	-	flaA	fliS	flaA	fliS	-	-	-	-	2.24E-125	2.32E-41	50.31%	59.06%	1	1
-	-	-	-	flaB	fliS	flaA	fliS	-	-	-	-	1.11E-124	2.32E-41	49.61%	59.06%	0	1
-	-	flaB	fliS	-	-	flaA	fliS	-	-	-	-	1.11E-124	2.32E-41	49.61%	59.06%	0	1
-	-	-	-	flaB	fliS	flaA	fliS	-	-	-	-	1.11E-124	2.32E-41	49.61%	59.06%	0	1
-	-	flaB	fliS	-	-	flaA	fliS	-	-	-	-	1.11E-124	2.32E-41	49.61%	59.06%	0	1
-	-	-	-	flaB	fliS	flaB	fliS	-	-	-	-	1.79E-106	2.32E-41	40.48%	59.06%	0	1
-	-	flaB	fliS	-	-	flaB	fliS	-	-	-	-	1.79E-106	2.32E-41	40.48%	59.06%	0	1
-	-	-	-	flaB	fliS	flaB	fliS	-	-	-	-	1.79E-106	2.32E-41	40.48%	59.06%	0	1
-	-	flaB	fliS	-	-	flaB	fliS	-	-	-	-	1.79E-106	2.32E-41	40.48%	59.06%	0	1
-	-	flaA	fliS	-	-	flaB	fliS	-	-	-	-	8.60E-101	2.32E-41	39.37%	59.06%	0	1
-	-	-	-	flaA	fliS	flaB	fliS	-	-	-	-	8.60E-101	2.32E-41	39.37%	59.06%	0	1
-	-	-	-	flaA	fliS	flaB	fliS	-	-	-	-	8.60E-101	2.32E-41	39.37%	59.06%	0	1
-	-	flaA	fliS	-	-	flaB	fliS	-	-	-	-	8.60E-101	2.32E-41	39.37%	59.06%	0	1
-	-	rplB	fliM	-	-	-	-	rplB	fliM	-	-	9.41E-82	4.28E-29	55.37%	26.13%	1	1
-	-	rplB	fliM	-	-	-	-	rplB	fliM	-	-	9.41E-82	4.28E-29	55.37%	26.13%	1	1
-	-	-	-	flaC	fliS	flaA	fliS	-	-	-	-	6.51E-11	2.32E-41	16.56%	59.06%	0	1
-	-	flaC	fliS	-	-	flaA	fliS	-	-	-	-	6.51E-11	2.32E-41	16.56%	59.06%	0	1
-	-	flaC	fliS	-	-	flaA	fliS	-	-	-	-	6.51E-11	2.32E-41	16.56%	59.06%	0	1
-	-	-	-	flaC	fliS	flaA	fliS	-	-	-	-	6.51E-11	2.32E-41	16.56%	59.06%	0	1
-	-	-	-	-	-	rpoBC	fliA	rpoB	fliA	-	-	0.00E+00	1.12E-26	32.50%	29.98%	1	1
-	-	-	-	-	-	rpoBC	fliA	rpoB	fliA	-	-	0.00E+00	1.12E-26	32.50%	29.98%	1	1
-	-	-	-	flaC	fliS	flaB	fliS	-	-	-	-	3.14E-13	2.32E-41	15.93%	59.06%	0	1
-	-	flaC	fliS	-	-	flaB	fliS	-	-	-	-	3.14E-13	2.32E-41	15.93%	59.06%	0	1
-	-	-	-	flaC	fliS	flaB	fliS	-	-	-	-	3.14E-13	2.32E-41	15.93%	59.06%	0	1
-	-	flaC	fliS	-	-	flaB	fliS	-	-	-	-	3.14E-13	2.32E-41	15.93%	59.06%	0	1
-	-	-	-	-	-	gltX	fliA	-	-	gltX	fliA	3.32E-52	1.12E-26	27.04%	29.98%	0	1
-	-	-	-	-	-	gltX	fliA	-	-	gltX	fliA	3.32E-52	1.12E-26	27.04%	29.98%	0	1
-	-	-	-	-	-	rpoBC	fliA	rpoC	fliA	-	-	0.00E+00	1.12E-26	26.63%	29.98%	0	1
-	-	-	-	-	-	rpoBC	fliA	rpoC	fliA	-	-	0.00E+00	1.12E-26	26.63%	29.98%	0	1
proS	flaB2	-	-	-	-	-	-	proS	fliC	-	-	5.28E-110	1.20E-25	39.73%	17.75%	1	0
proS	flaB2	-	-	-	-	-	-	-	-	proS	fliC	5.28E-110	1.20E-25	39.73%	17.75%	1	0
proS	flaB2	-	-	-	-	-	-	proS	fliC	-	-	5.28E-110	1.20E-25	39.73%	17.75%	1	0
proS	flaB2	-	-	-	-	-	-	-	-	proS	fliC	5.28E-110	1.20E-25	39.73%	17.75%	1	0
-	-	-	-	motB	Cj1648	motB	HP1464	-	-	-	-	4.52E-37	3.22E-14	42.07%	16.73%	1	1
-	-	motB	Cj1648	-	-	motB	HP1464	-	-	-	-	4.52E-37	3.22E-14	42.07%	16.73%	1	1
-	-	-	-	motB	Cj1648	motB	HP1464	-	-	-	-	4.52E-37	3.22E-14	42.07%	16.73%	1	1

Table A.4 | continued...

TPA A	TPA B	CJE ALL A	CJE ALL B	CJE HCF A	CJE HCF B	HPY A	HPY B	ECO SPK A	ECO SPK B	ECO SAI A	ECO SAI B	E-Value AA	E-Value BB	Identity AA	Identity BB	BEST HIT AA	BEST HIT BB
-	-	motB	Cj1648	-	-	motB	HP1464	-	-	-	-	4.52E-37	3.22E-14	42.07%	16.73%	1	1
flaB2	fliS	-	-	flaC	fliS	-	-	-	-	-	-	1.30E-14	8.30E-07	24.36%	26.88%	0	1
flaB2	fliS	flaC	fliS	-	-	-	-	-	-	-	-	1.30E-14	8.30E-07	24.36%	26.88%	0	1
flaB2	fliS	-	-	flaC	fliS	-	-	-	-	-	-	1.30E-14	8.30E-07	24.36%	26.88%	0	1
flaB2	fliS	flaC	fliS	-	-	-	-	-	-	-	-	1.30E-14	8.30E-07	24.36%	26.88%	0	1
flaB1	fliS	-	-	-	-	flaA	fliS	-	-	-	-	2.53E-25	5.10E-09	19.11%	28.56%	0	1
flaB1	fliS	-	-	-	-	flaA	fliS	-	-	-	-	2.53E-25	5.10E-09	19.11%	28.56%	0	1
flaB1	fliS	-	-	flaC	fliS	-	-	-	-	-	-	3.00E-11	8.30E-07	19.86%	26.88%	0	1
flaB1	fliS	flaC	fliS	-	-	-	-	-	-	-	-	3.00E-11	8.30E-07	19.86%	26.88%	0	1
flaB1	fliS	-	-	flaC	fliS	-	-	-	-	-	-	3.00E-11	8.30E-07	19.86%	26.88%	0	1
flaB1	fliS	flaC	fliS	-	-	-	-	-	-	-	-	3.00E-11	8.30E-07	19.86%	26.88%	0	1
flaB3	fliS	-	-	-	-	flaA	fliS	-	-	-	-	2.52E-25	5.10E-09	18.10%	28.56%	1	1
flaB3	fliS	-	-	-	-	flaA	fliS	-	-	-	-	2.52E-25	5.10E-09	18.10%	28.56%	1	1
flaB3	fliS	-	-	flaC	fliS	-	-	-	-	-	-	2.07E-12	8.30E-07	18.77%	26.88%	0	1
flaB3	fliS	flaC	fliS	-	-	-	-	-	-	-	-	2.07E-12	8.30E-07	18.77%	26.88%	0	1
flaB3	fliS	-	-	flaC	fliS	-	-	-	-	-	-	2.07E-12	8.30E-07	18.77%	26.88%	0	1
flaB3	fliS	flaC	fliS	-	-	-	-	-	-	-	-	2.07E-12	8.30E-07	18.77%	26.88%	0	1
flaB2	fliS	-	-	-	-	flaA	fliS	-	-	-	-	2.14E-24	5.10E-09	17.54%	28.56%	0	1
flaB2	fliS	-	-	-	-	flaA	fliS	-	-	-	-	2.14E-24	5.10E-09	17.54%	28.56%	0	1
flaB3	fliS	flaB	fliS	-	-	-	-	-	-	-	-	4.61E-25	8.30E-07	18.58%	26.88%	0	1
flaB3	fliS	-	-	flaB	fliS	-	-	-	-	-	-	4.61E-25	8.30E-07	18.58%	26.88%	0	1
flaB3	fliS	-	-	flaB	fliS	-	-	-	-	-	-	4.61E-25	8.30E-07	18.58%	26.88%	0	1
flaB3	fliS	flaB	fliS	-	-	-	-	-	-	-	-	4.61E-25	8.30E-07	18.58%	26.88%	0	1
flaB1	fliS	-	-	-	-	flaB	fliS	-	-	-	-	8.18E-24	5.10E-09	16.95%	28.56%	1	1
flaB1	fliS	-	-	-	-	flaB	fliS	-	-	-	-	8.18E-24	5.10E-09	16.95%	28.56%	1	1
flaB3	fliS	-	-	flaA	fliS	-	-	-	-	-	-	1.48E-23	8.30E-07	17.57%	26.88%	0	1
flaB3	fliS	flaA	fliS	-	-	-	-	-	-	-	-	1.48E-23	8.30E-07	17.57%	26.88%	0	1
flaB3	fliS	-	-	flaA	fliS	-	-	-	-	-	-	1.48E-23	8.30E-07	17.57%	26.88%	0	1
flaB3	fliS	flaA	fliS	-	-	-	-	-	-	-	-	1.48E-23	8.30E-07	17.57%	26.88%	0	1
flaB2	fliS	-	-	-	-	flaB	fliS	-	-	-	-	3.11E-23	5.10E-09	15.91%	28.56%	0	1
flaB2	fliS	-	-	-	-	flaB	fliS	-	-	-	-	3.11E-23	5.10E-09	15.91%	28.56%	0	1
flaB3	fliS	-	-	-	-	flaB	fliS	-	-	-	-	6.91E-23	5.10E-09	15.68%	28.56%	0	1
flaB3	fliS	-	-	-	-	flaB	fliS	-	-	-	-	6.91E-23	5.10E-09	15.68%	28.56%	0	1
flaB1	fliS	flaB	fliS	-	-	-	-	-	-	-	-	2.29E-24	8.30E-07	14.59%	26.88%	0	1
flaB1	fliS	-	-	flaB	fliS	-	-	-	-	-	-	2.29E-24	8.30E-07	14.59%	26.88%	0	1
flaB1	fliS	-	-	flaB	fliS	-	-	-	-	-	-	2.29E-24	8.30E-07	14.59%	26.88%	0	1
flaB1	fliS	flaB	fliS	-	-	-	-	-	-	-	-	2.29E-24	8.30E-07	14.59%	26.88%	0	1
flaB2	fliS	flaB	fliS	-	-	-	-	-	-	-	-	5.10E-24	8.30E-07	14.34%	26.88%	0	1
flaB2	fliS	flaB	fliS	-	-	-	-	-	-	-	-	5.10E-24	8.30E-07	14.34%	26.88%	0	1
flaB2	fliS	-	-	flaB	fliS	-	-	-	-	-	-	5.10E-24	8.30E-07	14.34%	26.88%	0	1

Table A.4 | continued...

