Biotechnology and Genetic Engineering Reviews - Vol. 24, 213-242 (2007)

Practical Applications of Bacterial Functional Genomics

GARRET SUEN, BRADLEY I. ARSHINOFF, RION G. TAYLOR, AND ROY D. WELCH*

Department of Biology, Syracuse University, Syracuse NY 13244, USA

Introduction

Microbial genome sequencing started in the late 1990s, and now, one decade later, we researchers have access to hundreds of genome sequences. This advance has revolutionized the way research is conducted, and microbial biology has solidly transitioned into the era of post-genomics. Researchers routinely have access to the full catalog of the genes within a genome, thereby eliminating any misperception that genes function in isolation, and thus facilitating the discovery and characterization of genes as parts of genetic networks. This paradigm shift, inspired by DNA sequencing, has led to the development of other high-throughput genomic techniques such as the DNA microarray. The major challenge of the post-genomic era is to interpret the overwhelming amount of data that is now available to all microbial life scientists. Functional genomics and systems biology seek to address this challenge by utilizing enormous genome-scale datasets to predict functional interactions between genes, both within a genetic network and within a genome. Many of the most advanced techniques and algorithms used to make these predictions require a large variety of genomic datasets that are currently only available to the most well-studied model organisms; for the majority of prokaryotic model organisms, the only available types of genomic data exist in the form of a genome sequence and a DNA microarray. Given this practical limitation, how can data be effectively incorporated into the

*To whom correspondence may be addressed (rowelch@syr.edu)

Abbreviations: 2D, two-dimensional; BLAST, Basic Local Alignment Search Tool; COG, clusters of orthologous groups; GO, gene ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; MOD, Model Organism Database; MS, mass spectrometry; NCBI, National Center for Biotechnological Information; ORF, open reading frame; Pfam, Protein families; PLEX, Protein Link Explorer; TR, transcriptional regulator; Y2H, yeast two-hybrid.

experimental framework of microbial molecular life science laboratories? The effective translation of genomic data into a practical experimental tool requires the end-user to understand the fundamental assumptions that frame each type of genomic data, as well as the quantitative limitations of each high-throughput protocol.

The vision of functional genomics is profound: all of the genes involved in a particular network, response, or phenotype can be rapidly identified, all but eliminating traditional "fishing expeditions" through mutagenesis screens. However, even with the availability of hundreds of genomes and a plethora of predictive algorithms, this vision is proving difficult to realize. As the field moves to increasingly complex methods of analysis, researchers continue to struggle with a lack of a consensus for many of the fundamental assumptions in functional genomics, and it is difficult to establish a generally accepted definition for function within this framework.

In this review, we will focus on the practical applications of bacterial functional genomics by describing how the most common types of prokaryotic genomic data can be utilized to construct an experimental pipeline. We will summarize the different forms of genomics datasets, with emphasis on how these data are used to make functional predictions. We also present a case study for the integration of datasets into an experimental pipeline for the identification and verification of functional interactions, including how these datasets can be applied to bacterial systems biology. Finally, we describe the relationship between functional genomics and the evolution of model organism databases, with emphasis on genome annotation. Throughout this review, we will provide examples, when appropriate, for the application of functional genomics to the prokaryote *Myxococcus xanthus*.

M. xanthus, is a soil-dwelling, gram-negative, δ-proteobacteria which can exist as a single-species biofilm and as free-living cells (Dworkin, 1993). Each bacterium is autonomous with respect to metabolism and reproduction, and the biofilm is a self-organizing predatory swarm that exhibits many characteristics of a multi-cellular organism (Kaiser, 1986). Under starvation conditions, *M. xanthus* undergoes a complex developmental cycle which culminates in the formation of fruiting bodies, spherical structures which contain environmentally stress-resistant myxospores (Shimkets, 1999; Kaiser, 2004). The highly complex lifecycle of *M. xanthus* is reflected in its 9.14 Mb genome (Goldman *et al.*, 2006), one of the largest prokaryotic genomes sequenced to date (Galperin, 2006), and analysis of this genome shows that *M. xanthus* has a highly developed set of signaling pathways. *M. xanthus* has also recently received considerable attention for its important practical applications in medicine and agriculture. *M. xanthus* produces hundreds of potentially valuable secondary metabolites (Reichenbach and Hofle, 1993; Reichenbach, 2001), including several anti-microbial drugs such as Myxalamid (Gerth *et al.*, 1983) and Myxovirescin A (Simunovic *et al.*, 2006). In addition, *M. xanthus* may also have application as a bacterial biological control agent to inhibit pathogenic fungi in plants (Bull *et al.*, 2002). The size and complexity of the *M. xanthus* genome make it an excellent example model organism for this review.

The rise of functional genomics

The genomics revolution began with the first reported complete genome sequence, that of the bacterial pathogen *Haemophilus influenza* in 1995 (Fleischmann *et al.,*

1995). For the first time, researchers were in possession of the unabridged catalogue of DNA required for an organism to survive in its ecological niche; within this sequence was the complete genetic blueprint for every response and phenotype. More than a decade later, we have hundreds of prokaryotic genome sequences, and the initial steps of analyses have been largely systematized. Suites of standard protocols are now applied to each new genome, and this information is made publicly accessible through GenBank (Benson *et al.*, 2006), the European Molecular Biology Laboratory (Cochrane *et al.*, 2006), and the DNA Data Bank of Japan (Okubo *et al.*, 2006). For any single model organism, this 'annotation pipeline' is largely independent of the scientists who conduct research on the organism. From a post-genomics perspective, this initial annotation is the starting point and it facilitates the development of high-throughput genomics technologies.

Functional genomics data can be subdivided into sequenced-based and experiment-based sets. Sequence-based datasets apply fundamental genetic principles to entire genomes through computer algorithms that rely, either directly or indirectly, on sequence homology and the overall structure and composition of the genome. For example, the principle of conserved operons (Overbeek *et al.*, 1999) can be used to predict the function and functional interactions of unknown open reading frames (ORFs) based on the clustering of ORFs into putative operons. Experiment-based datasets, on the other hand, are adaptations of established molecular biology protocols scaled-up to become 'high throughput'. For example, DNA microarrays represent the adaptation of standard hybridization techniques applied on a genomic scale. While there is a clear delineation between sequenced-based and experiment-based datasets, the primary utilization of these datasets is identical: the large-scale prediction of functional interactions. From a systems biology perspective, the de facto structure of a functional interaction can be reduced to a binary construct between two proteins or genes. Although this construct is a drastic oversimplification, it allows for both sequence-based and experiment-based datasets to be converted and combined into interaction sets. Stripped of all subtlety that exists in the interaction between genes within a genome, every pair of genes either interacts (1) or does not (0).

Because this binary characterization of an interaction bears little resemblance to reality, any interaction predicted using a functional genomics dataset must be experimentally verified and characterized with respect to spatial and temporal variables; at present, this process still occurs one interaction at a time. Further complicating this process is a lack of consensus regarding the definition of the term "functional interaction". The problem lies in the literal interpretation of the word "interaction" and, by extension, the purpose of its modifier, "functional". A functional interaction can be narrowly defined as only those proteins that engage in direct physical contact (Cusick *et al.*, 2005), or it can be broadly defined to include all of the proteins that are involved in a response or pathway (Eisenberg *et al.*, 2000). Both of these definitions represent extremes and, as such, contradictory examples abound. This lack of consensus does not represent a failure of the scientific community; it is possible that no rigorous definition exists. The difficulty in constructing a universally accepted definition for 'functional interaction' is similar to the current controversy in finding a definition for a bacterial species (Cohan, 2002; Moreno, 2002; Ochman *et al.*, 2005) or a planetary body (McCaughrean *et al.*, 2001; Basri and Brown, 2006; Sheppard, 2006).

Rather than commit to specific pairwise functional interactions, ontologies have been developed to categorize proteins into functional groups including Clusters of Orthologous Groups (COGs) (Tatusov *et al.*, 2000), Gene Ontology (GO) (Ashburner *et al.*, 2000), the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa *et al.*, 2004), and Protein Families (Pfam) (Finn *et al.*, 2006). Although every ontology uses an independent schema for their classifications, the fact that a computer script can convert data from one ontology to another is a strong indication of convergence. Ontological analysis provides insight into the relationship that exists between a gene and its genome by reducing the set of possible interaction partners from the whole genome to a subset. In the following section, we characterize the most common types of sequence-based and experiment-based functional genomics datasets available for prokaryotes.

Sequence-based genomic datasets

One of the major strengths of the sequence-based genomic datasets is that they are based upon a technology that is well advanced and produces robust data; current sequencing technology has advanced to the point where a genome sequence is generally accepted as "correct". This high degree of accuracy permits the detailed comparison of genome sequences across species, a type of analysis that becomes increasingly powerful as the number of available genomes increases. Prokaryotic genomes are relatively small when compared to eukaryotes, and so bacteria have the largest number of completed genome sequences. A recent listing of the completed prokaryote genome sequences deposited in NCBI shows that over 400 bacterial sequences are publicly available in contrast to the 22 available for eukaryotes. This number is expected to increase rapidly over the next few years, as a query of the Genomes OnLine Database (Liolios *et al.*, 2006) revealed a total of 1,070 ongoing bacterial sequencing projects.

The release of any given prokaryote genome sequence is accompanied by an initial genome annotation. The first, and perhaps most important annotation, is the mapping of predicted ORFs on the genome using gene finding programs such as GLIMMER (Delcher *et al.*, 1999). Advancements in this technology, as applied to prokaryotic genomes, is more evolutionary than revolutionary (Stein, 2001; Mathe *et al.*, 2002; Azad and Borodovsky, 2004; Wang *et al.*, 2004; Brent, 2005), and the prediction of ORFs is typically considered accurate in the same way the genome sequence itself is considered accurate. Each predicted ORF is then assigned a putative annotation based on various methods that rely on homology. The application of the Basic Local Alignment Search Tool (BLAST) (Altschul *et al.*, 1997) to compare each ORF against a database of characterized genes provides a basic annotation, and further methods of characterization are employed, such as classification using COGs (Tatusov *et al.*, 2000), the association of ORFs in known pathways using KEGG (Kanehisa *et al.*, 2004), and the homologous identification of protein domains using profile hidden Markov models (Eddy, 1998) against the Pfam database (Finn *et al.*, 2006). In addition to the annotation of ORFs, the genome sequence is also analyzed for additional features including non-coding regulatory RNAs (Eddy, 2002; Gottesman, 2005; Schattner *et al.*, 2005; Vogel and Sharma, 2005; Winkler, 2005) such as transfer-RNAs, ribosomal-

RNAs, micro-RNAs, and small, untranslated-RNAs. This wealth of information and analysis are considered by experimentalists as a starting point for the characterization of specific genes or genetic networks. These data can also be further exploited to construct functional genomics datasets that rely on a large number of different genome sequences. The three most common sequence-based genomic datasets available for prokaryotes include the enumeration of conserved operons, gene fusions as 'Rosetta Stones', and phylogenetic profiles and phylogenomic mapping.

