3

HIGH-THROUGHPUT TECHNOLOGIES AND FUNCTIONAL GENOMICS

Xiu-Feng Wan[1] and Dorothea K. Thompson[2]

[1]*Department of Microbiology, Miami University, Oxford, Ohio 45056*

[2]*Department of Biological Sciences, Purdue University, West Lafayette, Indiana 47907*

3.1 INTRODUCTION

Traditional biological studies generally target the structure and function of a specific gene or protein. Generally, a specific hypothesis is generated for a specific biological problem and then tested by an experimental design (Fig. 3-1). In the 1990s, the first high-throughput technologies were invented for biological studies and included genome sequencing, proteomics, DNA chips, and protein chips. These technological advancements have created a new field of bioinformatics and computational biology. The combination of bioinformatics and high-throughput technologies has re-shaped traditional biological studies; through these technologies, biologists will be able to generate better biological hypotheses, and also streamline the traditional methods, which has proven to be much more efficient than traditional biological study (Fig. 3-1). As these technologies become increasingly more mature and economically feasible, more and more laboratories are using these methods. In this chapter, we briefly introduce four high-throughput technologies and then focus on the details of three technologies: genomic sequencing, proteomics, and DNA and protein chip technologies. We also illustrate chip technologies using two applications.

Systems Biology and Synthetic Biology Edited by Pengcheng Fu and Sven Panke

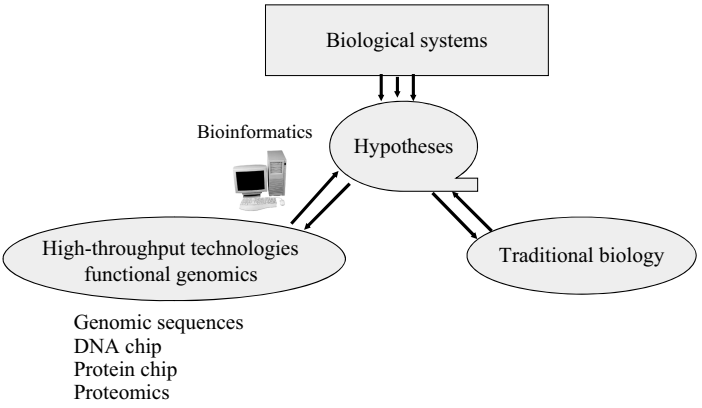

Figure 3-1 New scientific technologies for high-throughput measurements and functional genomics.

3.2 HIGH-THROUGHPUT TECHNOLOGIES

3.2.1 Genomic Sequencing

Genomic sequencing, which ultimately revolutionized the field of biology, was invented by Nobel Laureate Frederick Sanger in 1981 [1]. This technique involves the separation of fluorescently labeled DNA fragments according to length on *p*oly*a*crylamide *g*els via *e*lectrophoresis (PAGE). Through automation, each sequencing run can yield 500 bp to 1 kb of sequence data with a modern sequence machine. DNA-sequencing technology is another milestone in understanding the evolution, structure, and function of biological systems since the discovery of the DNA structure by Watson and Crick in 1953. However, additional complementary technologies are required for sequencing complete genomes since genomic sequences may be as long as billions of bases (Table 3-1). For instance, the human genome is about 3.3 billion bases. A typical bacterial genome ranges from several hundred kilobases to more than 10 million bases. The model bacterium *Escherichia coli* K12, for example, has a genome comprising about 4.67 million bases.

In 1983, Frederick Sanger invented the shotgun-sequencing strategy and sequenced the first complete genome, bacteriophage λ, which has 48,502 bases [2]. Without parallel advances in super computational techniques, the applications of shotgun

Table 3-1 Wide ranges in genome size

Species	Genome Size (Mb)	Species	Genome Size (Mb)
HIV	0.0097	*S. cerevisiae*	11.72
SARS-CoV	0.030	*C. elegans*	~100
Mycoplasma genitalium	0.59	*A. thaliana*	~125
Escherichia coli	4.67	*Homo sapien*	~3,300

genome sequencing were not fully realized until 12 years later. In 1995, through the use of super computational facilities, the *Haemophilus influenzae* genomic sequence was published by J. Craig Venter from the Institute for Genomic Research (TIGR) and Nobel Laureate Hamilton Smith of Johns Hopkins University [3]. The human genome sequence project was launched in 1990 and was eventually completed in the year 2003 with a large collaboration of international effort by the International Sequencing Consortium (http://www.intlgenome.org/). It should be noted that the contribution of computational biology was crucial to the successful completion of the human genome sequence project.