TPA A	TPA B	CJE ALL A	CJE ALL B	CJE HCF A	CJE HCF B	HPY A	HPY B	ECO SPK A	ECO SPK B	ECO SAI A	ECO SAI B	E-Value AA	E-Value BB	Identity AA	Identity BB	BEST HIT AA	BEST HIT BB
flaB2	fliS	-	-	flaB	fliS	-	-	-	-	-	-	5.10E-24	8.30E-07	14.34%	26.88%	0	1
flaB1	TP0658	-	-	-	-	flaA	HP1377	-	-	-	-	2.53E-25	1.04E-05	19.11%	19.60%	0	1
flaB1	TP0658	-	-	-	-	flaA	HP1377	-	-	-	-	2.53E-25	1.04E-05	19.11%	19.60%	0	1
flaB1	fliS	-	-	flaA	fliS	-	-	-	-	-	-	1.48E-23	8.30E-07	13.71%	26.88%	0	1
flaB1	fliS	flaA	fliS	-	-	-	-	-	-	-	-	1.48E-23	8.30E-07	13.71%	26.88%	0	1
flaB1	fliS	flaA	fliS	-	-	-	-	-	-	-	-	1.48E-23	8.30E-07	13.71%	26.88%	0	1
flaB1	fliS	-	-	flaA	fliS	-	-	-	-	-	-	1.48E-23	8.30E-07	13.71%	26.88%	0	1
flaB2	fliS	flaA	fliS	-	-	-	-	-	-	-	-	5.10E-24	8.30E-07	13.59%	26.88%	1	1
flaB2	fliS	-	-	flaA	fliS	-	-	-	-	-	-	5.10E-24	8.30E-07	13.59%	26.88%	1	1
flaB2	fliS	flaA	fliS	-	-	-	-	-	-	-	-	5.10E-24	8.30E-07	13.59%	26.88%	1	1
flaB2	fliS	-	-	flaA	fliS	-	-	-	-	-	-	5.10E-24	8.30E-07	13.59%	26.88%	1	1
flaB1	TP0658	-	-	-	-	flaA	HP1154	-	-	-	-	2.53E-25	1.55E-03	19.11%	18.97%	0	0
flaB1	TP0658	-	-	-	-	flaA	HP1154	-	-	-	-	2.53E-25	1.55E-03	19.11%	18.97%	0	0
nrdB	flaB3	-	-	-	-	-	-	-	-	nrdF	flgL	1.37E-21	2.50E-07	23.61%	15.07%	1	0
nrdB	flaB3	-	-	-	-	-	-	-	-	nrdF	flgL	1.37E-21	2.50E-07	23.61%	15.07%	1	0
flaB3	TP0658	-	-	-	-	flaA	HP1377	-	-	-	-	2.52E-25	1.04E-05	18.10%	19.60%	1	1
flaB3	TP0658	-	-	-	-	flaA	HP1377	-	-	-	-	2.52E-25	1.04E-05	18.10%	19.60%	1	1
flaB2	TP0658	-	-	-	-	flaA	HP1377	-	-	-	-	2.14E-24	1.04E-05	17.54%	19.60%	0	1
flaB2	TP0658	-	-	-	-	flaA	HP1377	-	-	-	-	2.14E-24	1.04E-05	17.54%	19.60%	0	1
flaB3	TP0658	-	-	-	-	flaA	HP1154	-	-	-	-	2.52E-25	1.55E-03	18.10%	18.97%	1	0
flaB3	TP0658	-	-	-	-	flaA	HP1154	-	-	-	-	2.52E-25	1.55E-03	18.10%	18.97%	1	0
flaB2	TP0658	-	-	-	-	flaA	HP1154	-	-	-	-	2.14E-24	1.55E-03	17.54%	18.97%	0	0
flaB2	TP0658	-	-	-	-	flaA	HP1154	-	-	-	-	2.14E-24	1.55E-03	17.54%	18.97%	0	0
-	-	-	-	-	-	cheW	tlpA	cheW	tsr	-	-	2.19E-16	1.51E-10	25.30%	11.97%	1	1
-	-	-	-	-	-	cheW	tlpA	-	-	cheW	tsr	2.19E-16	1.51E-10	25.30%	11.97%	1	1
-	-	-	-	-	-	cheW	tlpA	cheW	tsr	-	-	2.19E-16	1.51E-10	25.30%	11.97%	1	1
-	-	-	-	-	-	cheW	tlpA	-	-	cheW	tsr	2.19E-16	1.51E-10	25.30%	11.97%	1	1
-	-	xerD	flgG2	-	-	-	-	-	-	intC	flgG	1.77E-04	5.71E-20	10.80%	27.74%	0	0
-	-	xerD	flgG2	-	-	-	-	intC	flgG	-	-	1.77E-04	5.71E-20	10.80%	27.74%	0	0
-	-	xerD	flgG2	-	-	-	-	-	-	intC	flgG	1.77E-04	5.71E-20	10.80%	27.74%	0	0
-	-	xerD	flgG2	-	-	-	-	intC	flgG	-	-	1.77E-04	5.71E-20	10.80%	27.74%	0	0
-	-	-	-	-	-	cheW	tlpA	cheW	tar	-	-	2.19E-16	2.76E-08	25.30%	11.36%	1	0
-	-	-	-	-	-	cheW	tlpA	cheW	tar	-	-	2.19E-16	2.76E-08	25.30%	11.36%	1	0
flaB3	TP0981	-	-	-	-	-	-	-	-	fliC	b1490	4.10E-30	1.30E-19	20.70%	12.56%	1	0
flaB3	TP0981	-	-	-	-	-	-	-	-	fliC	b1490	4.10E-30	1.30E-19	20.70%	12.56%	1	0
proS	flgL	-	-	-	-	-	-	-	-	proS	fliC	5.28E-110	1.20E-06	39.73%	6.47%	1	0
proS	flgL	-	-	-	-	-	-	proS	fliC	-	-	5.28E-110	1.20E-06	39.73%	6.47%	1	0
proS	flgL	-	-	-	-	-	-	-	-	proS	fliC	5.28E-110	1.20E-06	39.73%	6.47%	1	0
proS	flgL	-	-	-	-	-	-	proS	fliC	-	-	5.28E-110	1.20E-06	39.73%	6.47%	1	0
-	-	nuoC	fliM	-	-	-	-	nuoC	fliM	-	-	1.10E-16	4.28E-29	8.79%	26.13%	0	1

Table A.4 | continued...

TPA A	TPA B	CJE ALL A	CJE ALL B	CJE HCF A	CJE HCF B	HPY A	HPY B	ECO SPK A	ECO SPK B	ECO SAI A	ECO SAI B	E-Value AA	E-Value BB	Identity AA	Identity BB	BEST HIT AA	BEST HIT BB
-	-	nuoC	fliM	-	-	-	-	-	-	nuoC	fliM	1.10E-16	4.28E-29	8.79%	26.13%	0	1
-	-	nuoC	fliM	-	-	-	-	nuoC	fliM	-	-	1.10E-16	4.28E-29	8.79%	26.13%	0	1
-	-	nuoC	fliM	-	-	-	-	-	-	nuoC	fliM	1.10E-16	4.28E-29	8.79%	26.13%	0	1
mcp2-3	flaB3	-	-	-	-	-	-	-	-	aer	fliC	1.38E-11	4.10E-30	10.93%	20.70%	0	1
mcp2-3	flaB3	-	-	-	-	-	-	-	-	aer	fliC	1.38E-11	4.10E-30	10.93%	20.70%	0	1
TP0100	flgD	Cj0864	flgD	-	-	-	-	-	-	-	-		2.00E-09		15.78%		1
-	-	-	-	fliL	Cj0579c	-	-	fliL	ybeC	-	-	1.10E-03		18.00%		1	
-	-	fliL	Cj0579c	-	-	-	-	fliL	ybeC	-	-	1.10E-03		18.00%		1	
-	-	groEL	fliH	-	-	-	-	mopA	fliH	-	-	2.86E-166		60.38%		1	
-	-	-	-	-	-	cag26	flaA	-	-	aer	fliC		4.45E-33		27.38%		1
TP0939	flgB	-	-	-	-	HP0589	flgB	-	-	-	-		1.85E-06		28.26%		1
-	-	-	-	-	-	cag26	flaA	-	-	aer	fliC		4.45E-33		27.38%		1
-	-	fliL	Cj0579c	-	-	-	-	fliL	ybeC	-	-	1.10E-03		18.00%		1	
TP0100	flgD	Cj0864	flgD	-	-	-	-	-	-	-	-		2.00E-09		15.78%		1
mcp2-3	flaB3	-	-	-	-	cag26	flaA	-	-	-	-		2.52E-25		18.10%		1
mcp2-3	flaB3	-	-	-	-	cag26	flaA	-	-	-	-		2.52E-25		18.10%		1
TP0939	flgB	-	-	-	-	HP0589	flgB	-	-	-	-		1.85E-06		28.26%		1
-	-	-	-	fliL	Cj0579c	-	-	fliL	ybeC	-	-	1.10E-03		18.00%		1	
-	-	groEL	fliH	-	-	-	-	mopA	fliH	-	-	2.86E-166		60.38%		1	