THE ENUMERATION OF CONSERVED OPERONS

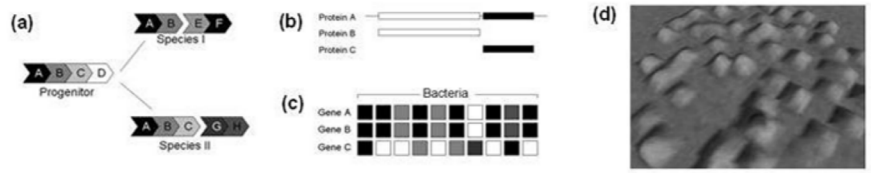
The very nature of bacterial function is embodied in the structure of the operon, a genetic unit which is under the control of a single transcriptional element (Jacob and Monod, 1961). Genes within a single operon likely participate in the same function. Operons, however, are not static elements, and the gene content of an operon is prone to rearrangement as prokaryotes evolve under selective pressure (Mushegian and Koonin, 1996; Itoh *et al.*, 1999). An analysis of related operons in different bacterial lineages indicates that they can contain non-identical genes (Dandekar *et al.*, 1998; Lathe *et al.*, 2000). Since genes within these operons are thought to participate in the same function, the same classification can be applied to all genes within the operon as shown in Figure 1a; previously unannotated genes receive a putative function (and putative functional interaction partners) through the principle of 'guilt-by-association' (Aravind, 2000; Oliver, 2000).

Gene clustering within operons has been successfully applied to the reconstruction of several metabolic pathways in a number of bacteria by Overbeek and colleagues (Overbeek *et al.*, 1999). In this study, the complete genome sequences of over 30 bacteria were compared by systematically querying the gene content of operons within these genomes. The order in which genes appear within operons was taken into account by computing a bidirectional best hit score with higher scores conferred to pairs of genes that were found adjacent to each other within multiple species. The phylogenetic relatedness of each genome was also taken into account as it is expected that the occurrence of genes within operons of closely-related species are similar. Using this approach, they reconstructed a number of metabolic pathways, including the purine biosynthesis and glycolysis pathways, and further identified additional genes within many of these operons that were not known to participate in the particular pathway. More recently, the application of gene clustering within operons was applied to over 190 sequenced bacterial genomes by Janga and colleagues (Janga *et al.*, 2005), and these predictions are publicly accessible through the Nebulon database (available at: http://tikal.cifn.unam.mx/~ediaz/NebulonNetView/).

GENE FUSIONS AS 'ROSETTA STONES'

The close association of genes within an operon can also be exploited in a different manner: gene fusions as 'Rosetta Stones' (Marcotte *et al.*, 1999). Gene fusions are based on the observation that a protein encoded by a single gene in one organism is sometimes found as two or more proteins encoded by multiple genes in other organisms (Figure 1b). Through selective pressure, sets of tightly coupled proteins can be fused

into a single protein in a different species. From this arrangement, the functional linkage of the genes encoding the protein domains of a fused protein can be established as a molecular 'Rosetta Stone'.



Database	URL	Datasets
Nebulon	http://tikal.cifn.unam.mx/~ediaz/NebulonNetView/	CO
FusionDB	http://igs-server.cnrs-mrs.fr/FusionDB/	RS
PhyloMap	http://myxococcus.syr.edu/Research/NBT_2005/index.htm	PM
PLEX	http://bioinformatics.icmb.utexas.edu/plex/	CO, RS, PP
PhydBac	http://igs-server.cnrs-mrs.fr/phydbac/	CO, RS, PP
Prolinks	http://dip.doe-mbi.ucla.edu/pronav	CO, RS, PP
STRING	http://string.embl.de/	CO, RS, PP

Figure 1. Common types of sequence-based functional genomic datasets available for prokaryotes. The enumeration of conserved operons (a) is based on the principle that the gene composition of operons that diverge across different lineages are related to the same function. In this example, genes A—H, found in related operons within different species, are functionally linked. Rosetta stones through gene fusions (b) are based on the observation that two different proteins can be found as a fused protein in another organism. In this example, protein B and C are fused together into Protein A. Phylogenetic profiling (c) is based on the idea of coinheritance where the presence of genes found across the same bacterial lineages are functionally related. In this example, Genes A and B have the same phylogenetic profile and are thought to be functionally linked. Phylogenomic mapping (d) is the application of clustering to phylogenetic profiles to produce a topographical map. A list of online databases for each of these datasets is also listed (CO = conserved operon; RS = Rosetta Stone; PP = phylogenetic profiling; PM = phylogenomic mapping).

The first application of gene fusions for the prediction of function in complete bacterial genomes was reported by Marcotte and colleagues (Marcotte *et al.*, 1999) for the model prokaryote *Escherichia coli*. Alignment of all proteins in the genome of *E. coli* with known protein sequences from a variety of protein databases produced a total of 6,809 pairs of non-homologous sequences. Using these putative gene fusion pairs, Marcotte and colleagues were able to reconstruct both the shikimate and purine biosynthesis pathways. A further application of the gene fusion model was presented by Enright and colleagues (Enright *et al.*, 1999) for *E. coli*, *H. influenzae*, and *Methanococcus jannaschii*. They found a total of 215 genes in all three genomes involved in 88 fusion events (64 of which are unique). Interestingly, an analysis of the gene fusion pairs showed that the majority were not found adjacent to each other in their respective genomes, as would be predicted based on the nature of operons. The prediction of gene fusions for any set of genes is available through a number of online databases including the Protein Link Explorer (PLEX) database (available at: http://bioinformatics.icmb.utexas.edu/plex/) (Date and Marcotte, 2005) and FusionDB (available at: http://igs-server.cnrs-mrs.fr/FusionDB/) (Suhre and Claverie, 2004).

PHYLOGENETIC PROFILING AND PHYLOGENOMIC MAPPING

The availability of numerous sequenced prokaryotic genomes enabled the construction of phylogenetic profiles and phylogenomic maps. Both of these functional genomics

datasets are based on the same underlying principle: groups of genes that are found to be conserved across bacterial lineages are thought to be functionally linked. As bacteria evolve due to the selective pressures imposed by their ecological niches, there is a considerable amount of genetic sharing that occurs through the process of horizontal gene transfer (Ochman *et al.*, 2000; Koonin *et al.*, 2001; Brown, 2003; Thomas and Nielsen, 2005). Since bacterial genomes rarely retain genes that do not confer a selective advantage within their environment (Ochman *et al.*, 2000; Gogarten *et al.*, 2002; Simonson *et al.*, 2005), comparisons of the gene content within genomes across different bacterial species can infer functional linkages between genes that are 'co-inherited', as shown in Figure 1c.

The construction of phylogenetic profiles has been successfully applied to the prediction of protein function in the genome of *E. coli*. Pellegrini and colleagues (Pellegrini *et al.*, 1999) compared the protein composition of *E. coli* against the protein composition of 16 other genomes by generating a raw data matrix with rows corresponding to proteins and columns representing the 16 genomes. The matrix cells contained data values that indicate either the presence or absence of the particular *E. coli* protein in the compared genome. Clustering of these profiles showed that proteins with similar phylogenetic profiles have a strong likelihood of being functionally linked. For example, they found that flagellar and cell-wall maintenance proteins share similar phylogenetic profiles, an indication of their functional linkage as flagella are known to be incorporated in the cell wall. A further extension of this model was applied to the prokaryotes *Caulobacter crescentus*, *Pseudomonas aeruginosa*, *Staphylococcus aureus*, and *Vibrio cholerae* by Date and Marcotte (Date and Marcotte, 2003). They compiled comparisons of these bacteria against 57 other sequenced genomes and were able to reconstruct a number of pathways including the SoxR oxidative stress response pathway in *V. cholerae*, a membrane transport pathway in *C. crescentus*, and a fatty acid synthesis pathway in *S. aureus*. Through their analysis, they were also able to link a number of uncharacterized proteins to these pathways; putative annotations of these proteins suggest that they are related to their respective pathways. The calculation of phylogenetic profiles for bacterial genome sequences is publicly available through the Phydbac database (available at: http://igs-server.cnrs-mrs.fr/phydbac/) (Enault *et al.*, 2004), where users can input protein sequences, or obtain pre-computed phylogenetic profiles for proteins in a number of bacteria.

More recently, Srinivasan and colleagues (Srinivasan *et al.*, 2005) have applied the principle of phylogenetic profiling to over 200 sequenced bacterial genomes. In this approach, termed 'phylogenomic mapping', phylogenetic profiles were constructed for each protein in all bacterial genomes used in the study. Profiles for each prokaryote's protein were clustered using Spearman's rank correlation to generate a similarity matrix. This similarity matrix was then processed using a combination of force-directed placement and multi-dimensional scaling in order to assign a two-dimensional coordinate to each protein. The visualization program *VxInsight* (Davidson *et al.*, 1998) was then used to display these proteins as a topographical map with mountains representing clusters of related proteins (Figure 1d). The phylogenomic map generated for *M. xanthus* was then used to generate predictions for genes involved in cell motility. A total of 15 such genes were selected and inactivated using homologous recombination. Motility assays were then applied to

these mutants and a total of 12 (or 80%) inactivated genes were found to be defective for motility, thus experimentally demonstrating the ability of this approach to predict function through clustering. Individual phylogenomic maps for all 200 bacteria used in this study are publicly available at: http://myxococcus.syr.edu/Research/NBT_2005/index.htm.

ONLINE RESOURCES FOR SEQUENCE-BASED DATASETS

In addition to the databases that exist for the calculation of specific sequence-based functional genomics datasets, there are a number of online resources that combine all of the approaches outlined in this section. As shown in Figure 1, PhydBac (Enault *et al.*, 2004), Prolinks (Bowers *et al.*, 2004), PLEX (Date and Marcotte, 2005), and STRING (von Mering *et al.*, 2005) combine all three sequence-based methods for the prediction of functional interactions in bacterial genomes. All of these databases contain pre-computed predictions for a number of sequenced bacterial genomes, and can also be searched for specific proteins. In addition, these databases allow users to input their own query sequences and obtain the calculated predictions for all three methods.

FURTHER CONSIDERATIONS ON SEQUENCE-BASED APPROACHES

For bacteria, the effectiveness of sequence-based functional genomics is increased by the large number of sequenced genomes and the relatively less complex structure of the prokaryotic genome. However, we must consider some issues that complicate the application of these approaches. Although there are many sequenced bacterial genomes, there is also a distinct sequencing bias. As shown in Table 1, not all groups of bacteria are equally represented in this set of sequenced genomes, and the disparity is largely due to the fact that researchers sequence the model organisms they study. For example, the large numbers of sequenced γ-proteobacteria are probably due to the medical relevance of this group.