Table A.5 | Interacting proteins with phenotype

COG	ECO MUT	BSU MUT	TPA	CJE ALL	CJE HCF	HPY	ECO SPK	ECO SAI
COG0477	ydeF	ydjK	-	Cj0339	Cj0339	-	-	-
COG0477	-	ydeG	-	Cj0461c	-	-	-	-
COG0477	-	ybfB	-	Cj0987c	-	-	-	-
COG0477	-	ycnB	-	Cj0080	-	-	-	-
COG0513	deaD	yfmL	-	-	-	-	-	deaD
COG0642	cpxA	ybdK	-	Cj0793	Cj0793	-	-	-
COG0642	rcsC	-	-	Cj0254	-	-	-	-
COG0642	-	-	-	Cj1222c	-	-	-	-
COG0642	-	-	-	Cj1492c	-	-	-	-
COG0643	cheA	cheA	-	cheA	-	cheA	cheA	cheA
COG0784	cheY	cheY	-	cheV	cheV	cheY	cheY	cheY
COG0784	rcsC	cheV	-	cheA	-	cheA	-	-
COG0784	-	-	-	-	-	cheV	-	-
COG0835	cheW	cheW	-	cheV	cheV	cheW	cheW	cheW
COG0835	-	-	-	-	-	cheV	-	-
COG0840	tsr	mcpA	mcp2-3	Cj0700	Cj1190c	cag26	tsr	tap
COG0840	tap	mcpC	-	Cj0092	-	tlpB	tap	tsr
COG0840	-	mcpB	-	Cj0202c	-	tlpA	-	-
COG0840	-	tlpB	-	Cj1190c	-	-	-	-
COG0840	-	tlpA	-	Cj0246c	-	-	-	-
COG1157	fliI	fliI	fliI	-	-	fliI	fliI	fliI
COG1191	fliA	sigD	TP0709	fliA	fliA	fliA	fliA	fliA
COG1256	flgK	flgK	flgK	flgK	flgK	flgK	flgK	flgK
COG1291	motA	motA	motA	motA	motA	-	motA	motA
COG1298	flhA	flhA	flhA	flhA	-	flhA	flhA	flhA
COG1317	fliH	fliH	fliH	fliH	fliH	fliH	fliH	fliH
COG1344	fliC	hag	flgL	flaC	flaC	flaB	flgL	fliC
COG1344	flgL	flgL	flaB2	flaA	flaA	flaA	fliC	flgL
COG1344	-	yvzB	flaB1	flaB	flaB	fla	-	-
COG1344	-	-	flaB3	-	-	-	-	-
COG1345	fliD	fliD	-	fliD	fliD	fliD	fliD	fliD
COG1352	cheR	cheR	cheR	cheR	-	-	cheR	cheR
COG1360	motB	motB	-	motB	motB	motB	motB	motB
COG1377	flhB	flhB	flhB	flhB	flhB	flhB	-	-
COG1396	nadR	ydcN	-	-	-	-	-	-
COG1516	fliS	fliS	fliS	fliS	fliS	fliS	fliS	fliS
COG1536	fliG	fliG	fliG-2	fliG	fliG	fliG	fliG	fliG
COG1536	-	-	fliG-1	-	-	-	-	-
COG1558	flgC	flgC	flgC	flgC	flgC	-	flgC	flgC
COG1684	fliR	fliR	fliR	fliR	fliR	-	-	-
COG1705	flgJ	yubE	-	-	-	-	flgJ	flgJ
COG1766	fliF	fliF	fliF	fliF	-	fliF	fliF	fliF
COG1815	flgB	flgB	flgB	flgB	flgB	flgB	flgB	flgB
COG1843	flgD	ylxG	flgD	flgD	flgD	-	-	-
COG1868	fliM	fliM	fliM	fliM	fliM	-	fliM	fliM
COG1886	fliN	fliY	fliY	fliY	fliY	fliN	-	-
COG1886	-	-	-	fliN	fliN	-	-	-
COG1987	fliQ	fliQ	fliQ	fliQ	-	-	fliQ	fliQ
COG2201	cheB	cheB	-	-	-	-	cheB	cheB

Table A.5 | continued...

COG	ECO MUT	BSU MUT	TPA	CJE ALL	CJE HCF	HPY	ECO SPK	ECO SAI
COG2747	flgM	flgM	-	-	-	-	flgM	flgM
COG2882	fliJ	fliJ	-	-	-	-	fliJ	fliJ
COG3144	fliK	fliK	-	-	-	-	fliK	fliK
COG4786	flgG	flhO	flgG-2	flgG2	flgG2	flgG	flgG	flgG
COG4786	-	flgE	-	flgG	flgG	-	-	-
COG0030	ksgA	-	-	ksgA	-	-	-	-
COG0036	rpe	-	cfxE	rep	-	-	-	-
COG0055	ygbF	-	-	-	-	-	atpD	atpD
COG0055	yhiF	-	-	-	-	-	-	-
COG0055	atpD	-	-	-	-	-	-	-
COG0082	aroC	-	-	aroC	-	aroC	-	-
COG0112	glyA	-	-	glyA	-	-	-	-
COG0158	fbp	-	-	fbp	-	-	-	-
COG0226	pstS	-	-	pstS	-	-	-	-
COG0250	rfaH	-	nusG	-	-	-	-	-
COG0254	rpmE	-	-	-	-	-	rpmE	-
COG0257	rpmJ	-	rpmJ-1	-	-	-	-	-
COG0265	htrA	-	htrA-1	-	-	-	-	-
COG0279	gmhA	-	-	gmhA	-	-	-	-
COG0343	tgt	-	-	tgt	-	-	-	-
COG0354	ygfZ	-	-	-	-	-	-	ygfZ
COG0399	wecE	-	-	wlaK	wlaK	-	-	-
COG0443	dnaK	-	-	-	-	HP0149	dnaK	-
COG0451	rfaD	-	-	Cj1427c	-	-	-	-
COG0451	-	-	-	fcl	-	-	-	-
COG0468	recA	-	-	recA	-	-	recA	recA
COG0468	-	-	-	Cj1009c	-	-	-	-
COG0484	dnaJ	-	-	-	-	-	dnaJ	-
COG0582	fimE	-	-	xerD	-	-	-	-
COG0583	ydhB	-	-	-	-	-	-	-
COG0691	smpB	-	-	-	-	smpB	-	-
COG0745	arcA	-	-	Cj1223c	-	-	-	arcA
COG0809	queA	-	-	queA	-	-	-	-
COG0834	yhdW	-	-	peb1A	peb1A	omp28	-	-
COG0834	-	-	-	hisJ	hisJ	-	-	-
COG0848	tolR	-	-	exbD3	-	-	-	-
COG0848	-	-	-	exbD1	-	-	-	-
COG0848	-	-	-	exbD2	-	-	-	-
COG0859	rfaF	-	-	waaF	waaF	-	-	-
COG0859	-	-	-	waaC	-	-	-	-
COG1076	yfhE	-	-	Cj0954c	Cj0954c	-	-	-
COG1261	flgA	-	-	-	-	-	flgA	flgA
COG1294	cydB	-	-	cydB	cydB	-	-	-
COG1508	rpoN	-	-	rpoN	-	-	-	-
COG1539	ygiG	-	-	-	-	-	-	-
COG1706	flgI	-	-	flgI	flgI	flgI	-	-
COG1749	flgE	-	flgE	flgE2	-	flgE	flgE	-
COG1749	-	-	-	-	-	flgE	-	-
COG1826	b3838	-	-	Cj0579c	Cj0579c	-	-	-

Table A.5 | continued...

COG	ECO MUT	BSU MUT	TPA	CJE ALL	CJE HCF	HPY	ECO SPK	ECO SAI
COG1923	hfq	-	-	-	-	-	hfq	hfq
COG2009	sdhC	-	-	frdC	-	-	-	-
COG2186	fadR	-	-	-	-	-	-	-
COG2194	yjgX	-	-	Cj0256	Cj0256	-	-	-
COG2200	yhjH	-	-	-	-	-	-	-
COG2771	yhiF	-	-	-	-	-	-	-
COG2916	hns	-	-	-	-	-	-	hns
COG2956	yciM	-	-	-	-	HP0660	-	-
COG3112	yacL	-	-	-	-	-	-	yacL
COG3143	cheZ	-	-	-	-	-	cheZ	cheZ
COG3391	b1452	-	TP0421	-	-	-	-	-
COG3417	ycfM	-	-	Cj0091	Cj0091	-	-	-
COG3418	flgN	-	-	-	-	-	flgN	flgN
COG3951	flgJ	-	TP0959	-	-	-	flgJ	flgJ
COG4787	flgF	-	-	-	-	-	flgF	flgF
NOG07455	flhD	-	-	-	-	-	flhD	flhD
NOG14307	yqeJ	-	-	-	-	-	-	yqeJ
COG0315	-	ydiG	-	moaC	-	-	-	-
COG0346	-	ydfO	-	Cj1301	-	-	-	-
COG0455	-	ylxH	ylxH-1	Cj0063c	Cj0063c	ylxH	-	-
COG0457	-	rapG	TP0648	Cj0390	Cj0390	pflA	-	-
COG0457	-	-	-	Cj1034c	Cj1034c	HP1479	-	-
COG0457	-	-	-	Cj0055c	-	-	-	-
COG0457	-	-	-	Cj0497	-	-	-	-
COG0457	-	-	-	Cj1637c	-	-	-	-
COG0463	-	csbB	-	Cj1434c	waaV	-	wcaA	wcaA
COG0463	-	-	-	Cj1422c	-	-	-	-
COG0463	-	-	-	waaV	-	-	-	-
COG0463	-	-	-	Cj1135	-	-	-	-
COG0463	-	-	-	Cj1136	-	-	-	-
COG0500	-	ybaJ	-	Cj1426c	-	-	bioC	bioC
COG0500	-	-	-	bioC	-	-	yebH	-
COG0500	-	-	-	Cj1326	-	-	-	-
COG0500	-	-	-	Cj0976	-	-	-	-
COG0500	-	-	-	Cj1420c	-	-	-	-
COG0628	-	ydbI	-	amaA	-	-	-	-
COG0673	-	idh	-	Cj0504c	-	-	-	-
COG0726	-	yxkH	-	-	-	-	ycdR	ycdR
COG0726	-	-	-	-	-	-	yadE	yadE
COG0791	-	lytF	-	Cj1653c	Cj1653c	HP0087	-	ydhO
COG0791	-	-	-	-	-	-	-	yiiX
COG0842	-	yfiM	-	-	-	-	ybhS	ybhS
COG1012	-	gabD	-	-	-	HP0056	-	-
COG1087	-	galE	-	galE	-	-	-	-
COG1136	-	yclH	-	Cj1663	-	ftsE	-	-
COG1334	-	yvyC	-	flaG	flaG	flaG	-	-
COG1419	-	flhF	flhF	flhF	flhF	flhF	-	-
COG1475	-	yyaA	-	Cj0101	-	-	-	-
COG1512	-	ydjH	TP0561	-	-	-	-	-

Table A.5 | continued...

COG	ECO MUT	BSU MUT	TPA	CJE ALL	CJE HCF	HPY	ECO SPK	ECO SAI
COG1580	-	fliL	fliL	fliL	fliL	HP0809	fliL	fliL
COG1664	-	yhbF	TP0048	-	-	HP1542	-	-
COG1664	-	yhbE	-	-	-	-	-	-
COG1677	-	fliE	fliE	fliE	fliE	-	-	-
COG1699	-	yviF	TP0658	-	-	HP1154	-	-
COG1699	-	-	-	-	-	HP1377	-	-
COG1774	-	yaaT	TP0046	-	-	-	-	-
COG1776	-	fliY	fliY	-	-	-	-	-
COG1776	-	cheC	-	-	-	-	-	-
COG2001	-	yllB	TP0383	-	-	-	-	-
COG2213	-	mtlA	-	-	-	-	cmtA	cmtA
COG2814	-	ybcL	-	Cj1241	-	-	-	-
COG3190	-	fliZ	-	-	-	-	fliO	fliO
COG3334	-	ylxF	-	Cj1496c	-	-	-	-
COG4606	-	yclN	-	ceuB	-	-	-	-