In fact, all sequence-based functional genomics skew toward model organisms. For example, the accuracy of gene fusions depends on proteins from bacterial genomes, so that the sequencing bias limits the successful application of gene fusion to those prokaryotes whose target fusion protein exists in a bacterial species for which at least one complete sequence is available. Phylogenetic profiling and phylogenomics are similarly hamstrung by this bias; proteins specific to a bacterial taxonomic group will not have strong phylogenetic profiles if that group is largely not sequenced. As more bacterial sequences become available, the impact of sequencing bias will be mitigated, but for now these issues must be considered when applying sequence-based these techniques for the prediction of functional interactions in any bacteria. In addition, to this sequencing bias, there is also a disproportionate amount of information available for certain species of prokaryotes (Galvez *et al.*, 1998; Hugenholtz, 2002), such as *E. coli* and *Bacillus subtilis*, so that the interpretation of any predicted interaction is skewed toward our understanding of function within these well-studied model organisms.

Table 1. Sequencing Bias amongst completed Prokaryote Sequencing Projects deposited in the National Center for Biotechnological Information (NCBI). The number of sequenced genomes within a particular taxonomic group may reflect the research interests of the prokaryotic research community. For example, the large number sequenced Gammaproteobacteria may be due to the medical relevancy of many of the prokaryotes in this group. The values presented in this table are current as of 10/15/2006.

Taxonomic Group	Number of Sequenced Genomes
Acidobacteria	1
Aquificae	1
Fusobacteria	1
Nanoarchaeota	1
Planctomycetes	1
Thermotogae	1
Chloroflexi	2
Deinococcus-Thermus	4
Crenarchaeota	5
Spirochaetes	7
Bacteroidetes/Chlorobi	9
Epsilonproteobacteria	9
Chlamydiae/Verrucomicrobia	11
Deltaproteobacteria	11
Cyanobacteria	19
Euryarchaeota	22
Actinobacteria	25
Betaproteobacteria	28
Alphaproteobacteria	52
Firmicutes	82
Gammaproteobacteria	91

Experiment-based genomics datasets

Complete genome sequences have enabled the development of experiment-based technologies. Much of the analysis of experiment-based data is performed using methods that are the same or similar to those used for sequence-based data. While the greatest variety of experiment-based technologies are currently being applied to model organisms such as yeast, mouse, worm, and fly, a number of these technologies are beginning to appear for the most widely studied model prokaryotes. For example, a number of experiment-based functional genomics datasets are available for *E. coli* based on proteomics (Butland *et al.*, 2005; Han and Lee, 2006), DNA microarrays (Lucchini *et al.*, 2001), chromatin immunoprecipitation chips (ChIP-chip) (Herring *et al.*, 2005), localization studies (Arita *et al.*, 2005), gene essentiality studies (Joyce *et al.*, 2006), and metabolomics (Saito *et al.*, 2006). In this section we will focus on the two most common experiment-based genomic datasets available for many bacteria: proteomics, and DNA microarrays.

PROTEOMICS

The field of proteomics has rapidly evolved with the availability of whole genome sequences, and a number of techniques have been developed to probe the interactions that occur between proteins within a cell. The goal of proteomics is to provide evidence for protein-protein interactions that occur within the cell, and one of the most difficult challenges to this approach lies in the ability to extract and screen large numbers of proteins. In the bacterial realm, two major types of proteomics approaches have been applied for the genome-wide study of protein-protein interactions: yeast two-hybridization and protein profiling. Both approaches provide fundamentally different types of information about protein-protein interactions, and each have been successfully utilized to characterize the protein interactions that occur within the bacterial cell.

Yeast two-hybridization

The yeast two-hybridization method (Y2H), first proposed by Fields and Song (Fields and Song, 1989), remains one of the most applied techniques for the detection of direct interactions between proteins. In this approach, the direct interaction of two proteins is detected by the activation of a specific reporter in an organism such as yeast. The protein of interest, known as the 'prey' protein, is fused to a DNA-binding domain of a specific transcription factor such as GAL4; a second protein, the 'bait' protein, is fused to the activator domain of the same transcription factor. Both protein fusions are brought together through the yeast mating system and the interaction of both proteins causes the activation of the reporter transcription factor. Yeast colonies are plated on media that require the activation of the reporter gene through the successful interaction of both proteins. As a result, protein-protein interactions are indicated if growth of yeast colonies is observed. The specifics of this technology and variations in Y2H approaches are discussed in great detail elsewhere (Phizicky and Fields, 1995; Zhu *et al.*, 2003; Causier, 2004; Miller and Stagljar, 2004; Parrish *et al.*, 2006). We will focus our discussion on how this method is applied at the genome and subgenome scale to predict functional interactions in bacteria.

To date, only two whole-genome protein-protein interaction maps have been reported for prokaryotes: *Helicobacter pylori* (Rain *et al.*, 2001) and *Rickettsia sibirica* (Malek *et al.*, 2004). The *H. pylori* study used a modified version of the Y2H system and a total of 285 screens were conducted using 261 bait proteins against a protein prey library containing randomly sheared portions of the genome. Bait proteins were chosen either at random, or based on their known involvement in specific complexes or roles in pathogenicity. More than 1,200 protein-protein interactions were identified in this manner, representing connections between 47% of the *H. pylori* genome. Further analysis of these connections revealed novel interactions between proteins involved in the *H. pylori* chemotaxis and urease pathways. In addition, the detected interactions also allowed for the putative assignment of function to many of the previously uncharacterized ORFs within the *H. pylori* genome. The second study on a whole-genome interaction map based on the Y2H approach in prokaryotes was applied to the pathogen *R. sibirica*, a bacterium responsible for spotted fever (Fournier *et al.*, 1998). In this study, a total of 284 protein-protein interactions were obtained

between 150 proteins, representing 12% of the total predicted ORFs in the genome of *R. sibirica*. Analysis of these interactions showed that 24 unannotated proteins were found to interact with the subunits of the type IV secretion pathway, which is known to be integral to the virulence of this prokaryote. In addition, new protein-protein interactions between known and suspected virulence genes were also established, highlighting the utility of this approach for the identification of putative functional interactions.

A more common approach to the use of the Y2H system in bacterial genomics is to conduct a subset genomic screen on a small number of specific proteins of interest, rather than generating data on a whole genome. Using this method, numerous examples of protein-protein interactions discovered through Y2H screens have been reported for a variety of prokaryotes, such as *Agrobacterium tumefaciens* (Liu *et al.*, 2001), *B. subtilis* (Noirot-Gros *et al.*, 2002; Dervyn *et al.*, 2004), *C. caulobacter* (Ohta and Newton, 2003), *E. coli* (Hall *et al.*, 1998), *Mycobacterium tuberculosis* (Steyn *et al.*, 2002), and *Xanthomonas axonopodis* (Alegria *et al.*, 2004). Recently, this approach was used for the detection of proteins that interact with MglA, a cytoplasmic GTPase that is required for motility in *M. xanthus*. Yang *et al.* (2004) utilized the *M. xanthus* genome sequence to construct a prey protein library and probed this library using MglA as bait. As a result, they were able to detect the interaction of MglA with AglZ, a type-2 myosin-like protein. Further verification of this interaction suggested that these two proteins interact to control the dual motility systems in *M. xanthus*; interestingly, this was also the first reported interaction between a GTPase and a myosin-like protein in a prokaryote.

While the utility of the Y2H system is self-evident, there are significant concerns about the reproducibility of data. Interactions predicted by the Y2H system are notorious for their high rates of false positives; some reports have suggested that close to 50% of all observed interactions are false (von Mering *et al.*, 2002; Sprinzak *et al.*, 2003). Even more alarming is the low overlaps observed between independently produced genome-wide Y2H studies. For example, analysis of two large-scale Y2H experiments performed for yeast showed only 20% overlap (Hazbun and Fields, 2001; Ito *et al.*, 2001), and similar results have been reported for Y2H experiments conducted for *Drosophila melanogaster* (Giot *et al.*, 2003; Formstecher *et al.*, 2005). These results have prompted one of the pioneers of the Y2H system, Stanley Fields, to openly question whether or not such large-scale undertakings are even worth pursuing (Fields, 2005). Y2H limitations do raise some important issues regarding reliability and reproducibility thresholds. The results obtained from large-scale Y2H studies are best viewed as predictions that can then be used to formulate hypotheses about protein-protein interactions and suggest further studies. Y2H produces a large amount of useful information and represents only one type of protein-protein interaction indicator, which becomes more meaningful if combined with other functional genomics datasets. Finally, Y2H technology continues to evolve (Suter *et al.*, 2006), so that the ability to both detect and analyze protein-protein interactions is improving, and will eventually increase the reliability of this approach.

Protein profiling

In contrast to the Y2H system, where individual protein-protein interactions are

predicted, protein profiling serves to identify those proteins that are expressed by a cell under specific conditions. The ability to profile an organism's protein output can reveal how it responds to certain environmental conditions, and provide clues to the function of these expressed proteins. The traditional method employed to detect the full complement of proteins expressed by a cell is through the combination of 2D gel technology and mass spectrometry (MS). In this approach, a cell is harvested for its complement of proteins and separated using a 2D slab gel. This results in the separation of the proteins based on their isoelectric charge and molecular mass. MS is used to further identify the proteins separated in the 2D gel. More recent advances in MS technology, such as the development of the matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) method, have further increased the ability to identify proteins extracted from a gel. There are significant technological limitations to this method, and even the most sophisticated experiments are only capable of identifying a portion of the total expressed proteins. Specific aspects of these technologies have been summarized (Zhu *et al.*, 2003; Causier, 2004; Jessani and Cravatt, 2004; Phillips and Bogyo, 2005) and we will focus our discussion on bacterial systems.

The application of protein profiling to bacterial systems is roughly divided into two classes of experiments: vegetatively growing cells, and cells undergoing a specific stress response (Wolff *et al.*, 2006). The goal of both experiments, however, is the same: the identification of those proteins that contribute to a specific cell state. So far, protein profiles have been generated for the vegetative growth of a number of bacteria, including *B. subtilis* (Eymann *et al.*, 2004), *Deinococcus radiodurans* (Lipton *et al.*, 2002), *E. coli* (Taoka *et al.*, 2004; Butland *et al.*, 2005), *Mycoplasma genitalium* (Wasinger *et al.*, 2000), and *M. mobile* (Jaffe *et al.*, 2004). An important finding from these studies was the ability to estimate the number of actively transcribed genes within the genome. For example, an analysis of the *M. mobile* protein profile reveals that about 88% of its 635 ORFs can be detected during vegetative growth. Given the small size of the *M. mobile* genome, this provides some insight into the minimal set of proteins required by a cell (Jaffe *et al.*, 2004). Recently, the protein profiling method has been applied to vegetative growth in cultures of *M. xanthus* by Schley and colleagues (Schley *et al.*, 2006). In this study, the authors were able to identify 631 unique proteins, corresponding to approximately 10% of the predicted ORFs in the *M. xanthus* genome. Analysis of the identified proteins revealed that a number of previously uncharacterized polyketide synthases were present in the cells. A survey of the *M. xanthus* genome reveals that a total of 18 polyketide synthase clusters are present, however, very few of these clusters were thought to be active, as *M. xanthus* is not known as a major producer of bioactive products. The presence of these polyketide synthases during the vegetative growth of *M. xanthus* is counterintuitive, since it is metabolically expensive to express these proteins, and it seems unnecessary to produce them under nutrient-rich conditions.

The vegetative protein profile of a bacterial system can also be used a reference set for the comparison of a stress response, and thereby reveal the specific proteins that participate in that response. This approach has been successfully applied to the characterization of proteins that are involved in biofilm formation in *B. subtilis* (Vilain and Brozel, 2006), *P. aeruginosa* (Sauer *et al.*, 2002), and *Streptococcus mutans* (Luppens and ten Cate, 2005). The identification of proteins that are expressed in response to a specific stress provides information regarding the function of these

proteins, resulting in the prediction of functional interactions. As additional large-scale protein profiling experiments become available, the ability to cluster the protein expression patterns in a manner similar to that of phylogenomic mapping becomes possible. Proteins with similar profiles would indicate tightly coupled proteins whose expression is correlated under multiple conditions. As our ability to detect the complete protein content in a cell advances, this approach will undoubtedly provide an invaluable experiment-based functional genomics dataset.

DNA MICROARRAYS

DNA microarrays are by far the most common experiment-based functional genomics technique utilized by prokaryotic researchers. The purpose of a microarray experiment is to compare the steady-state mRNA levels between two samples that have been exposed to different conditions, either environmental or genetic. Messenger RNA is harvested from both samples and converted to cDNA with reverse transcriptase. Fluorescent dyes are used to label both samples, one with green cyanine dye (Cy3) and the other with red cyanine dye (Cy5). Both samples are then combined and hybridized to the same microarray that typically contains thousands of regularly spaced DNA spots. The DNA in each spot contains the complement of a fragment of one ORF in the genome, and any gene that was actively transcribed in either sample will hybridize with the complementary DNA spot on the microarray. Because the two samples are labeled with different colored dyes, the relative steady-state mRNA levels of each gene can be measured for each of the two samples. Further explanation of DNA microarray technology as applied to bacterial genomes is available elsewhere (Lucchini *et al.*, 2001; Southern, 2001; Rhodius *et al.*, 2002; Ehrenreich, 2006). We will focus the rest of our discussion on the use of microarray data in experimental pipelines.

Comparative microarray analaysis

The analysis of DNA microarrays typically revolves around the fundamental observation that genes sharing similar expression patterns from a collection of microarray experiments have a high likelihood of being functionally related (Ge *et al.*, 2001). Multiple experiments in a large number of model organisms have confirmed this hypothesis (Jansen *et al.*, 2002; Walhout *et al.*, 2002). One of the first applications of DNA microarrays to bacteria was the comparison of two genetically identical populations under different environmental conditions. For example, analyses of gene expression for *E. coli* have been conducted by comparing microarray data for cells in minimal vs. rich media (Tao *et al.*, 1999) and for groups of *E. coli* containing small and large populations (Liu *et al.*, 2000). Similar microarray experiments have also been reported for other bacteria under comparative conditions such as drug treatment in *M. tuberculosis* (Wilson et al., 1999) and *Streptomyces coelicolor* (Huang *et al.*, 2001), and stress conditions in both *B. subtilis* (Ye *et al.*, 2000) and *P. putida* (Reva *et al.*, 2006).

Many studies have used microarrays to probe differential gene expression between two strains of the same bacteria. In one type of experiment, a specific gene of interest

is inactivated and a comparative microarray analysis is conducted between the mutant strain and wild type. This type of analysis has been utilized effectively in a number of bacteria including *B. subtilis* (Cao *et al.*, 2003), *E. coli* (Barbosa and Levy, 2000; Salmon *et al.*, 2003), *Lactococcus lactis* (den Hengst *et al.*, 2005), and *M. xanthus* (Diodati *et al.*, 2006), where it was applied to the discovery of genes differentially expressed due to the inactivation of the NtrC-like transcriptional activator *nla18*. Diodati and colleagues (Diodati *et al*., 2006) utilized a whole-genome cDNA microarray for *M. xanthus* and compared the differential expression of genes between wild type and a strain of *M. xanthus* inactivated for *nla18* under vegetative growth. They found that over 700 ORFs were differentially expressed, and an analysis of these ORFs revealed that the majority encoded membrane and membrane-associated proteins. Results from this set of experiments therefore indicated a possible role for *nla18*.

Gene expression mapping

The wide-spread use of microarrays as a tool for the large-scale investigation of differential gene expression in bacteria has resulted in the accumulation of a substantial amount of publicly available microarray data. A number of online databases exist to house this data, such as ArrayExpress (Brazma *et al.*, 2006), the Gene Expression Omnibus (Barrett *et al.*, 2005), and the Stanford Microarray Database (Ball *et al.*, 2005). The availability of a large number of microarray experiments under a wide variety of experimental conditions has stimulated the development of tools aimed at analyzing this overabundance of data.

While the focus of the majority of microarray experiments is to determine those genes that are differentially expressed under a specific condition, there is also value in identifying those genes that are similarly expressed under all conditions. Combining the differential gene expression patterns for an organism under a multitude of experimental conditions provides insight into genes whose expression is tightly correlated, a strong indication of a functional interaction (Ge *et al.*, 2001). The first application of this approach was reported by Kim and colleagues (Kim *et al.*, 2001) for the model organism *Caenorhabditis elegans*. They obtained over 500 microarray experiments from a number of sources conducted over a wide-range of experimental conditions and constructed a gene expression map. Construction of this map followed the same process used to construct the phylogenomic map presented earlier. A data matrix containing Cy3/Cy5 values was constructed in which rows corresponded to a specific gene in *C. elegans* and columns represented the individual microarray experiments. Clustering of these gene expression profiles was applied using Pearson's correlation and a combination of force-directed placement and multi-dimensional scaling was used to place each gene onto a topographical map. Analysis of the resulting map showed that those genes exhibiting similar annotations, such as genes involved in sperm production, were found to cluster together into a single mountain. Gene expression mapping has only recently been applied to bacterial genomes, and currently maps exist for *E. coli*, *M. tuberculosis*, *M. xanthus*, and *S. coelicolor* (Suen *et al.*, 2006).

The reproducibility and reliability of microarray data

One of the major criticisms of the DNA microarray platform is the large amount of

noise that complicates the prediction of functional interactions. As a result, the issues of the reliability and reproducibility of microarray experiments have become the focal point of two large-scale studies involving major consortiums. The May 2005 issue of *Nature Methods* featured three articles that presented the findings of the Toxicogenomics Research Consortium (Bammler *et al.*, 2005; Irizarry *et al.*, 2005; Larkin *et al.*, 2005). This study examined the correlation of microarray experiments conducted across multiple laboratories using different platforms and protocols. They concluded that the best reproducibility was found between commercially available microarray platforms; reproducibility was found to be poorest when commercial and non-commercial arrays were compared, with correlation values as low as 0.11. More recently, *Nature Biotechnology* devoted their entire September 2006 issue to the findings of the Microarray Quality Control Group. In this study, the authors used seven different microarray platforms in a variety of labs to conduct more than 1,300 tests that sampled the expression levels of more than 12,000 genes. They found that their expression data overlapped significantly, with reproducibility in the order of 70% to 90%.

With such contradictory results in the analysis of the reproducibility of microarray data in these studies, there is still great controversy surrounding the reliability of the microarray platform. One significant difference in the way these two studies were conducted was that, in the study that reported low reproducibility, a specific set of protocols for performing microarray experiments was not enforced across different laboratories. Instead, each participating laboratory used its own protocol. In contrast, the study that reported high reproducibility enforced a standard set of protocols for all experiments. If that important difference is responsible for the difference in observed results then, for most bacterial research communities, the results of microarray experiments is most likely closer to the those reported by the Toxicogenomics Research Consortium for two reasons. First, it is not possible to enforce a standard set of protocols and second, the Microarray Quality Control Group study does not begin to address if their set of protocols produces 'correct' results. There are similar concerns regarding the validity of underlying data in other experiment-based datasets, such as Y2H, and the same general principles also apply there. These issues underscore a crucial difference between sequenced-based and experiment-based datasets, because there are relatively few concerns about the validity of the underlying DNA sequence data. As microarray platform technology continues to evolve, and as new methods for normalizing and processing microarray data are developed, these issues will eventually be minimized.

Case study: using microarray data as part of an experimental pipeline

In the following section, we present a case study for the integration of microarray data into an experimental pipeline for the identification of functional interactions in *M. xanthus*. The main challenge in integrating functional genomics datasets into an experimental setting is the large amount of data that must be processed. The typical use of functional genomics datasets within an experimental context is to obtain an extensive list of predictions and then test these predictions one at a time. While this approach provides a manageably small test set, it is not uncommon for the test set to grow extensively large, numbering in the many hundreds. The testing of these

predictions by standard experimental assays such as genetic analysis thus becomes both time and cost prohibitive.

We have developed an experimental pipeline that incorporates different interpretations of large-scale microarray datasets for the identification of the downstream targets of particular transcriptional regulators (Figure 2). A survey of the *M. xanthus* genome (Goldman *et al.*, 2006) revealed some striking peculiarities, specifically with respect to signaling pathways. The genome encodes for 99 serine/threonine kinases, 137 sensor and hybrid histidine kinases, 48 different sigma-factors, and over 50 different NtrC-like activators. Analysis of both one- and two-component regulators in *M. xanthus* reveals that its regulatory networks are atypical among prokaryotes. *M. xanthus*' signaling networks branch, often having at least 2 components, with sites for sensory input upstream of the regulators of transcription. This is attributed to the large number of multi-site DNA binding transcriptional regulators (TR) found in *M. xanthus*. This large and complex genome makes understanding the connection between TR cascades and behavior difficult to resolve.

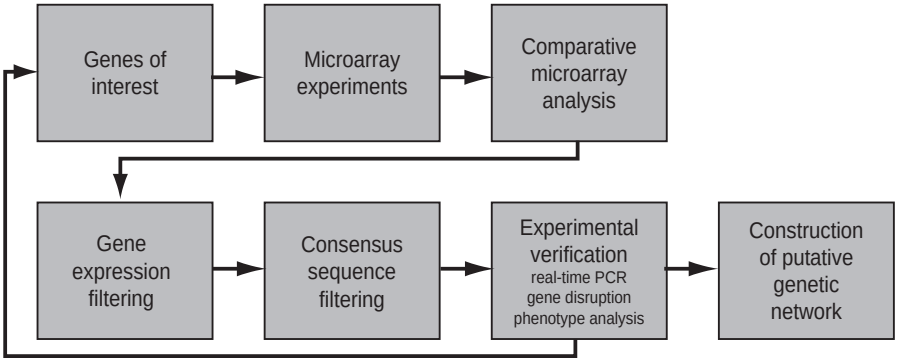


Figure 2. The integration of experiment-based functional genomics datasets to an experimental pipeline. In this example, a strain of the target organism inactivated for a specific TR is subjected to comparative microarray analysis over a time course. Sets of up- and down-regulated genes are retained and filtered through a gene expression map constructed for the target organism. This reduced set is further filtered by searching the upstream intergenic regions of each predicted gene for the presence of a putative binding site. Experimental verification of these predictions are then used to construct a putative genetic network, and specific genes of interest characterized in this manner can be used as the specific gene of interest for the next cycle of the pipeline.