Table A.6 | Aligned protein networks

Source		Node A			Node B		Blast Result	
Set A	Set B	Gene 1	Gene 2	Type	Gene 3	Gene 4	E Value A	E Value B
ECO SAI	CJE ALL	potG	livF	11	fliM	fliM	3.14E-13	4.28E-29
ECO SAI	CJE ALL	fliY	peb1A	12	livG	livF	3.67E-19	4.74E-22
ECO SAI	CJE ALL	fliM	fliM	11	nuoC	nuoC	4.28E-29	1.10E-16
ECO SAI	CJE ALL	tar	Cj1190c	12	tsr	Cj0246c	6.26E-07	9.44E-07
ECO SAI	CJE ALL	potG	kpsT	11	fliM	fliM	2.65E-09	4.28E-29
ECO SAI	CJE ALL	tar	Cj0246c	12	tsr	Cj1190c	1.72E-08	3.10E-06
ECO SAI	CJE ALL	potG	Cj1663	12	fliM	fliM	5.81E-29	4.28E-29
ECO SAI	CJE ALL	ispA	ispA	12	fliY	peb1A	9.57E-36	3.67E-19
ECO SAI	CJE ALL	fliY	peb1A	12	livG	Cj1663	3.67E-19	2.09E-23
ECO SAI	CJE ALL	tar	Cj1190c	10	tsr	Cj1190c	6.26E-07	3.10E-06
ECO SAI	CJE ALL	purB	purB	12	fliC	flaC	1.94E-20	3.74E-09
ECO SAI	CJE ALL	potG	Cj1538c	12	fliM	fliM	7.77E-17	4.28E-29
ECO SAI	CJE ALL	fliY	peb1A	12	livG	iamA	3.67E-19	2.95E-15
ECO SAI	CJE ALL	tar	Cj0246c	10	tsr	Cj0246c	1.72E-08	9.44E-07
ECO SAI	CJE ALL	potG	iamA	11	fliM	fliM	3.55E-20	4.28E-29
ECO SAI	CJE ALL	fliY	peb1A	12	b2865	Cj1215	3.67E-19	4.21E-16
ECO SAI	CJE ALL	ompA	pal	21	flgC	flgC	7.08E-12	4.92E-18
ECO SAI	CJE ALL	cheA	cheA	10	basR	cheA	2.06E-99	3.31E-08
ECO SAI	CJE ALL	potG	Cj1587c	12	fliM	fliM	2.85E-14	4.28E-29
ECO SPK	CJE ALL	flgB	flgB	21	cheY	cheV	1.50E-09	6.17E-11
ECO SPK	CJE ALL	fliM	fliM	11	nuoC	nuoC	4.28E-29	1.10E-16
ECO SPK	CJE ALL	fliM	fliM	11	rplB	rplB	4.28E-29	9.41E-82
ECO SPK	CJE ALL	fliM	fliM	12	mopA	groEL	4.28E-29	2.86E-166
ECO SPK	CJE ALL	fliY	peb1A	12	livG	Cj1663	3.67E-19	2.09E-23
ECO SPK	CJE ALL	fliY	peb1A	12	livG	livF	3.67E-19	4.74E-22
ECO SPK	CJE ALL	cheR	cheR	21	fliM	fliM	1.51E-19	4.28E-29
ECO SPK	CJE ALL	flgB	flgB	21	cheW	cheV	1.50E-09	5.34E-08

Table A.6 | continued...

Source		Node A			Node B		Blast Result	
Set A	Set B	Gene 1	Gene 2	Type	Gene 3	Gene 4	E Value A	E Value B
ECO SPK	CJE ALL	fliY	peb1A	12	livG	iamA	3.67E-19	2.95E-15
ECO SPK	CJE ALL	fliC	flaC	21	fliS	fliS	3.74E-09	1.63E-07
ECO SPK	CJE ALL	ispA	ispA	12	fliY	peb1A	9.57E-36	3.67E-19
ECO SPK	CJE ALL	fliC	flaB	21	fliS	fliS	6.64E-35	1.63E-07
ECO SPK	CJE ALL	fliM	fliM	21	aer	Cj1189c	4.28E-29	1.55E-24
ECO SPK	CJE ALL	minD	Cj0063c	12	cheW	cheV	1.79E-15	5.34E-08
ECO SPK	CJE ALL	fliC	flaA	21	fliS	fliS	1.06E-36	1.63E-07
HPY	CJE ALL	flaB	flaA	11	fliS	fliS	8.60E-101	2.32E-41
HPY	CJE ALL	ftsE	livF	12	flgE	flgG2	4.52E-09	2.96E-08
HPY	CJE ALL	flaB	flaC	11	fliS	fliS	3.14E-13	2.32E-41
HPY	CJE ALL	ftsE	iamA	12	flgE	flgG	3.32E-19	2.21E-24
HPY	CJE ALL	HP0809	fliL	21	motB	motB	3.95E-23	4.52E-37
HPY	CJE ALL	fliS	fliS	12	flhF	flhF	2.32E-41	2.29E-86
HPY	CJE ALL	flaB	flaB	11	fliS	fliS	1.79E-106	2.32E-41
HPY	CJE ALL	flaA	flaB	11	fliS	fliS	1.11E-124	2.32E-41
HPY	CJE ALL	aroC	aroC	21	flhA	flhA	3.87E-99	7.35E-163
HPY	CJE ALL	cheA	cheV	10	cheY	cheV	7.23E-14	7.45E-12
HPY	CJE ALL	ftsE	livF	12	flgE	flgG	4.52E-09	2.21E-24
HPY	CJE ALL	motB	motB	11	HP1464	Cj1648	4.52E-37	3.22E-14
HPY	CJE ALL	flaA	flaA	11	fliS	fliS	2.24E-125	2.32E-41
HPY	CJE ALL	cheW	cheV	10	cheA	cheV	2.61E-09	7.23E-14
HPY	CJE ALL	flaA	flaC	11	fliS	fliS	6.51E-11	2.32E-41
HPY	CJE ALL	ftsE	Cj1587c	11	flgE	flgG2	1.31E-08	2.96E-08
HPY	CJE ALL	cheA	cheA	10	cheY	cheA	0.00E+00	9.17E-15
HPY	CJE ALL	ftsE	iamA	12	flgE	flgG2	3.32E-19	2.96E-08
HPY	CJE ALL	ftsE	Cj1538c	11	flgE	flgG2	4.52E-14	2.96E-08
HPY	CJE ALL	HP0595	Cj0017c	11	flgE	flgG2	2.99E-84	2.96E-08
CJE HCF	ECO SAI	fliM	fliM	11	livF	potG	4.28E-29	3.14E-13
CJE HCF	ECO SAI	Cj1190c	tar	1	Cj1190c	tsr	6.26E-07	3.10E-06
CJE HCF	ECO SAI	fliM	fliM	11	iamA	potG	4.28E-29	3.55E-20
CJE HCF	ECO SPK	cheV	cheY	12	flgB	flgB	6.17E-11	1.50E-09
CJE HCF	ECO SPK	fliS	fliS	12	flaA	fliC	1.63E-07	1.06E-36
CJE HCF	ECO SPK	fliS	fliS	12	flaB	fliC	1.63E-07	6.64E-35
CJE HCF	ECO SPK	cheV	cheW	12	flgB	flgB	5.34E-08	1.50E-09
CJE HCF	ECO SPK	fliM	fliM	21	groEL	mopA	4.28E-29	2.86E-166
CJE HCF	ECO SPK	fliM	fliM	12	Cj1189c	aer	4.28E-29	1.55E-24
CJE HCF	ECO SPK	fliS	fliS	12	flaC	fliC	1.63E-07	3.74E-09
CJE HCF	ECO SPK	Cj0063c	minD	21	cheV	cheW	1.79E-15	5.34E-08
CJE HCF	HPY	cheV	cheA	1	cheV	cheY	7.23E-14	7.45E-12
CJE HCF	HPY	fliS	fliS	11	flaA	flaA	2.32E-41	2.24E-125
CJE HCF	HPY	motB	motB	12	fliL	HP0809	4.52E-37	3.95E-23
CJE HCF	HPY	fliS	fliS	11	flaB	flaB	2.32E-41	1.79E-106
CJE HCF	HPY	fliS	fliS	11	flaC	flaA	2.32E-41	6.51E-11
CJE HCF	HPY	fliS	fliS	11	flaB	flaA	2.32E-41	1.11E-124
CJE HCF	HPY	motB	motB	11	Cj1648	HP1464	4.52E-37	3.22E-14
CJE HCF	HPY	fliS	fliS	11	flaA	flaB	2.32E-41	8.60E-101
CJE HCF	HPY	fliS	fliS	11	flaC	flaB	2.32E-41	3.14E-13
CJE HCF	HPY	cheV	cheW	1	cheV	cheA	2.61E-09	7.23E-14
CJE HCF	HPY	flhF	flhF	21	fliS	fliS	2.29E-86	2.32E-41

Table A.6 | continued...

Source		Node A			Node B		Blast Result	
Set A	Set B	Gene 1	Gene 2	Type	Gene 3	Gene 4	E Value A	E Value B
HPY	ECO SAI	cheA	cheB	10	cheY	cheB	1.32E-07	1.05E-07
HPY	ECO SAI	fla	flgL	21	flgK	flgK	5.12E-09	3.97E-25
HPY	ECO SAI	gltX	gltX	11	fliA	fliA	3.32E-52	1.12E-26
HPY	ECO SAI	tlpA	tar	12	cheW	cheW	2.76E-08	2.19E-16
HPY	ECO SAI	cheA	arcA	10	cheY	arcA	8.43E-10	7.98E-11
HPY	ECO SAI	cheA	cheA	12	cheY	cheB	4.25E-88	1.05E-07
HPY	ECO SAI	cheA	phoB	12	cheY	uvrY	5.67E-11	7.28E-06
HPY	ECO SAI	cheA	cheA	1	cheA	basR	4.25E-88	2.77E-06
HPY	ECO SAI	cheA	phoB	10	cheY	phoB	5.67E-11	2.58E-14
HPY	ECO SAI	tlpA	tar	1	tlpA	tsr	2.76E-08	1.51E-10
HPY	ECO SAI	cheA	cheY	10	cheY	cheY	1.02E-09	1.14E-27
HPY	ECO SAI	tlpB	tsr	21	ylxH	minD	6.16E-07	6.20E-12
HPY	ECO SAI	tlpA	tsr	11	cheW	cheW	1.51E-10	2.19E-16
HPY	ECO SPK	cheA	cheB	10	cheY	cheB	1.32E-07	1.05E-07
HPY	ECO SPK	cheA	cheA	12	cheY	cheY	4.25E-88	1.14E-27
HPY	ECO SPK	tlpA	tsr	11	cheW	cheW	1.51E-10	2.19E-16
HPY	ECO SPK	cheW	cheW	12	cheA	cheY	2.19E-16	1.02E-09
HPY	ECO SPK	cheA	cheY	12	cheY	cheB	1.02E-09	1.05E-07
HPY	ECO SPK	fliA	fliA	11	rpoBC	rpoB	1.12E-26	0.00E+00
HPY	ECO SPK	flaB	fliC	12	fliS	fliS	2.07E-26	1.70E-09
HPY	ECO SPK	cheA	cheB	11	cheY	uvrY	1.32E-07	7.28E-06
HPY	ECO SPK	cheA	cheA	12	cheY	cheB	4.25E-88	1.05E-07
HPY	ECO SPK	flaA	fliC	12	fliS	fliS	4.45E-33	1.70E-09
HPY	ECO SPK	cheW	cheW	12	cheA	cheB	2.19E-16	1.32E-07
HPY	ECO SPK	cheY	cheB	1	cheY	uvrY	1.05E-07	7.28E-06
HPY	ECO SPK	cheA	phoB	12	cheY	uvrY	5.67E-11	7.28E-06
HPY	ECO SPK	cheW	cheW	12	cheA	cheA	2.19E-16	4.25E-88
HPY	ECO SPK	fliA	fliA	11	rpoBC	rpoC	1.12E-26	0.00E+00
HPY	ECO SPK	gltX	gltX	12	fliA	fliA	3.32E-52	1.12E-26
HPY	ECO SPK	tlpA	tar	11	cheW	cheW	2.76E-08	2.19E-16
HPY	ECO SPK	cheA	phoB	10	cheY	phoB	5.67E-11	2.58E-14
HPY	ECO SPK	cheA	cheB	12	cheY	cheY	1.32E-07	1.14E-27
HPY	ECO SPK	cheA	cheY	10	cheY	cheY	1.02E-09	1.14E-27
TPA	CJE ALL	flaB1	flaA	11	fliS	fliS	1.48E-23	8.30E-07
TPA	CJE ALL	fliY	fliY	21	fliM	fliM	1.87E-19	7.64E-57
TPA	CJE ALL	fliM	fliM	21	flgG-2	flgG	7.64E-57	4.46E-31
TPA	CJE ALL	fliG-1	fliG	12	fliY	fliN	1.36E-22	1.92E-12
TPA	CJE ALL	mcp2-3	Cj0246c	21	flgG-2	flgG2	1.66E-08	5.02E-30
TPA	CJE ALL	fliG-1	fliG	12	flgG-2	flgG	1.36E-22	4.46E-31
TPA	CJE ALL	flaB2	flaC	11	fliS	fliS	1.30E-14	8.30E-07
TPA	CJE ALL	fliG-1	fliG	12	fliY	fliY	1.36E-22	1.87E-19
TPA	CJE ALL	flaB2	flaA	11	fliS	fliS	5.10E-24	8.30E-07
TPA	CJE ALL	mcp2-3	Cj1190c	21	flgG-2	flgG2	2.52E-09	5.02E-30
TPA	CJE ALL	fliY	fliY	1	fliY	fliN	1.87E-19	1.92E-12
TPA	CJE ALL	flgC	flgC	12	fliY	fliN	1.59E-19	1.92E-12
TPA	CJE ALL	flgC	flgC	12	fliY	fliY	1.59E-19	1.87E-19
TPA	CJE ALL	flaB1	flaB	11	fliS	fliS	2.29E-24	8.30E-07
TPA	CJE ALL	flaB1	flaC	11	fliS	fliS	3.00E-11	8.30E-07
TPA	CJE ALL	fliG-1	fliG	12	fliM	fliM	1.36E-22	7.64E-57