One important class of TRs is the NtrC-like activators, also known as sigma-54 enhancer binding proteins (Morett and Segovia, 1993; Caberoy *et al.*, 2003; Jakobsen *et al.*, 2004; Jelsbak *et al.*, 2005), and it has been shown that these activators are important for the expression of over half of the genes required for normal development (Caberoy *et al.*, 2003). We have recently begun an integrated genomics study into the downstream targets of the NtrC-like activator *nla6* in collaboration with the laboratory of Professor Anthony Garza (Syracuse University). A time-course of microarray experiments was conducted by comparing the gene expression levels of almost all genes in the genome between a strain of *M. xanthus* inactivated for *nla6* and wild type under starvation conditions. For this time course, only those genes that showed at least a four-fold change in steady-state RNA levels (i.e. $-0.5 < \log2(Cy3/Cy5) > 2$) were retained. This produced a list of 846 ORFs, which represents

approximately 10% of the total genome. This list therefore includes all ORFs affected by the inactivation of *nla6* both directly and indirectly.

Ideally, we would like to identify the direct downstream targets of *nla6* from this extensive list of predictions. In order to accomplish this, we constructed a gene expression map for *M. xanthus* based on 212 different microarray experiments. Since the gene expression map clusters ORFs that share similar gene expression patterns across a wide variety of experimental conditions, we hypothesize that the functional predictions for *nla6* have a higher likelihood of including direct targets. We compared the set of 846 ORFs predicted using the time-course experiment by filtering against the top 100 predictions retained for *nla6* using gene expression mapping. A total of 50 ORFs were found to overlap between these two predicted sets, and these predictions became the subject of further analysis.

To further characterize the putative downstream targets selected by comparative microarray analysis and gene expression filtering, we searched the upstream intergenic regions of these predictions for putative NtrC-like binding sites using the online resource PromScan (available at: http://www.promscan.uklinux.net/). The application of this filtering step reduced the total number of predictions to 9 targets. DNA binding assays were then employed to confirm the ability of *nla6* to binding to the upstream regions of these targets, and our current success rate is greater than 50% (data not shown). Further characterization of these targets through gene inactivation and phenotypic analysis also confirm these results. The utility of this approach is exhibited through the systematic application of filters to reduce the number of predictions obtained through comparative microarray analysis. In this way, a small set of highly probable predictions can be obtained and further characterization of these ORFs can lead to the construction of a putative genetic network. This approach is iterative, and can also be successively applied to the confirmed predictions as further comparative microarray time-course experiments are conducted and filtered in the same manner.

Integration of functional genomics datasets: systems biology

The accumulation of a wide variety of different functional genomics datasets has resulted in the development of the field of systems biology (Kitano, 2002). Systems biology attempts to decipher the complex behaviors exhibited by an organism by understanding how these behaviors are derived from the interactions of its underlying components. While this is a seemingly broad statement, when viewed within the framework of functional genomics much of this description can be reduced to the specific interactions that exist between proteins within the cell; in essence, systems biology is a method for the interpretation of function as it pertains to these interactions. The view that function can be encapsulated into modules of genetic networks represents a paradigm shift in the way we represent the genome of any living organism (Hartwell *et al.*, 1999). The ability to quantify the inputs and outputs that exist for genes and proteins within and between modules represents a critical step in describing the function of a genome. Each genomics dataset represents one kind of functional relationship that defines the modules and shapes the composition of genetic nodes in these networks. Systems biology strives to provide a unifying contextual framework to these functional relationships through the application of integration algorithms to

multiple genomics datasets. The current generation of these algorithms focuses on probabilistic integration, and the most recent of these have demonstrated the ability to recapitulate known pathways such as the galactose utilization pathway in yeast (Hwang *et al.*, 2005a,b).

Since all functional genomics datasets represent a method for the prediction of a functional interaction, the integration of many such datasets is thought to provide greater confidence in the individual predictions (Marcotte and Date, 2001; Vidal, 2001; Gerstein *et al.*, 2002; Ge *et al.*, 2003). In other words, an interaction that is predicted to occur between two genes through DNA microarray experiments and phylogenetic profiling has a higher probability of being a 'real' interaction then if it were predicted by only one of these methods. This postulate has gained wide acceptance within the systems biology community, and an ever-increasing number of new algorithms for the integration of functional genomics datasets are being reported (Marcotte *et al.*, 1999; Ge *et al.*, 2003; Joyce and Palsson, 2006). Interestingly, the most profitable advances in the application of these algorithms are found in the integration of a large number of functional genomics datasets. For model organisms such as yeast, mouse, and fly, where a large number of functional genomics datasets are available, it is not surprising that the majority of integration algorithms are being reported in these systems. The trend within the field of bacterial functional genomics, however, is to apply only a handful of these functional genomics datasets for the characterization and verification of genetic networks. This is due in large part to the unavailability of multiple datasets for the majority of the sequenced bacterial genomes.

The utilization of functional genomics datasets in bacterial systems reflects the clear delineation that exists between sequence-based and experiment-based genomics datasets, as the majority of studies report the use of one or the other, but seldom both simultaneously. For example, the application of sequence-based functional genomics datasets such as gene fusions, conserved operon, and phylogenetic profiles for the prediction of functional interactions have been reported for both *E. coli* (Wu *et al.*, 2005) and *M. tuberculosis* (Strong *et al.*, 2003a,b). Similarly, the combination of different experiment-based functional genomics datasets such as DNA microarrays and protein profiling has been used to predict functional interactions in B. *subtilis* (Budde *et al.*, 2006), *Desulfovibrio vulgaris* (Mukhopadhyay *et al.*, 2006), *E. coli* (Corbin *et al.*, 2003), and *Shewanella oneidensis* (Kolker *et al.*, 2005). The utilization of exclusively sequence-based or experiment-based genomic datasets to make functional predictions may be due to the inherently complementary nature of the data. For example, DNA microarray data measures the ratio of steady state mRNA for a given gene under a specific condition. Protein profiling of the same cell under the same conditions provides information about the amount of expressed protein. Both of these datasets can be used to confirm the expression of a gene and its translated protein under that specific condition. It is likely that the functional predictions made through the combination of these two related experiment-based methods are stronger than combining either with a sequence-based method, such as phylogenomic profiling.

A recent statistical analysis of functional genomics datasets in yeast found that functional predictions made using similar types of data were more accurate than predictions made by disparate sets (Lu *et al.*, 2005). A similar observation was made by Suen and colleagues (2006) in the comparison of functional predictions made by phylogenomic and gene expression mapping. In this study, genome-scale interactions

were predicted for four prokaryotes: *E. coli, M. tuberculosis*, *M. xanthus*, and *S. coelicolor*, by combining different quantities of either phylogenomic (sequence-based) or microarray expression (experiment-based) data, or both. The authors found that the actual number of overlapping predictions made by the two datasets was low, and did not improve significantly as more genomic data was incorporated into the predictions. Since researchers working on most bacterial model organisms currently have access to only a genome sequence and a DNA microarray, the application of these datasets for functional predictions may be limited. As the availability of new datasets increases for bacterial systems, the application of probabilistic integration algorithms for the prediction of functional interactions should eventually approach the level of success reported for eukaryotic model organisms.

Functional genomics as genome annotation

Functional genomics is an advanced form of genome annotation – the process of attaching biological meaning to a DNA sequence. Annotation can be summarized as the application of two largely independent steps: the identification of a list of genetic elements on a genome (a process called gene finding, or structural annotation) followed by the attachment of biological meaning to these elements (a process called functional annotation) (Stein, 2001). For prokaryotes, structural annotation is largely automated, and while these annotations can change as the body of experimental data grows or algorithms for gene predictions improve, these changes will most likely be evolutionary, so that most prokaryotic structural annotations are considered relatively stable. In contrast, functional annotations are dynamic and subject to frequent change.

The functional annotation of a gene is traditionally assigned by one of two methods. In the first method, a novel annotation is produced through experimental research conducted on a specific gene, and this accumulated body of knowledge is used to assign roles to these genes. The second method is annotation by homology, where a functional annotation is assigned to one gene in a genome because it is homologous to a second gene that has been annotated in another organism; for obvious reasons involving the accumulation of error, it is optimal if that second (or third, or fourth) gene was not also annotated by homology, thus producing a phenomenon known as 'genome rot' (Galperin and Koonin, 1998; Gerlt and Babbitt, 2000). The annotation of genes through homology is routinely applied to newly sequenced genomes and approximately 65% of the predicted ORFs can be assigned some sort of annotation (Karaoz *et al.*, 2004); the challenge is how best to assign a functional annotation to the remaining 35%. The topics described in this review represent a method for the high-throughput prediction of annotation assignments, and the critical step lies in confirming these predictions. Thus, the role of the experimentalist as an expert in the biology of a given bacteria is required for the confirmation of any predicted interaction. The application of functional genomics to experimental research also results in the rapid accumulation of a large amount of data, and new methods are needed to manage and disseminate this data to the research community.

To address this challenge, the construction of Model Organism Databases (MODs) has been applied to the bacterial genome. These databases serve as a common repository for the collected knowledge regarding the structural and functional

annotation of the model organism. Most MODs also include a multitude of analysis tools, and often provide links to other annotation resources such as COG (Tatusov *et al.*, 2000), GO (Ashburner *et al.*, 2000), KEGG (Kanehisa *et al.*, 2004), and Pfam (Finn *et al.*, 2006). The MOD is now a common tool employed by many bacterial research communities. A survey of the 2006 Molecular Biology Database Collection (available at: http://nar.oxfordjournals.org/cgi/content/full/34/suppl_1/D3) reveals a list of 858 genomic databases, 46 of which are specific to prokaryotes. Some communities, such as *E. coli*, even have multiple MODs, each geared toward a specific aspect of the organism such as metabolism. A survey of the evolution of these databases reveals that many are providing a forum for the development of dynamic functional annotations; for the most part, MODs are not static.

As a result, expert curation is integral to the correct annotation of any model organism's genome. The traditional curation model is centered on a group of professional annotators who attempt to synthesize the vast amount of scientific knowledge for each model organism in order to annotate its genome (Stein, 2001). Other types of curation models include the party model, where a community of scientists will meet periodically for the purposes of joint genome annotation, and the cottage industry model, where experts within a field are recruited on a part-time basis for annotation efforts. A novel curation model for the *M. xanthus* MOD has recently been reported by Arshinoff and colleagues (Arshinoff *et al.*, 2007). Specifically, xanthusBase (available at: http://xanthusbase.org), incorporates standard automated annotation methods such as GO, KEGG, and Pfam, in combination with a non-standard annotation method that leverages the expertise of researchers within the *M. xanthus* scientific community. In this annotation method, xanthusBase uses "wiki editing principles" utilized by the highly-successful wikipedia online encyclopedia (http://www.wikipedia.com). In this way, all *M. xanthus* researchers are granted full access to edit the contents of the database, while simultaneously protecting the data from accidental, malicious, or overly contentious changes.