Table A.6 | continued...

Source		Node A			Node B		Blast Result	
Set A	Set B	Gene 1	Gene 2	Type	Gene 3	Gene 4	E Value A	E Value B
TPA	CJE ALL	pyrG	pyrG	12	cheR	cheR	1.47E-124	1.27E-11
TPA	CJE ALL	flgE	flgG2	12	flgD	flgD	2.41E-06	2.00E-09
TPA	CJE ALL	fliG-1	fliG	10	fliG-2	fliG	1.36E-22	1.73E-57
TPA	CJE ALL	fliG-1	fliG	21	TP0100	Cj1207c	1.36E-22	9.52E-11
TPA	CJE ALL	flgK	flgK	12	fliY	fliY	1.48E-28	1.87E-19
TPA	CJE ALL	flaB3	flaA	11	fliS	fliS	1.48E-23	8.30E-07
TPA	CJE ALL	flaB3	flaB	11	fliS	fliS	4.61E-25	8.30E-07
TPA	CJE ALL	flaB3	flaC	11	fliS	fliS	2.07E-12	8.30E-07
TPA	CJE ALL	TP0100	trxA	21	flgK	flgK	7.79E-08	1.48E-28
TPA	CJE ALL	cheR	cheR	21	fliM	fliM	1.27E-11	7.64E-57
TPA	CJE ALL	flaB2	flaB	11	fliS	fliS	5.10E-24	8.30E-07
TPA	CJE ALL	nrdB	nrdB	12	flaB3	flaC	5.70E-69	2.07E-12
TPA	CJE HCF	fliY	fliY	21	fliM	fliM	1.87E-19	7.64E-57
TPA	CJE HCF	mcp2-3	Cj1190c	21	flgG-2	flgG2	2.52E-09	5.02E-30
TPA	CJE HCF	flaB2	flaB	11	fliS	fliS	5.10E-24	8.30E-07
TPA	CJE HCF	fliG-1	fliG	10	fliG-2	fliG	1.36E-22	1.73E-57
TPA	CJE HCF	flaB3	flaB	11	fliS	fliS	4.61E-25	8.30E-07
TPA	CJE HCF	flaB1	flaB	11	fliS	fliS	2.29E-24	8.30E-07
TPA	CJE HCF	flaB1	flaC	11	fliS	fliS	3.00E-11	8.30E-07
TPA	CJE HCF	flaB2	flaA	11	fliS	fliS	5.10E-24	8.30E-07
TPA	CJE HCF	flaB2	flaC	11	fliS	fliS	1.30E-14	8.30E-07
TPA	CJE HCF	flgC	flgC	12	fliY	fliN	1.59E-19	1.92E-12
TPA	CJE HCF	flaB3	flaC	11	fliS	fliS	2.07E-12	8.30E-07
TPA	CJE HCF	flgC	flgC	12	fliY	fliY	1.59E-19	1.87E-19
TPA	CJE HCF	flaB1	flaA	11	fliS	fliS	1.48E-23	8.30E-07
TPA	CJE HCF	flaB3	flaA	11	fliS	fliS	1.48E-23	8.30E-07
TPA	CJE HCF	fliY	fliY	1	fliY	fliN	1.87E-19	1.92E-12
TPA	CJE HCF	fliG-1	fliG	21	TP0100	Cj1207c	1.36E-22	9.52E-11
TPA	ECO SAI	flgK	flgK	21	flaB3	flgL	8.20E-27	2.50E-07
TPA	ECO SAI	nrdB	nrdF	11	flaB3	flgL	1.37E-21	2.50E-07
TPA	ECO SAI	fliI	atpD	11	mcp2-3	tsr	3.57E-43	2.81E-11
TPA	ECO SAI	mcp2-3	aer	21	flaB2	fliC	1.38E-11	1.20E-25
TPA	ECO SAI	proS	proS	11	flaB2	fliC	5.28E-110	1.20E-25
TPA	ECO SAI	fliG-1	fliG	10	fliG-2	fliG	5.22E-17	4.64E-45
TPA	ECO SAI	proS	proS	21	flaB3	fliC	5.28E-110	4.10E-30
TPA	ECO SAI	flgK	flgK	21	flaB1	flgL	8.20E-27	8.35E-07
TPA	ECO SAI	mcp2-3	tar	1	mcp2-3	tsr	2.23E-12	2.81E-11
TPA	ECO SAI	flaB3	fliC	11	TP0981	b1490	4.10E-30	1.30E-19
TPA	ECO SAI	proS	proS	11	flgL	fliC	5.28E-110	1.20E-06
TPA	ECO SAI	mcp2-3	aer	11	flaB3	fliC	1.38E-11	4.10E-30
TPA	ECO SAI	fliI	atpD	12	mcp2-3	tar	3.57E-43	2.23E-12
TPA	ECO SPK	flaB2	fliC	12	fliS	fliS	1.20E-25	7.69E-10
TPA	ECO SPK	fliG-1	fliG	10	fliG-2	fliG	5.22E-17	4.64E-45
TPA	ECO SPK	proS	proS	21	flaB3	fliC	5.28E-110	4.10E-30
TPA	ECO SPK	fliI	atpD	12	mcp2-3	tsr	3.57E-43	2.81E-11
TPA	ECO SPK	fliG-1	fliG	12	fliM	fliM	5.22E-17	6.45E-27
TPA	ECO SPK	fliG-1	fliG	12	flgG-2	flgG	5.22E-17	2.25E-33
TPA	ECO SPK	fliI	fliI	12	mcp2-3	aer	3.39E-98	1.38E-11
TPA	ECO SPK	flaB1	fliC	12	fliS	fliS	4.26E-27	7.69E-10

Table A.6 | continued...

Source		Node A			Node B		Blast Result	
Set A	Set B	Gene 1	Gene 2	Type	Gene 3	Gene 4	E Value A	E Value B
TPA	ECO SPK	flaB3	fliC	12	fliS	fliS	4.10E-30	7.69E-10
TPA	ECO SPK	proS	proS	11	flaB2	fliC	5.28E-110	1.20E-25
TPA	ECO SPK	nrdB	nrdF	12	flaB3	flgL	1.37E-21	2.50E-07
TPA	ECO SPK	fliG-1	fliG	12	flgG-2	flgF	5.22E-17	2.62E-14
TPA	ECO SPK	flgB	flgB	12	flaB3	fliC	2.47E-10	4.10E-30
TPA	ECO SPK	proS	proS	11	flgL	fliC	5.28E-110	1.20E-06
TPA	ECO SPK	cheR	cheR	12	flaB3	fliC	1.16E-29	4.10E-30
TPA	ECO SPK	fliG-1	fliG	12	flgG-2	flgE	5.22E-17	5.44E-07
TPA	ECO SPK	fliG-1	fliG	12	cheR	cheR	5.22E-17	1.16E-29
TPA	ECO SPK	fliI	atpD	12	mcp2-3	tar	3.57E-43	2.23E-12
TPA	ECO SPK	mcp2-3	aer	12	flaB3	fliC	1.38E-11	4.10E-30
TPA	HPY	ruvB	HP1026	21	flgB	flgB	3.57E-08	1.85E-06
TPA	HPY	TP0048	HP1542	21	flaB1	flaA	4.18E-06	2.53E-25
TPA	HPY	flaB1	flaA	11	fliS	fliS	2.53E-25	5.10E-09
TPA	HPY	fliG-1	fliG	10	fliG-2	fliG	3.82E-25	2.75E-63
TPA	HPY	flaB3	flaB	11	fliS	fliS	6.91E-23	5.10E-09
TPA	HPY	flaB2	flaA	11	fliS	fliS	2.14E-24	5.10E-09
TPA	HPY	flaB1	flaB	11	fliS	fliS	8.18E-24	5.10E-09
TPA	HPY	nrdB	nrdB	21	flgB	flgB	1.83E-67	1.85E-06
TPA	HPY	TP0048	HP1542	12	fliS	fliS	4.18E-06	5.10E-09
TPA	HPY	TP0048	HP1542	21	flaB3	flaA	4.18E-06	2.52E-25
TPA	HPY	flaB2	flaB	11	fliS	fliS	3.11E-23	5.10E-09
TPA	HPY	TP0048	HP1542	21	flaB2	flaA	4.18E-06	2.14E-24
TPA	HPY	flaB3	flaA	11	fliS	fliS	2.52E-25	5.10E-09

Table A.7 | A selection of predicted interactions

Species		COG A	COG B	SOURCE	SwissProt ID A	SwissProt ID B
Listeria F2365	*monocytogenes*	COG0008	COG1191	ECO, HPY	AAT03036	AAT03693
		COG0085	COG1191	ECO, HPY	AAT03061	AAT03693
		COG0086	COG1191	ECO, HPY	AAT03062	AAT03693
		COG0090	COG1868	CJE, ECO	AAT05367	AAT03516
		COG0208	COG1344	ECO, TPA	AAT04953	AAT03507
		COG0208	COG1344	ECO, TPA	AAT04953	AAT03523
		COG0442	COG1344	ECO, TPA	AAT04111	AAT03507
		COG0442	COG1344	ECO, TPA	AAT04111	AAT03523
		COG0643	COG0784	HPY, LIT	AAT03509	AAT03508
		COG0784	COG1536	LIT	AAT03508	AAT03531
		COG0784	COG1868	LIT	AAT03508	AAT03516
		COG0835	COG0840	ECO, HPY	AAT03506	AAT03540
		COG0835	COG0840	ECO, HPY	AAT03506	AAT04496
		COG0840	COG1344	ECO, TPA	AAT03540	AAT03507
		COG0840	COG1344	ECO, TPA	AAT03540	AAT03523
		COG0840	COG1344	ECO, TPA	AAT04496	AAT03507
		COG0840	COG1344	ECO, TPA	AAT04496	AAT03523

Table A.7 | continued...