The wide-spread availability of functional genomics datasets for a number of bacteria has also resulted in their incorporation into MODs. For example, a number of the *E. coli* MODs now include DNA microarray (Keseler *et al*., 2005) and protein profiling data (Janga *et al*., 2005). As more genomic datasets are incorporated into MODs, the processing of this data will likely move toward computer automation, resulting in the further annotation of the genome using algorithmic methods. In the future, it is conceivable that researchers will be able to upload genomic data to a MOD, where it will automatically be incorporated, processed, and used to derive new and more accurate annotations. Currently this is still a conceptual model, but its origins lie in the systematic application of genomic data to predict functional interactions. MODs are already storing the relevant data and the functional annotation, so it is only a matter of time until they are used to make predictions that help researchers design experiments.

Large-scale annotation efforts are moving away from a group of central curators to an amalgamation of different curation methodologies, so careful attention must be paid to the types of evidence used to prove an annotation. This type of curation is already in practice with the classification method used by the Gene Ontology Consortium (Ashburner *et al.*, 2000). Each ontological assignment is supported by a set of evidence codes that range from data reported in the literature to gene expression

profiles from DNA microarray experiments. These codes measure something real that is difficult to quantify, as it can be argued that researchers are more likely to trust an annotation based on careful biochemical characterization than on the correlated gene expression patterns observed from a handful of DNA microarray experiments. Nevertheless, the appeal of functional genomics lies not in its accuracy, but rather in its capacity to rapidly predict functional interactions for a large portion of a genome; any perceived degradation in accuracy should be viewed as the inherent cost of efficiency. Annotations created from genomics datasets will serve as starting points to direct the research of experimentalists who will be able to strengthen an annotation through more rigorous methods of characterization. Simultaneously, advancements will be made in the field of functional genomics and experiment-based datasets will increase in their reproducibility, so that functional predictions should significantly improve. Therefore, a basic understanding of the principles used to make predictions from functional genomics datasets should be a requisite part of the training program for experimental microbiologists of the future.

Acknowledgements

The authors would like to thank Barry S. Goldman, Kimberly A. Murphy and Sarah F.W. Brisbin for critical reading of this manuscript.

References

ALEGRIA, M.C., DOCENA, C., KHATER, L. *ET AL.* (2004). New protein-protein interactions identified for the regulatory and structural components and substrates of the type III secretion system of the phytopathogen Xanthomonas axonopodis Pathovar citri. *Journal of Bacteriology* **186**, 6186-97.

ALTSCHUL, S.F., MADDEN, T.L., SCHAFFER, A.A. *ET AL.* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389-402.

ARAVIND, L. (2000). Guilt by association: contextual information in genome analysis. *Genome Research* **10**, 1074-7.

ARITA, M., ROBERT, M. and TOMITA, M. (2005). All systems go: launching cell simulation fueled by integrated experimental biology data. *Current Opinion in Biotechnology* **16**, 344-9.

ARSHINOFF, B.I., SUEN, G., JUST, E.M. *ET AL.* (2007). XanthusBase: Adapting Wikipedia Principles to a Model Organism Database. *Nucleic Acids Research* **35**, D422-6.

ASHBURNER, M., BALL, C.A., BLAKE, J.A. *ET AL.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics* **25**, 25-9.

AZAD, R.K. and BORODOVSKY, M. (2004). Probabilistic methods of identifying genes in prokaryotic genomes: connections to the HMM theory. *Briefings in Bioinformatics* **5**, 118-30.

BALL, C.A., AWAD, I.A., DEMETER, J. *ET AL.* (2005). The Stanford Microarray Database accommodates additional microarray platforms and data formats. *Nucleic Acids Research* **33**, D580-2.

BAMMLER, T., BEYER, R.P., BHATTACHARYA, S. *ET AL.* (2005). Standardizing global gene

expression analysis between laboratories and across platforms. *Nature Methods* **2**, 351-6.

BARBOSA, T.M. and LEVY, S.B. (2000). Differential expression of over 60 chromosomal genes in Escherichia coli by constitutive expression of MarA. *Journal of Bacteriology* **182**, 3467-74.

BARRETT, T., SUZEK, T.O., TROUP, D.B. *ET AL.* (2005). NCBI GEO: mining millions of expression profiles—database and tools. *Nucleic Acids Research* **33**, D562-6.

BASRI, G. and BROWN, M.E. (2006). PLANETESIMALS TO BROWN DWARFS: What is a Planet? *Annual Review of Earth and Planetary Sciences* **34**, 193-216.

BENSON, D.A., KARSCH-MIZRACHI, I., LIPMAN, D.J., OSTELL, J. and WHEELER, D.L. (2006). GenBank. *Nucleic Acids Research* **34**, D16-20.

BOWERS, P.M., PELLEGRINI, M., THOMPSON, M.J. *ET AL.* (2004). Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biology* **5**, R35.

BRAZMA, A., KAPUSHESKY, M., PARKINSON, H., SARKANS, U. and SHOJATALAB, M. (2006). Data storage and analysis in ArrayExpress. *Methods in Enzymology* **411**, 370-86.

BRENT, M.R. (2005). Genome annotation past, present, and future: how to define an ORF at each locus. *Genome Research* **15**, 1777-86.

BROWN, J.R. (2003). Ancient horizontal gene transfer. *Nature Reviews Genetics* **4**, 121-32.

BUDDE, I., STEIL, L., SCHARF, C., VOLKER, U. and BREMER, E. (2006). Adaptation of Bacillus subtilis to growth at low temperature: a combined transcriptomic and proteomic appraisal. *Microbiology* **152**, 831-53.

BULL, C.T., SHETTY, K.G. and SUBBARAO, K.V. (2002). Interactions Between Myxobacteria, Plant Pathogenic Fungi, and Biocontrol Agents. *Plant Disease* **86**, 889-896.

BUTLAND, G., PEREGRIN-ALVAREZ, J.M., LI, J. *ET AL.* (2005). Interaction network containing conserved and essential protein complexes in Escherichia coli. *Nature* **433**, 531-7.

CABEROY, N.B., WELCH, R.D., JAKOBSEN, J.S., SLATER, S.C. and GARZA, A.G. (2003). Global mutational analysis of NtrC-like activators in Myxococcus xanthus: identifying activator mutants defective for motility and fruiting body development. *Journal of Bacteriology* **185**, 6083-94.

CAO, M., SALZBERG, L., TSAI, C.S. *ET AL.* (2003). Regulation of the Bacillus subtilis extracytoplasmic function protein sigma(Y) and its target promoters. *Journal of Bacteriology* **185**, 4883-90.

CAUSIER, B. (2004). Studying the interactome with the yeast two-hybrid system and mass spectrometry. *Mass Spectrometry Reviews* **23**, 350-67.

COCHRANE, G., ALDEBERT, P., ALTHORPE, N. *ET AL.* (2006). EMBL Nucleotide Sequence Database: developments in 2005. *Nucleic Acids Research* **34**, D10-5.

COHAN, F.M. (2002). What are bacterial species? *Annual Review of Microbiology* **56**, 457-87.

CORBIN, R.W., PALIY, O., YANG, F. *ET AL.* (2003). Toward a protein profile of Escherichia coli: comparison to its transcription profile. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 9232-7.

CUSICK, M.E., KLITGORD, N., VIDAL, M. and HILL, D.E. (2005). Interactome: gateway into systems biology. *Human Molecular Genetics* **14 Spec No. 2**, R171-81.

DANDEKAR, T., SNEL, B., HUYNEN, M. and BORK, P. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. *Trends in Biochemical Sciences* **23**, 324-8.

DATE, S.V. and MARCOTTE, E.M. (2003). Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nature Biotechnology* **21**, 1055-62.

DATE, S.V. and MARCOTTE, E.M. (2005). Protein function prediction using the Protein Link EXplorer (PLEX). *Bioinformatics* **21**, 2558-9.

DAVIDSON, G.S., HENDRICKSON, B., JOHNSON, D.K., MEYERS, C.E. and WYLIE, B.N. (1998). Knowledge Mining With VxInsight: Discovery Through Interaction. *Journal of Intelligent Information Systems* **11**, 259-285.

DELCHER, A.L., HARMON, D., KASIF, S., WHITE, O. and SALZBERG, S.L. (1999). Improved microbial gene identification with GLIMMER. *Nucleic Acids Research* **27**, 4636-41.

DEN HENGST, C.D., VAN HIJUM, S.A., GEURTS, J.M. *ET AL.* (2005). The Lactococcus lactis CodY regulon: identification of a conserved cis-regulatory element. *Journal of Biological Chemistry* **280**, 34332-42.

DERVYN, E., NOIROT-GROS, M.F., MERVELET, P. *ET AL.* (2004). The bacterial condensin/cohesin-like protein complex acts in DNA repair and regulation of gene expression. *Molecular Microbiology* **51**, 1629-40.

DIODATI, M.E., OSSA, F., CABEROY, N.B. *ET AL.* (2006). Nla18, a key regulatory protein required for normal growth and development of Myxococcus xanthus. *Journal of Bacteriology* **188**, 1733-43.

DWORKIN, M. (1993). *Myxobacteria II*. Washington, DC: ASM Press.

EDDY, S.R. (1998). Profile hidden Markov models. *Bioinformatics* **14**, 755-63.

EDDY, S.R. (2002). Computational genomics of noncoding RNA genes. *Cell* **109**, 137-40.

EHRENREICH, A. (2006). DNA microarray technology for the microbiologist: an overview. *Applied Microbiology and Biotechnology* **73**, 255-73.

EISENBERG, D., MARCOTTE, E.M., XENARIOS, I. and YEATES, T.O. (2000). Protein function in the post-genomic era. *Nature* **405**, 823-6.

ENAULT, F., SUHRE, K., POIROT, O., ABERGEL, C. and CLAVERIE, J.M. (2004). Phydbac2: improved inference of gene function using interactive phylogenomic profiling and chromosomal location analysis. *Nucleic Acids Research* **32**, W336-9.

ENRIGHT, A.J., ILIOPOULOS, I., KYRPIDES, N.C. and OUZOUNIS, C.A. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**, 86-90.

EYMANN, C., DREISBACH, A., ALBRECHT, D. *ET AL.* (2004). A comprehensive proteome map of growing Bacillus subtilis cells. *Proteomics* **4**, 2849-76.

FIELDS, S. (2005). High-throughput two-hybrid analysis. The promise and the peril. *Febs J* **272**, 5391-9.

FIELDS, S. and SONG, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature* **340**, 245-6.

FINN, R.D., MISTRY, J., SCHUSTER-BOCKLER, B. *ET AL.* (2006). Pfam: clans, web tools and services. *Nucleic Acids Research* **34**, D247-51.

FLEISCHMANN, R.D., ADAMS, M.D., WHITE, O. *ET AL.* (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269**, 496-512.

FORMSTECHER, E., ARESTA, S., COLLURA, V. *ET AL.* (2005). Protein interaction mapping: a Drosophila case study. *Genome Research* **15**, 376-84.

FOURNIER, P.E., ROUX, V. and RAOULT, D. (1998). Phylogenetic analysis of spotted fever group rickettsiae by study of the outer surface protein rOmpA. *International Journal*

of Systematic Bacteriology **48 Pt 3**, 839-49.