Species	COG A	COG B	SOURCE	SwissProt ID A	SwissProt ID B
	COG1157	COG1298	LIT	AAT03533	AAT03497
	COG1157	COG1344	LIT	AAT03533	AAT03507
	COG1157	COG1344	LIT	AAT03533	AAT03523
	COG1157	COG1749	LIT	AAT03533	AAT03514
	COG1291	COG1291	LIT, TPA	AAT03502	AAT03502
	COG1291	COG1360	LIT	AAT03502	AAT03503
	COG1291	COG1536	LIT	AAT03502	AAT03531
	COG1291	COG1868	CJE, LIT	AAT03502	AAT03516
	COG1298	COG1338	LIT	AAT03497	AAT03493
	COG1298	COG1766	LIT	AAT03497	AAT03530
	COG1298	COG1987	LIT	AAT03497	AAT03494
	COG1344	COG1344	LIT	AAT03507	AAT03507
	COG1344	COG1344	LIT	AAT03507	AAT03523
	COG1344	COG1344	LIT	AAT03523	AAT03507
	COG1344	COG1344	LIT	AAT03523	AAT03523
	COG1344	COG1377	LIT	AAT03507	AAT03496
	COG1344	COG1377	LIT	AAT03523	AAT03496
	COG1344	COG2199	ECO, TPA	AAT03507	AAT04973
	COG1344	COG2199	ECO, TPA	AAT03507	AAT03342
	COG1344	COG2199	ECO, TPA	AAT03507	AAT04711
	COG1344	COG2199	ECO, TPA	AAT03507	AAT04710
	COG1344	COG2199	ECO, TPA	AAT03523	AAT04973
	COG1344	COG2199	ECO, TPA	AAT03523	AAT03342
	COG1344	COG2199	ECO, TPA	AAT03523	AAT04711
	COG1344	COG2199	ECO, TPA	AAT03523	AAT04710
	COG1345	COG1345	LIT	AAT03524	AAT03524
	COG1360	COG1536	LIT	AAT03503	AAT03531
	COG1377	COG1843	LIT	AAT03496	AAT03513
	COG1419	COG1419	CJE, LIT	AAT03498	AAT03498
	COG1516	COG1516	LIT, TPA	AAT03525	AAT03525
	COG1536	COG1536	LIT, TPA	AAT03531	AAT03531
	COG1536	COG1766	LIT, TPA	AAT03531	AAT03530
	COG1536	COG1868	LIT, TPA	AAT03531	AAT03516
	COG1536	COG1886	LIT, TPA	AAT03531	AAT03515
	COG1536	COG1886	LIT, TPA	AAT03531	AAT03510
	COG1677	COG1677	LIT	AAT03529	AAT03529
	COG1677	COG1815	LIT, TPA	AAT03529	AAT03527
	COG1766	COG1868	LIT	AAT03530	AAT03516
	COG1868	COG1868	LIT	AAT03516	AAT03516
	COG1868	COG1886	CJE, LIT	AAT03516	AAT03515
	COG1868	COG1886	CJE, LIT	AAT03516	AAT03510
	COG1886	COG1886	CJE, LIT	AAT03515	AAT03515
	COG1886	COG1886	CJE, LIT	AAT03515	AAT03510
	COG1886	COG1886	CJE, LIT	AAT03510	AAT03515
	COG1886	COG1886	CJE, LIT	AAT03510	AAT03510
Bacillus anthracis	COG0008	COG1191	ECO, HPY	SYE_BACAA	RP28_BACAA
	COG0008	COG1191	ECO, HPY	SYE_BACAA	Q81YQ5
	COG0008	COG1191	ECO, HPY	SYE_BACAA	Q81W67
	COG0008	COG1191	ECO, HPY	SYE_BACAA	Q81WD6
	COG0008	COG1191	ECO, HPY	SYE_BACAA	RPSB_BACAA

Table A.7 | continued...

Species	COG A	COG B	SOURCE	SwissProt ID A	SwissProt ID B
	COG0008	COG1191	ECO, HPY	SYE_BACAA	Q81MF5
	COG0008	COG1191	ECO, HPY	SYE_BACAA	RP35_BACAA
	COG0085	COG1191	ECO, HPY	RPOB_BACAA	RP28_BACAA
	COG0085	COG1191	ECO, HPY	RPOB_BACAA	Q81YQ5
	COG0085	COG1191	ECO, HPY	RPOB_BACAA	Q81W67
	COG0085	COG1191	ECO, HPY	RPOB_BACAA	Q81WD6
	COG0085	COG1191	ECO, HPY	RPOB_BACAA	RPSB_BACAA
	COG0085	COG1191	ECO, HPY	RPOB_BACAA	Q81MF5
	COG0085	COG1191	ECO, HPY	RPOB_BACAA	RP35_BACAA
	COG0086	COG1191	ECO, HPY	RPOC_BACAA	RP28_BACAA
	COG0086	COG1191	ECO, HPY	RPOC_BACAA	Q81YQ5
	COG0086	COG1191	ECO, HPY	RPOC_BACAA	Q81W67
	COG0086	COG1191	ECO, HPY	RPOC_BACAA	Q81WD6
	COG0086	COG1191	ECO, HPY	RPOC_BACAA	RPSB_BACAA
	COG0086	COG1191	ECO, HPY	RPOC_BACAA	Q81MF5
	COG0086	COG1191	ECO, HPY	RPOC_BACAA	RP35_BACAA
	COG0208	COG1344	ECO, TPA	Q81TB4	Q81SF2
	COG0442	COG1344	ECO, TPA	Q81WL6	Q81SF2
	COG0442	COG1344	ECO, TPA	Q81Z76	Q81SF2
	COG0784	COG1536	LIT	Q81SI8	Q81SH3
	COG0784	COG1536	LIT	Q81JW3	Q81SH3
	COG0840	COG1344	ECO, TPA	Q81RN3	Q81SF2
	COG0840	COG1344	ECO, TPA	Q81JN0	Q81SF2
	COG0840	COG1344	ECO, TPA	Q81Z93	Q81SF2
	COG0840	COG1344	ECO, TPA	Q81TX6	Q81SF2
	COG0840	COG1344	ECO, TPA	Q81XC3	Q81SF2
	COG0840	COG1344	ECO, TPA	Q81XI1	Q81SF2
	COG0840	COG1344	ECO, TPA	Q81VC2	Q81SF2
	COG0840	COG1344	ECO, TPA	Q81XF7	Q81SF2
	COG0840	COG1344	ECO, TPA	Q81ZA3	Q81SF2
	COG0840	COG1344	ECO, TPA	Q81V20	Q81SF2
	COG0840	COG1344	ECO, TPA	Q81YS4	Q81SF2
	COG0840	COG1344	ECO, TPA	Q81ZA2	Q81SF2
	COG0840	COG1344	ECO, TPA	Q81NB9	Q81SF2
	COG1157	COG1298	LIT	Q81SH1	Q81SE4
	COG1157	COG1344	LIT	Q81SH1	Q81SF2
	COG1157	COG1749	LIT	Q81SH1	Q81SG7
	COG1291	COG1291	LIT, TPA	Q81L81	Q81L81
	COG1291	COG1291	LIT, TPA	Q81L81	Q81SJ0
	COG1291	COG1291	LIT, TPA	Q81SJ0	Q81L81
	COG1291	COG1291	LIT, TPA	Q81SJ0	Q81SJ0
	COG1291	COG1360	LIT	Q81L81	Q81SI9
	COG1291	COG1360	LIT	Q81SJ0	Q81SI9
	COG1291	COG1536	LIT	Q81L81	Q81SH3
	COG1291	COG1536	LIT	Q81SJ0	Q81SH3
	COG1298	COG1338	LIT	Q81SE4	Q81SE8
	COG1298	COG1987	LIT	Q81SE4	Q81SE7
	COG1344	COG1344	LIT	Q81SF2	Q81SF2
	COG1344	COG1377	LIT	Q81SF2	Q81SE5
	COG1344	COG2199	ECO, TPA	Q81SF2	Q81JN9

Table A.7 | continued...

Species	COG A	COG B	SOURCE	SwissProt ID A	SwissProt ID B
	COG1345	COG1345	LIT	Q81SH9	Q81SH9
	COG1360	COG1536	LIT	Q81SI9	Q81SH3
	COG1377	COG1843	LIT	Q81SE5	Q81SG8
	COG1419	COG1419	CJE, LIT	Q81SE2	Q81SE2
	COG1516	COG1516	LIT, TPA	Q81SH8	Q81SH8
	COG1536	COG1536	LIT, TPA	Q81SH3	Q81SH3
	COG1536	COG1886	LIT, TPA	Q81SH3	Q81SF0
	COG1677	COG1677	LIT	Q81SH4	Q81SH4
	COG1677	COG1815	LIT, TPA	Q81SH4	Q81SH6
	COG1886	COG1886	CJE, LIT	Q81SF0	Q81SF0
Shigella flexneri 2a 2457T	COG0090	COG1868	CJE, ECO	RL2_SHIFL	Q7UAA4
	COG0208	COG1344	ECO, TPA	Q83QG9	FLIC_SHIFL
	COG0208	COG1344	ECO, TPA	Q7UC73	FLIC_SHIFL
	COG0442	COG1344	ECO, TPA	Q7UDQ4	FLIC_SHIFL
	COG0643	COG0784	HPY, LIT	Q7UAB5	CHEY_ECOLI
	COG0643	COG3143	LIT	Q7UAB5	Q7UAB8
	COG0784	COG1536	LIT	CHEY_ECOLI	Q7UAA6
	COG0784	COG1868	LIT	CHEY_ECOLI	Q7UAA4
	COG0784	COG3143	LIT	CHEY_ECOLI	Q7UAB8
	COG0835	COG0840	ECO, HPY	CHEW_ECOLI	Q7UAB6
	COG0835	COG0840	ECO, HPY	CHEW_ECOLI	Q83P14
	COG0835	COG0840	ECO, HPY	CHEW_ECOLI	Q7UAB7
	COG0835	COG0840	ECO, HPY	CHEW_ECOLI	Q83KT9
	COG0840	COG1344	ECO, TPA	Q7UAB6	FLIC_SHIFL
	COG0840	COG1344	ECO, TPA	Q83P14	FLIC_SHIFL
	COG0840	COG1344	ECO, TPA	Q7UAB7	FLIC_SHIFL
	COG0840	COG1344	ECO, TPA	Q83KT9	FLIC_SHIFL
	COG1157	COG1298	LIT	Q83R33	Q7UDL7
	COG1157	COG1317	HPY, LIT	Q83R33	Q7UAA5
	COG1157	COG1344	LIT	Q83R33	FLIC_SHIFL
	COG1157	COG1749	LIT	Q83R33	Q7UCX0
	COG1291	COG1291	LIT, TPA	Q83R49	Q83R49
	COG1291	COG1360	LIT	Q83R49	Q83KP3
	COG1291	COG1360	LIT	Q83R49	Q7UDL6
	COG1291	COG1536	LIT	Q83R49	Q7UAA6
	COG1291	COG1868	CJE, LIT	Q83R49	Q7UAA4
	COG1298	COG1317	LIT	Q7UDL7	Q7UAA5
	COG1298	COG1987	LIT	Q7UDL7	FLIQ_ECOLI
	COG1298	COG3190	LIT	Q7UDL7	Q83ML3
	COG1317	COG1317	HPY, LIT	Q7UAA5	Q7UAA5
	COG1344	COG1344	LIT	FLIC_SHIFL	FLIC_SHIFL
	COG1344	COG1377	LIT	FLIC_SHIFL	Q83KP8
	COG1344	COG2199	ECO, TPA	FLIC_SHIFL	Q7UCG6
	COG1344	COG2199	ECO, TPA	FLIC_SHIFL	Q7UCK8
	COG1344	COG3418	ECO, LIT	FLIC_SHIFL	Q83LI6
	COG1345	COG1345	LIT	Q83R43	Q83R43
	COG1345	NOG08749	LIT	Q83R43	Q83R41
	COG1360	COG1463	CJE, HPY	Q83KP3	Q9RHA1
	COG1360	COG1463	CJE, HPY	Q7UDL6	Q9RHA1
	COG1360	COG1536	LIT	Q83KP3	Q7UAA6

Table A.7 | continued...

Species	COG A	COG B	SOURCE	SwissProt ID A	SwissProt ID B
	COG1360	COG1536	LIT	Q7UDL6	Q7UAA6
	COG1377	COG1843	LIT	Q83KP8	Q7UCX1
	COG1377	COG3144	LIT	Q83KP8	Q83R31
	COG1516	COG1516	LIT, TPA	Q83R42	Q83R42
	COG1536	COG1536	LIT, TPA	Q7UAA6	Q7UAA6
	COG1536	COG1868	LIT, TPA	Q7UAA6	Q7UAA4
	COG1536	COG1886	LIT, TPA	Q7UAA6	Q83R29
	COG1677	COG1677	LIT	FLIE_SHIFL	FLIE_SHIFL
	COG1677	COG1815	LIT, TPA	FLIE_SHIFL	FLGB_ECOLI
	COG1815	COG3951	LIT	FLGB_ECOLI	Q7UCW9
	COG1868	COG1868	LIT	Q7UAA4	Q7UAA4
	COG1868	COG1886	CJE, LIT	Q7UAA4	Q83R29
	COG1868	COG3144	LIT	Q7UAA4	Q83R31
	COG1886	COG1886	CJE, LIT	Q83R29	Q83R29
	COG1886	COG3143	LIT	Q83R29	Q7UAB8

Bibliography

[1] T. Fenchel. Microbial behavior in a heterogeneous world. *Science*, 296(5570):1068–1071, 2002.

[2] A. Bren and M. Eisenbach. How signals are heard during bacterial chemotaxis: protein-protein interactions in sensory signal propagation. *J Bacteriol.*, 182:6865–73, 2000.

[3] J. P. Armitage. Bacterial tactic responses. *Adv Microb Physiol*, 41:229–289, 1999.

[4] Wadhams, H. George, Armitage, and P. Judith. Making sense of it all: bacterial chemotaxis. *Nat Rev Mol Cell Biol*, 5(12):1024–1037, 2004.

[5] N. R. Francis, G. E. Sosinsky, D. Thomas, and D. J. DeRosier. Isolation, characterization and structure of bacterial flagellar motors containing the switch complex. *J Mol Biol*, 235(4):1261–1270, 1994.

[6] M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*, 34(Database issue):D354–7, 2006.

[7] R. M. Macnab. How bacteria assemble flagella. *Annu Rev Microbiol*, 57:77–100, 2003.

[8] M. L. DePamphilis and J. Adler. Purification of intact flagella from *Escherichia coli* and *Bacillus subtilis*. *J Bacteriol*, 105(1):376–383, 1971.

[9] H. Sockett, S. Yamaguchi, M. Kihara, V. M. Irikura, and R. M. Macnab. Molecular analysis of the flagellar switch protein FliM of *Salmonella typhimurium*. *J Bacteriol*, 174(3):793–806, 1992.

[10] Rajagopala S.V. *The Protein-protein Interaction Map of the Treponema pallidum Flagellar Apparatus*. PhD thesis, Ruprecht - Karls - Universität Heidelberg, 2006.

[11] J. C. Rain, L. Selig, H. De Reuse, V. Battaglia, C. Reverdy, S. Simon, G. Lenzen, F. Petel, J. Wojcik, V. Schachter, Y. Chemama, A. Labigne, and P. Legrain. The protein-protein interaction map of *Helicobacter pylori*. *Nature*, 409(6817):211–215, 2001.

[12] Arifuzzaman M., Maeda M., Itoh A., Nishikata K., Takita C., Saito R., Ara T., Nakahigashi K., Huang H. C., Hirai A., Tsuzuki K., Nakamura S., Altaf-Ul-Amin M., Oshima T., Baba T., Yamamoto N., Kawamura T., Ioka-Nakamichi T., Kitagawa M., Tomita M., Kanaya S., Wada C., and Mori H. Large-scale identification of protein-protein interaction of *Escherichia coli* K-12. *Genome Res*, 2006.

[13] J. Lewis M. Raff K. Roberts B., A. Johnson and P. W. Alberts. *Molecular Biology of the Cell, 4th edition*. Garland/Taylor and Francis, 2002.

[14] G. D. Bader, D. Betel, and C. W. Hogue. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res*, 31(1):248–250, Jan 2003.

[15] C. Stark, B. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res*, 34(Database issue):535–539, 2006.

[16] H. Hermjakob, L. Montecchi-Palazzi, C. Lewington, S. Mudali, S. Kerrien, S. Orchard, M. Vingron, B. Roechert, P. Roepstorff, A. Valencia, H. Margalit, J. Armstrong, A. Bairoch, G. Cesareni, D. Sherman, and R. Apweiler. IntAct: an open source molecular interaction database. *Nucleic Acids Res*, 32(Database issue):D452–5, 2004.

[17] K. R. Brown and I. Jurisica. Online predicted human interaction database. *Bioinformatics*, 21(9):2076–82, 2005.

[18] A. Zanzoni, L. Montecchi-Palazzi, M. Quondam, Ga. Ausiello, M. Helmer-Citterich, and G. Cesareni. MINT: a Molecular INTeraction database. *FEBS Lett*, 513(1):135–140, 2002.

[19] H. Berman, K. Henrick, and H. Nakamura. Announcing the worldwide Protein Data Bank. *Nat Struct Biol*, 10(12):980, 2003.

[20] M. Fromont-Racine, J. C. Rain, and P. Legrain. Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. *Nat Genet*, 16(3):277–82, 1997.

[21] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin,

D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J. M. Rothberg. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, 403(6770):623–627, 2000.

[22] S. Fields and O. Song. A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230):245–6, Jul 20 1989.

[23] P. L. Bartel, J. A. Roecklein, D. SenGupta, and S. Fields. A protein linkage map of *Escherichia coli* bacteriophage T7. *Nat Genet*, 12(1):72–7, 1996.

[24] P. Bartel and S. Fields. *The yeast two-hybrid sytem*. Oxford University Press, 1997.

[25] M. Mann, R. C. Hendrickson, and A. Pandey. Analysis of proteins and proteomes by mass spectrometry. *Annu Rev Biochem*, 70:437–73, 2001.

[26] P. Uetz. Two-hybrid arrays. *Curr Opin Chem Biol*, 6(1):57–62, 2002.

[27] A. Kumar and M. Snyder. Protein complexes take the bait. *Nature*, 415(6868):123–124, 2002.

[28] M. Cornell, N.W. Paton, and S.G. Oliver. A critical and integrated view of the yeast interactome. *Comp Funct Genom*, 5:382–402, 2004.

[29] C. von Mering, L. J. Jensen, B. Snel, S. D. Hooper, M. Krupp, M. Foglierini, N. Jouffre, M. A. Huynen, and P. Bork. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res*, 33(Database issue):D433–7, 2005.

[30] A. C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. J. Jensen, S. Bastuck, B. Dumpelfeld, A. Edelmann, M. A. Heurtier, V. Hoffman, C. Hoefert, K. Klein, M. Hudak, A. M. Michon, M. Schelder, M. Schirle, M. Remor, T. Rudi, S. Hooper, A. Bauer, T. Bouwmeester, G. Casari, G. Drewes, G. Neubauer, J. M. Rick, B. Kuster, P. Bork, R. B. Russell, and G. Superti-Furga. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 2006.

[31] G. D. Bader and C. W. Hogue. Analyzing yeast protein-protein interaction data obtained from different sources. *Nat Biotechnol*, 20(10):991–7, 2002.

[32] P. Aloy and R. B. Russell. Interrogating protein interaction networks through structural biology. *Proc Natl Acad Sci U S A*, 99(9):5896–901, 2002.

[33] G. Butland, J. M. Peregrin-Alvarez, J. Li, W. Yang, X. Yang, V. Canadien, A. Starostine, D. Richards, B. Beattie, N. Krogan, M. Davey, J. Parkinson, J. Greenblatt, and A. Emili. Interaction network containing conserved and essential protein complexes in *Escherichia coli*. *Nature*, 433(7025):531–7, 2005.

[34] D. J. LaCount, M. Vignali, R. Chettier, A. Phansalkar, R. Bell, J. R. Hesselberth, L. W. Schoenfeld, I. Ota, S. Sahasrabudhe, C. Kurschner, S. Fields, and R. E. Hughes. A protein interaction network of the malaria parasite *Plasmodium falciparum*. *Nature*, 438(7064):103–7, 2005.

[35] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A*, 98(8):4569–74, 2001.

[36] A. C. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A. M. Michon, C. M. Cruciat, M. Remor, C. Hofert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M. A. Heurtier, R. R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. Superti-Furga. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–7, 2002.

[37] Y. Ho, A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, S. L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A. R. Willems, H. Sassi, P. A. Nielsen, K. J. Rasmussen, J. R. Andersen, L. E. Johansen, L. H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B. D. Sorensen, J. Matthiesen, R. C. Hendrickson, F. Gleeson, T. Pawson, M. F. Moran, D. Durocher, M. Mann, C. W. Hogue, D. Figeys, and M. Tyers. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, 415(6868):180–3, 2002.

[38] N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A. P. Tikuisis, T. Punna, J. M. Peregrin-Alvarez, M. Shales, X. Zhang, M. Davey, M. D. Robinson, A. Paccanaro, J. E. Bray, A. Sheung, B. Beattie, D. P. Richards, V. Canadien, A. Lalev, F. Mena, P. Wong, A. Starostine, M. M. Canete, J. Vlasblom, S. Wu, C. Orsi, S. R. Collins, S. Chandran, R. Haw, J. Rilstone, K. Gandi, N. J. Thompson, G. Musso, P. St Onge, S. Ghanny, M. H. Y. Lam, G. Butland, A. M Altaf-Ul, S. Kanaya, A. Shilatifard, E. O'Shea, J. S. Weissman, C. J. Ingles, T. R. Hughes, J. Parkinson, M. Gerstein, S. J. Wodak, A. Emili, and J. F. Greenblatt. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature*, 440(7084):637–643, 2006.