GALPERIN, M.Y. (2006). A square archaeon, the smallest eukaryote and the largest bacteria. *Environmental Microbiology* **8**, 1683-7.

GALPERIN, M.Y. and KOONIN, E.V. (1998). Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. *In Silico Biology* **1**, 55-67.

GALVEZ, A., MAQUEDA, M., MATRINEZ-BUENO, M. and VALDIVIA, E. (1998). Publication rates reveal trends in microbiological research. *ASM News* **64,** 269-275.

GE, H., LIU, Z., CHURCH, G.M. and VIDAL, M. (2001). Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nature Genetics* **29**, 482-6.

GE, H., WALHOUT, A.J. and VIDAL, M. (2003). Integrating 'omic' information: a bridge between genomics and systems biology. *Trends in Genetics* **19**, 551-60.

GERLT, J.A. and BABBITT, P.C. (2000). Can sequence determine function? *Genome Biology* **1**, REVIEWS0005.

GERSTEIN, M., LAN, N. and JANSEN, R. (2002). Proteomics. Integrating interactomes. *Science* **295**, 284-7.

GERTH, K., JANSEN, R., REIFENSTAHL, G. *ET AL.* (1983). The myxalamids, new antibiotics from Myxococcus xanthus (Myxobacterales). I. Production, physico-chemical and biological properties, and mechanism of action. *Journal of Antibiotics (Tokyo)* **36**, 1150-6.

GIOT, L., BADER, J.S., BROUWER, C. *ET AL.* (2003). A protein interaction map of Drosophila melanogaster. *Science* **302**, 1727-36.

GOGARTEN, J.P., DOOLITTLE, W.F. and LAWRENCE, J.G. (2002). Prokaryotic evolution in light of gene transfer. *Molecular Biology and Evolution* **19**, 2226-38.

GOLDMAN, B.S., NIERMAN, W.C., KAISER, D. *ET AL.* (2006). Evolution of sensory complexity recorded in a myxobacterial genome. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 15200-5.

GOTTESMAN, S. (2005). Micros for microbes: non-coding regulatory RNAs in bacteria. *Trends in Genetics* **21**, 399-404.

HALL, M.C., JORDAN, J.R. and MATSON, S.W. (1998). Evidence for a physical interaction between the Escherichia coli methyl-directed mismatch repair proteins MutL and UvrD. *EMBO Journal* **17**, 1535-41.

HAN, M.J. and LEE, S.Y. (2006). The Escherichia coli proteome: past, present, and future prospects. *Microbiol Mol Biol Rev* **70**, 362-439.

HARTWELL, L.H., HOPFIELD, J.J., LEIBLER, S. and MURRAY, A.W. (1999). From molecular to modular cell biology. *Nature* **402**, C47-52.

HAZBUN, T.R. and FIELDS, S. (2001). Networking proteins in yeast. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 4277-8.

HERRING, C.D., RAFFAELLE, M., ALLEN, T.E. *ET AL.* (2005). Immobilization of Escherichia coli RNA polymerase and location of binding sites by use of chromatin immunoprecipitation and microarrays. *Journal of Bacteriology* **187**, 6166-74.

HUANG, J., LIH, C.J., PAN, K.H. and COHEN, S.N. (2001). Global analysis of growth phase responsive gene expression and regulation of antibiotic biosynthetic pathways in Streptomyces coelicolor using DNA microarrays. *Genes & Development* **15**, 3183-92.

HUGENHOLTZ, P. (2002). Exploring prokaryotic diversity in the genomic era. *Genome*

Biology **3**, REVIEWS0003.

HWANG, D., RUST, A.G., RAMSEY, S. *ET AL* (2005a). A data integration methodology for systems biology. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 17296-301.

HWANG, D., SMITH, J.J., LESLIE, D.M. *ET AL.* (2005b). A data integration methodology for systems biology: experimental verification. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 17302-7.

IRIZARRY, R.A., WARREN, D., SPENCER, F. *ET AL.* (2005). Multiple-laboratory comparison of microarray platforms. *Nature Methods* **2**, 345-50.

ITO, T., CHIBA, T., OZAWA, R. *ET AL.* (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 4569-74.

ITOH, T., TAKEMOTO, K., MORI, H. and GOJOBORI, T. (1999). Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. *Molecular Biology and Evolution* **16**, 332-46.

JACOB, F. and MONOD, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology* **3**, 318-56.

JAFFE, J.D., STANGE-THOMANN, N., SMITH, C. *ET AL.* (2004). The complete genome and proteome of Mycoplasma mobile. *Genome Research* **14**, 1447-61.

JAKOBSEN, J.S., JELSBAK, L., JELSBAK, L. *ET AL.* (2004). Sigma54 enhancer binding proteins and Myxococcus xanthus fruiting body development. *Journal of Bacteriology* **186**, 4361-8.

JANGA, S.C., COLLADO-VIDES, J. and MORENO-HAGELSIEB, G. (2005). Nebulon: a system for the inference of functional relationships of gene products from the rearrangement of predicted operons. *Nucleic Acids Research* **33**, 2521-30.

JANSEN, R., GREENBAUM, D. and GERSTEIN, M. (2002). Relating whole-genome expression data with protein-protein interactions. *Genome Research* **12**, 37-46.

JELSBAK, L., GIVSKOV, M. and KAISER, D. (2005). Enhancer-binding proteins with a forkhead-associated domain and the sigma54 regulon in Myxococcus xanthus fruiting body development. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 3010-5.

JESSANI, N. and CRAVATT, B.F. (2004). The development and application of methods for activity-based protein profiling. *Current Opinion in Chemical Biology* **8**, 54-9.

JOYCE, A.R. and PALSSON, B.O. (2006). The model organism as a system: integrating 'omics' data sets. *Nature Reviews Molecular Cell Biology* **7**, 198-210.

JOYCE, A.R., REED, J.L., WHITE, A. *ET AL.* (2006). Experimental and computational assessment of conditionally essential genes in E. coli. *Journal of Bacteriology* **188**, 8259-71.

KAISER, D. (1986). Control of multicellular development: Dictyostelium and Myxococcus. *Annual Review of Genetics* **20**, 539-66.

KAISER, D. (2004). Signaling in myxobacteria. *Annual Review of Microbiology* **58**, 75-98.

KANEHISA, M., GOTO, S., KAWASHIMA, S., OKUNO, Y. and HATTORI, M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Research* **32**, D277-80.

KARAOZ, U., MURALI, T.M., LETOVSKY, S. *ET AL.* (2004). Whole-genome annotation by using evidence integration in functional-linkage networks. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 2888-93.

KIM, S.K., LUND, J., KIRALY, M. *ET AL.* (2001). A gene expression map for Caenorhabditis elegans. *Science* **293**, 2087-92.

KITANO, H. (2002). Systems biology: a brief overview. *Science* **295**, 1662-4.

KESELER, I.M., GAMA-CASTRO, S., PERALTA-GIL, M. *ET AL.* EcoCyc: a comprehensive database resource for Escherichia coli. *Nucleic Acids Research* **34**, D394-7.

KOLKER, E., PICONE, A.F., GALPERIN, M.Y. *ET AL.* (2005). Global profiling of Shewanella oneidensis MR-1: expression of hypothetical genes and improved functional annotations. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 2099-104.

KOONIN, E.V., MAKAROVA, K.S. and ARAVIND, L. (2001). Horizontal gene transfer in prokaryotes: quantification and classification. *Annual Review of Microbiology* **55**, 709-42.

LARKIN, J.E., FRANK, B.C., GAVRAS, H., SULTANA, R. and QUACKENBUSH, J. (2005). Independence and reproducibility across microarray platforms. *Nature Methods* **2**, 337-44.

LATHE, W.C., 3RD, SNEL, B. and BORK, P. (2000). Gene context conservation of a higher order than operons. *Trends in Biochemical Sciences* **25**, 474-9.

LIOLIOS, K., TAVERNARAKIS, N., HUGENHOLTZ, P. and KYRPIDES, N.C. (2006). The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. *Nucleic Acids Research* **34**, D332-4.

LIPTON, M.S., PASA-TOLIC, L., ANDERSON, G.A. *ET AL.* (2002). Global analysis of the Deinococcus radiodurans proteome by using accurate mass tags. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 11049-54.

LIU, X., NG, C. and FERENCI, T. (2000). Global adaptations resulting from high population densities in Escherichia coli cultures. *Journal of Bacteriology* **182**, 4158-64.

LIU, Z., JACOBS, M., SCHAFF, D.A., MCCULLEN, C.A. and BINNS, A.N. (2001). ChvD, a chromosomally encoded ATP-binding cassette transporter-homologous protein involved in regulation of virulence gene expression in Agrobacterium tumefaciens. *Journal of Bacteriology* **183**, 3310-7.

LU, L.J., XIA, Y., PACCANARO, A., YU, H. and GERSTEIN, M. (2005). Assessing the limits of genomic data integration for predicting protein networks. *Genome Research* **15**, 945-53.

LUCCHINI, S., THOMPSON, A. and HINTON, J.C. (2001). Microarrays for microbiologists. *Microbiology* **147**, 1403-14.

LUPPENS, S.B. and TEN CATE, J.M. (2005). Effect of biofilm model, mode of growth, and strain on streptococcusmutans protein expression as determined by two-dimensional difference gel electrophoresis. *Journal of Proteome Research* **4**, 232-7.

MALEK, J.A., WIERZBOWSKI, J.M., TAO, W. *ET AL.* (2004). Protein interaction mapping on a functional shotgun sequence of Rickettsia sibirica. *Nucleic Acids Research* **32**, 1059-64.

MARCOTTE, E. and DATE, S. (2001). Exploiting big biology: integrating large-scale biological data for function inference. *Briefings in Bioinformatics* **2**, 363-74.

MARCOTTE, E.M., PELLEGRINI, M., NG, H.L. *ET AL.* (1999). Detecting protein function and protein-protein interactions from genome sequences. *Science* **285**, 751-3.

MARCOTTE, E.M., PELLEGRINI, M., THOMPSON, M.J., YEATES, T.O. and EISENBERG, D. (1999). A combined algorithm for genome-wide prediction of protein function. *Nature* **402**, 83-6.

MATHE, C., SAGOT, M.F., SCHIEX, T. and ROUZE, P. (2002). Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Research* **30**, 4103-17.

MCCAUGHREAN, M., REID, N., TINNEY, C. *ET AL.* (2001). What Is a Planet? *Science* **291**, 1487-1488.

MILLER, J. and STAGLJAR, I. (2004). Using the Yeast Two-Hybrid System to Identify Interacting Proteins. In: *Protein-protein interactions: Methods and Applications*. Eds. Fu, H., pp 247-262. Totowa, NJ: Humana Press.

MORENO, E. (2002). In search of a bacterial species definition. *Revista de Biologia Tropical* **50**, 803-21.

MORETT, E. and SEGOVIA, L. (1993). The sigma 54 bacterial enhancer-binding protein family: mechanism of action and phylogenetic relationship of their functional domains. *Journal of Bacteriology* **175**, 6067-74.