[39] S. Li, C. M. Armstrong, N. Bertin, H. Ge, S. Milstein, M. Boxem, P. O. Vidalain, J. D. Han, A. Chesneau, T. Hao, D. S. Goldberg, N. Li, M. Martinez, J. F. Rual,

P. Lamesch, L. Xu, M. Tewari, S. L. Wong, L. V. Zhang, G. F. Berriz, L. Jacotot, P. Vaglio, J. Reboul, T. Hirozane-Kishikawa, Q. Li, H. W. Gabel, A. Elewa, B. Baumgartner, D. J. Rose, H. Yu, S. Bosak, R. Sequerra, A. Fraser, S. E. Mango, W. M. Saxton, S. Strome, S. Van Den Heuvel, F. Piano, J. Vandenhaute, C. Sardet, M. Gerstein, L. Doucette-Stamm, K. C. Gunsalus, J. W. Harper, M. E. Cusick, F. P. Roth, D. E. Hill, and M. Vidal. A map of the interactome network of the metazoan *C. elegans*. *Science*, 303(5657):540–3, 2004.

[40] L. Giot, J. S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y. L. Hao, C. E. Ooi, B. Godwin, E. Vitols, G. Vijayadamodar, P. Pochart, H. Machineni, M. Welsh, Y. Kong, B. Zerhusen, R. Malcolm, Z. Varrone, A. Collis, M. Minto, S. Burgess, L. McDaniel, E. Stimpson, F. Spriggs, J. Williams, K. Neurath, N. Ioime, M. Agee, E. Voss, K. Furtak, R. Renzulli, N. Aanensen, S. Carrolla, E. Bickelhaupt, Y. Lazovatsky, A. DaSilva, J. Zhong, C. A. Stanyon, Jr. Finley R. L., K. P. White, M. Braverman, T. Jarvie, S. Gold, M. Leach, J. Knight, R. A. Shimkets, M. P. McKenna, J. Chant, and J. M. Rothberg. A protein interaction map of *Drosophila melanogaster*. *Science*, 302(5651):1727–36, 2003.

[41] E. Formstecher, S. Aresta, V. Collura, A. Hamburger, A. Meil, A. Trehin, C. Reverdy, V. Betin, S. Maire, C. Brun, B. Jacq, M. Arpin, Y. Bellaiche, S. Bellusci, P. Benaroch, M. Bornens, R. Chanet, P. Chavrier, O. Delattre, V. Doye, R. Fehon, G. Faye, T. Galli, J. A. Girault, B. Goud, J. de Gunzburg, L. Johannes, M. P. Junier, V. Mirouse, A. Mukherjee, D. Papadopoulo, F. Perez, A. Plessis, C. Rosse, S. Saule, D. Stoppa-Lyonnet, A. Vincent, M. White, P. Legrain, J. Wojcik, J. Camonis, and L. Daviet. Protein interaction mapping: a Drosophila case study. *Genome Res*, 15(3):376–84, 2005.

[42] J. F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D. S. Goldberg, L. V. Zhang, S. L. Wong, G. Franklin, S. Li, J. S. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R. S. Sikorski, J. Vandenhaute, H. Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M. E. Cusick, D. E. Hill, F. P. Roth, and M. Vidal. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–8, 2005.

[43] U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde, E. Toksoz, A. Droege, S. Krobitsch, B. Korn, W. Birchmeier, H. Lehrach, and E. E. Wanker. A human

protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122(6):957–68, 2005.

[44] P. Cramer, D. A. Bushnell, and R. D. Kornberg. Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. *Science*, 292(5523):1863–76, 2001.

[45] A. Flores, J. F. Briand, O. Gadal, J. C. Andrau, L. Rubbi, V. Van Mullem, C. Boschiero, M. Goussot, C. Marck, C. Carles, P. Thuriaux, A. Sentenac, and M. Werner. A protein-protein interaction map of yeast RNA polymerase III. *Proc Natl Acad Sci U S A*, 96(14):7815–20, 1999.

[46] A. M. Edwards, B. Kus, R. Jansen, D. Greenbaum, J. Greenblatt, and M. Gerstein. Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends Genet*, 18(10):529–36, 2002.

[47] H. W. Mewes, D. Frishman, K. F. Mayer, M. Munsterkotter, O. Noubibou, P. Pagel, T. Rattei, M. Oesterheld, A. Ruepp, and V. Stumpflen. MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Res*, 34(Database issue):D169–72, 2006.

[48] C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887):399–403, 2002.

[49] C. M. Deane, L. Salwinski, I. Xenarios, and D. Eisenberg. Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol Cell Proteomics*, 1(5):349–56, 2002.

[50] C. von Mering, M. Huynen, D. Jaeggi, S. Schmidt, P. Bork, and B. Snel. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res*, 31(1):258–61, 2003.

[51] E. M. Marcotte, M. Pellegrini, H. L. Ng, D. W. Rice, T. O. Yeates, and D. Eisenberg. Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285(5428):751–3, 1999.

[52] W. C. 3rd Lathe, B. Snel, and P. Bork. Gene context conservation of a higher order than operons. *Trends Biochem Sci*, 25(10):474–479, 2000.

[53] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A*, 96(8):4285–8, 1999.

[54] G. Cesareni, A. Ceol, C. Gavrila, L. M. Palazzi, M. Persico, and M. V. Schneider. Comparative interactomics. *FEBS Lett*, 579(8):1828–33, 2005.

[55] L. R. Matthews, P. Vaglio, J. Reboul, H. Ge, B. P. Davis, J. Garrels, S. Vincent, and M. Vidal. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Res*, 11(12):2120–6, 2001.

[56] R. L. Tatusov, E. V. Koonin, and D. J. Lipman. A genomic perspective on protein families. *Science*, 278(5338):631–7, 1997.

[57] M. A. Huynen, B. Snel, and V. van Noort. Comparative genomics for reliable protein-function prediction from genomic data. *Trends Genet*, 20(8):340–4, 2004.

[58] H. Yu, N. M. Luscombe, H. X. Lu, X. Zhu, Y. Xia, J. D. Han, N. Bertin, S. Chung, M. Vidal, and M. Gerstein. Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res*, 14(6):1107–18, 2004.

[59] B. P. Kelley, R. Sharan, R. M. Karp, T. Sittler, D. E. Root, B. R. Stockwell, and T. Ideker. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc Natl Acad Sci U S A*, 100(30):11394–11399, 2003.

[60] S. Suthram, T. Sittler, and T. Ideker. The plasmodium protein network diverges from those of other eukaryotes. *Nature*, 438(7064):108–12, 2005.

[61] Roded Sharan and Trey Ideker. Modeling cellular machinery through biological network comparison. *Nat Biotechnol*, 24(4):427–433, 2006.

[62] R. L. Tatusov, N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, and D. A. Natale. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4(1):41, 2003.

[63] K. D. Pruitt, T. Tatusova, and D. R. Maglott. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, 33(Database issue):501–504, 2005.

[64] J. L. Gardy, M. R. Laird, F. Chen, S. Rey, C. J. Walsh, M. Ester, and F. S. L. Brinkman. PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics*, 21(5):617–623, 2005. Evaluation Studies.

[65] A. Stein, R. B. Russell, and P. Aloy. 3did: interacting protein domains of known three-dimensional structure. *Nucleic Acids Res*, 33(Database issue):413–417, 2005.

[66] Xizeng Mao, Tao Cai, John G Olyarchuk, and Liping Wei. Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. *Bioinformatics*, 21(19):3787–3793, 2005. Evaluation Studies.

[67] W. Schumann. *Functional Analysis of Bacterial Genes (A practical Manual)*. John wiley and sons, LTD, 2001.

[68] K. van Amsterdam and A. van der Ende. *Helicobacter pylori* HP1034 (ylxH) is required for motility. *Helicobacter*, 9(5):387–395, 2004.

[69] B. Shepherd N. R. Salama and S. Falkow. Global transposon mutagenesis and essential gene analysis of *Helicobacter pylori*. *J Bacteriol*, 186(23):7926–7935, 2004.

[70] B. M. Pruss, J. W. Campbell, T. K. Van Dyk, C. Zhu, Y. Kogan, and P. Matsumura. FlhD/FlhC is a regulator of anaerobic respiration and the Entner-Doudoroff pathway through induction of the methyl-accepting chemotaxis protein Aer. *J Bacteriol*, 185(2):534–543, 2003.

[71] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3402, 1997.

[72] B. P. Kelley, B. Yuan, F. Lewitter, R. Sharan, B. R. Stockwell, and T. Ideker. PathBLAST: a tool for alignment of protein interaction networks. *Nucleic Acids Res*, 32(Web Server issue):W83–8, 2004.

[73] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11):2498–504, 2003.

[74] R. Chenna, H. Sugawara, T. Koike, R. Lopez, T. J. Gibson, D. G. Higgins, and J. D. Thompson. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res*, 31(13):3497–3500, 2003.

[75] J. Castresana. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol*, 17(4):540–552, 2000.

[76] J. Felsenstein. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*, 39:783–791, 1985.

[77] D. L. Swofford, editor. *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)*. Sinauer Associates: Sunderland, Massachusetts, 2003.

[78] J. Felsenstein. *PHYLIP (Phylogeny Inference Package) version 3.6.* 2005.

[79] H. A. Schmidt, K. Strimmer, M. Vingron, and A. von Haeseler. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, 18(3):502–4, 2002.

[80] O. R. P. Bininda-Emonds, J. L. Gittleman, and M. A. Steel. The (super) tree of life: procedures, problems, and prospects. *Ann. Rev. Ecol. Syst.*, 32:265–289, 2002.

[81] O. R. P. Bininda-Emonds. The evolution of supertrees. *Trends Ecol Evol*, 19(6):315–322, 2004.

[82] C. J. Creevey and J. O. McInerney. Clann: investigating phylogenetic information through supertree analyses. *Bioinformatics*, 21(3):390–392, 2005.

[83] B. R. Baum. Combining trees as a way of combining data sets for phylogenetic inference,and the desirability of combining gene trees. *Taxon*, 41:3–10, 1992.

[84] Müller J. and Müller K. TreeGraph: automated drawing of complex tree figures using an extensible tree description format. *Molecular Ecology Notes*, 4:786–788, 2004.

[85] A. Bateman, L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. L. Sonnhammer, D. J. Studholme, C. Yeats, and S. R. Eddy. The Pfam protein families database. *Nucleic Acids Res*, 32(Database issue):138–141, 2004.

[86] R. Belas and R. Suvanasuthi. The ability of *Proteus mirabilis* to sense surfaces and regulate virulence gene expression involves FliL, a flagellar basal body protein. *J Bacteriol*, 187(19):6789–6803, 2005.

[87] C Chang, A Mooser, A Pluckthun, and A Wlodawer. Crystal structure of the dimeric C-terminal domain of TonB reveals a novel fold. *J Biol Chem*, 276(29):27535–27540, 2001.

[88] G. J. Olsen, C. R. Woese, and R. Overbeek. The winds of (evolutionary) change: breathing new life into microbiology. *J Bacteriol*, 176(1):1–6, 1994.

[89] J. R. Brown, C. J. Douady, M. J. Italia, W. E. Marshall, and M. J. Stanhope. Universal trees based on large combined protein sequence data sets. *Nat Genet*, 28(3):281–285, 2001.