MUKHOPADHYAY, A., HE, Z., ALM, E.J. *ET AL.* (2006). Salt stress in Desulfovibrio vulgaris Hildenborough: an integrated genomics approach. *Journal of Bacteriology* **188**, 4068-78.

MUSHEGIAN, A.R. and KOONIN, E.V. (1996). Gene order is not conserved in bacterial evolution. *Trends in Genetics* **12**, 289-90.

NOIROT-GROS, M.F., SOULTANAS, P., WIGLEY, D.B. *ET AL.* (2002). The beta-propeller protein YxaL increases the processivity of the PcrA helicase. *Molecular Genetics and Genomics* **267**, 391-400.

OCHMAN, H., LAWRENCE, J.G. and GROISMAN, E.A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299-304.

OCHMAN, H., LERAT, E. and DAUBIN, V. (2005). Examining bacterial species under the specter of gene transfer and exchange. *Proceedings of the National Academy of Sciences of the United States of America* **102 Suppl 1**, 6595-9.

OHTA, N. and NEWTON, A. (2003). The core dimerization domains of histidine kinases contain recognition specificity for the cognate response regulator. *Journal of Bacteriology* **185**, 4424-31.

OKUBO, K., SUGAWARA, H., GOJOBORI, T. and TATENO, Y. (2006). DDBJ in preparation for overview of research activities behind data submissions. *Nucleic Acids Research* **34**, D6-9.

OLIVER, S. (2000). Guilt-by-association goes global. *Nature* **403**, 601-3.

OVERBEEK, R., FONSTEIN, M., D'SOUZA, M., PUSCH, G.D. and MALTSEV, N. (1999). The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 2896-901.

PARRISH, J.R., GULYAS, K.D. and FINLEY, R.L., JR. (2006). Yeast two-hybrid contributions to interactome mapping. *Current Opinion in Biotechnology* **17**, 387-93.

PELLEGRINI, M., MARCOTTE, E.M., THOMPSON, M.J., EISENBERG, D. and YEATES, T.O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 4285-8.

PHILLIPS, C.I. and BOGYO, M. (2005). Proteomics meets microbiology: technical advances in the global mapping of protein expression and function. *Cellular Microbiology* **7**, 1061-76.

PHIZICKY, E.M. and FIELDS, S. (1995). Protein-protein interactions: methods for detection and analysis. *Microbiological Reviews* **59**, 94-123.

RAIN, J.C., SELIG, L., DE REUSE, H. *ET AL.* (2001). The protein-protein interaction map of

Helicobacter pylori. *Nature* **409**, 211-5.

REICHENBACH, H. (2001). Myxobacteria, producers of novel bioactive substances. *Journal of Industrial Microbiology & Biotechnology* **27**, 149-56.

REICHENBACH, H. and HOFLE, G. (1993). Biologically active secondary metabolites from myxobacteria. *Biotechnology Advances* **11**, 219-77.

REVA, O.N., WEINEL, C., WEINEL, M. *ET AL.* (2006). Functional genomics of stress response in Pseudomonas putida KT2440. *Journal of Bacteriology* **188**, 4079-92.

RHODIUS, V., VAN DYK, T.K., GROSS, C. and LAROSSA, R.A. (2002). Impact of genomic technologies on studies of bacterial gene expression. *Annual Review of Microbiology* **56**, 599-624.

SAITO, N., ROBERT, M., KITAMURA, S. *ET AL.* (2006). Metabolomics approach for enzyme discovery. *Journal of Proteome Research* **5**, 1979-87.

SALMON, K., HUNG, S.P., MEKJIAN, K. *ET AL.* (2003). Global gene expression profiling in Escherichia coli K12. The effects of oxygen availability and FNR. *Journal of Biological Chemistry* **278**, 29837-55.

SAUER, K., CAMPER, A.K., EHRLICH, G.D., COSTERTON, J.W. and DAVIES, D.G. (2002). Pseudomonas aeruginosa displays multiple phenotypes during development as a biofilm. *Journal of Bacteriology* **184**, 1140-54.

SCHATTNER, P., BROOKS, A.N. and LOWE, T.M. (2005). The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Research* **33**, W686-9.

SCHLEY, C., ALTMEYER, M.O., SWART, R., MULLER, R. and HUBER, C.G. (2006). Proteome Analysis of Myxococcus xanthus by Off-Line Two-Dimensional Chromatographic Separation Using Monolithic Poly-(styrene-divinylbenzene) Columns Combined with Ion-Trap Tandem Mass Spectrometry. *Journal of Proteome Research* **5**, 2760-8.

SHEPPARD, S.S. (2006). Solar system: a planet more, a planet less? *Nature* **439**, 541-2.

SHIMKETS, L.J. (1999). Intercellular signaling during fruiting-body development of Myxococcus xanthus. *Annual Review of Microbiology* **53**, 525-49.

SIMONSON, A.B., SERVIN, J.A., SKOPHAMMER, R.G. *ET AL.* (2005). Decoding the genomic tree of life. *Proceedings of the National Academy of Sciences of the United States of America* **102 Suppl 1**, 6608-13.

SIMUNOVIC, V., ZAPP, J., RACHID, S. *ET AL.* (2006). Myxovirescin A biosynthesis is directed by hybrid polyketide synthases/nonribosomal peptide synthetase, 3-hydroxy-3-methylglutaryl-CoA synthases, and trans-acting acyltransferases. *Chembiochem* **7**, 1206-20.

SOUTHERN, E.M. (2001). DNA Arrays. In: *DNA Arrays: Methods and Protocols*. Eds. Rampal, J.S., pp 1-15. Totowa, NJ: Humana Press.

SPRINZAK, E., SATTATH, S. and MARGALIT, H. (2003). How reliable are experimental protein-protein interaction data? *Journal of Molecular Biology* **327**, 919-23.

SRINIVASAN, B.S., CABEROY, N.B., SUEN, G. *ET AL.* (2005). Functional genome annotation through phylogenomic mapping. *Nature Biotechnology* **23**, 691-8.

STEIN, L. (2001). Genome annotation: from sequence to biology. *Nature Reviews Genetics* **2**, 493-503.

STEYN, A.J., COLLINS, D.M., HONDALUS, M.K. *ET AL.* (2002). Mycobacterium tuberculosis WhiB3 interacts with RpoV to affect host survival but is dispensable for in vivo growth. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 3147-52.

STRONG, M., GRAEBER, T.G., BEEBY, M. *ET AL.* (2003a). Visualization and interpretation of protein networks in Mycobacterium tuberculosis based on hierarchical clustering of genome-wide functional linkage maps. *Nucleic Acids Research* **31**, 7099-109.

STRONG, M., MALLICK, P., PELLEGRINI, M., THOMPSON, M.J. and EISENBERG, D. (2003b). Inference of protein function and protein linkages in Mycobacterium tuberculosis based on prokaryotic genome organization: a combined computational approach. *Genome Biology* **4**, R59.

SUEN, G., JAKOBSEN, J.S., GOLDMAN, B.S. *ET AL.* (2006). Bacterial Post-Genomics: The Promise and Peril of Systems Biology. *Journal of Bacteriology* **188**, 7999-8003

SUHRE, K. and CLAVERIE, J.M. (2004). FusionDB: a database for in-depth analysis of prokaryotic gene fusion events. *Nucleic Acids Research* **32**, D273-6.

SUTER, B., AUERBACH, D. and STAGLJAR, I. (2006). Yeast-based functional genomics and proteomics technologies: the first 15 years and beyond. *Biotechniques* **40**, 625-44.

TAO, H., BAUSCH, C., RICHMOND, C., BLATTNER, F.R. and CONWAY, T. (1999). Functional genomics: expression analysis of Escherichia coli growing on minimal and rich media. *Journal of Bacteriology* **181**, 6425-40.

TAOKA, M., YAMAUCHI, Y., SHINKAWA, T. *ET AL.* (2004). Only a small subset of the horizontally transferred chromosomal genes in Escherichia coli are translated into proteins. *Molecular & Cellular Proteomics* **3**, 780-7.

TATUSOV, R.L., GALPERIN, M.Y., NATALE, D.A. and KOONIN, E.V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research* **28**, 33-6.

THOMAS, C.M. and NIELSEN, K.M. (2005). Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature Reviews Microbiology* **3**, 711-21.

VIDAL, M. (2001). A biological atlas of functional maps. *Cell* **104**, 333-9.

VILAIN, S. and BROZEL, V.S. (2006). Multivariate approach to comparing whole-cell proteomes of Bacillus cereus indicates a biofilm-specific proteome. *Journal of Proteome Research* **5**, 1924-30.

VOGEL, J. and SHARMA, C.M. (2005). How to find small non-coding RNAs in bacteria. *Biological Chemistry* **386**, 1219-38.

VON MERING, C., JENSEN, L.J., SNEL, B. *ET AL.* (2005). STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research* **33**, D433-7.

VON MERING, C., KRAUSE, R., SNEL, B. *ET AL.* (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**, 399-403.

WALHOUT, A.J., REBOUL, J., SHTANKO, O. *ET AL.* (2002). Integrating interactome, phenome, and transcriptome mapping data for the C. elegans germline. *Current Biology* **12**, 1952-8.

WANG, Z., CHEN, Y. and LI, Y. (2004). A brief review of computational gene prediction methods. *Genomics Proteomics Bioinformatics* **2**, 216-21.

WASINGER, V.C., POLLACK, J.D. and HUMPHERY-SMITH, I. (2000). The proteome of Mycoplasma genitalium. Chaps-soluble component. *European Journal of Biochemistry* **267**, 1571-82.

WILSON, M., DERISI, J., KRISTENSEN, H.H. *ET AL.* (1999). Exploring drug-induced alterations in gene expression in Mycobacterium tuberculosis by microarray hybridization. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 12833-8.

WINKLER, W.C. (2005). Riboswitches and the role of noncoding RNAs in bacterial metabolic control. *Current Opinion in Chemical Biology* **9**, 594-602.

WOLFF, S., ANTELMANN, H., ALBRECHT, D. *ET AL.* (2006). Towards the entire proteome of the model bacterium Bacillus subtilis by gel-based and gel-free approaches. *Journal of chromatography. B, Analytical technologies in the biomedical and life sciences* **849**, 129-40.

WU, H., SU, Z., MAO, F., OLMAN, V. and XU, Y. (2005). Prediction of functional modules based on comparative genome analysis and Gene Ontology application. *Nucleic Acids Research* **33**, 2822-37.

YANG, R., BARTLE, S., OTTO, R. *ET AL.* (2004). AglZ is a filament-forming coiled-coil protein required for adventurous gliding motility of Myxococcus xanthus. *Journal of Bacteriology* **186**, 6168-78.

YE, R.W., TAO, W., BEDZYK, L. *ET AL.* (2000). Global gene expression profiles of Bacillus subtilis grown under anaerobic conditions. *Journal of Bacteriology* **182**, 4458-65.

ZHU, H., BILGIN, M. and SNYDER, M. (2003). Proteomics. *Annual Review of Biochemistry* **72**, 783-812.