Figure 3-2 shows a simplified procedure for the shotgun genome-sequencing strategy. To sequence a large sequence, shotgun genomic sequencing first breaks the sequence randomly several times into small fragments of about 1,500 bases by enzymes or physical shearing and then sequences these individual fragments. The computer is able to connect the sequences based on the overlapping ends between these sequences. The size of this large sequence will be less than 150 kb. This is because, for a large genome, the sequence needs to be separated into smaller fragments of about 150 kb, each of which will be cloned into bacterial artificial chromosome (BAC) vectors. Each of these large fragments is called a contig and can be sequenced using the shotgun-sequencing method. By using BACs, these contigs can be mapped, as BAC records the positions where the contigs come from the genomic sequence.

3.2.2 DNA Microarray

The DNA microarray, also known as a DNA chip, contains thousands of arrayed probes, each composed of a short oligonucleotide or cDNA fragment. The invention of DNA chip technology has made it possible to study the functions of thousands of genes at the same time, allowing for biological study in a more systematic way. The

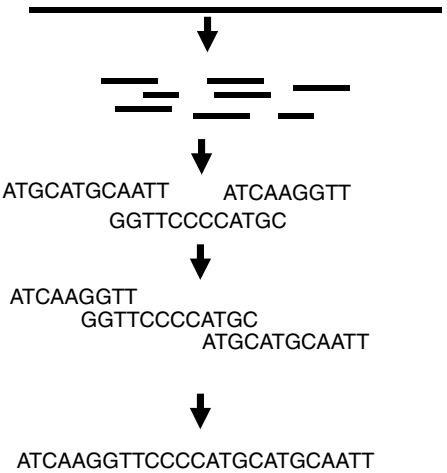

Figure 3-2 Simplified shotgun genome sequencing strategy.

fundamental mechanism underlying the DNA chip methodology is nucleotide hybridization, which had been previously deduced. However, the most important concept for DNA chips (as well as protein chips) is that these technologies facilitate the automation of evenly spotted DNA molecules onto a surface, which will allow for quantification of the hybridization signal. Thus, the first DNA arrays originated from the development in the late 1980s of robotic devices (gridding robots) that make it possible to array bacterial colonies in compact and regular patterns [4]. The original DNA chip had approximately 10,000 spots on a $22 \times 22\,\text{cm}^2$ surface. This array allowed for rapid genomic library scanning. The functional genomics for expression analysis with quantitative acquisition of hybridization signals was first reported in 1992 [5]. This technology was based in part on integrated mapping and sequencing analysis of genomes.

The massive DNA sequences generated by shotgun sequencing have given us an opportunity as well as a challenge to study the evolution, structure, and function of these genes. Most notably, the complete genomic sequences allow us to evaluate expression patterns on a genomic scale. DNA chip technologies and functional genomics have been applied widely in many different fields, such as pathogenesis, drug discovery, cancer research, cell development, cell structure, agricultural seed selection, and even in the environmental community study [6–18]. For instance, in drug discovery, functional genomics can be applied in basic research and target discovery, biomarker determination, pharmacology, toxicogenomics, target selectivity, development of prognostic tests, and disease subclass determination [6]. Further details regarding DNA chips and functional genomics are discussed in Section 3.3.

3.2.3 Protein Microarrays

DNA microarrays are used to monitor global gene expression levels based on intracellular RNA concentration. However, the corresponding protein expression may be different from RNA abundance due to gene regulation at the translational level and alternative gene splicing. Protein chip technology was invented for this purpose. Different from DNA chips, protein chips have been used to detect the quantity of specific proteins by measuring signals from the interactions between protein versus protein and protein versus antibody. The target molecules can not only be traditional protein molecules but also be other types of molecules, such as artificial proteins [19], RNA or DNA aptamers [20], allosteric ribozymes [21], peptides, and other small molecules [22,23]. With these extensions, protein chips can be applied to monitor the interactions between protein versus ligand, protein versus drug, enzyme versus substrate, and so on.

Haab et al. [24] printed a set of 115 antibody–antigen pairs to evaluate the use of protein microarrays for specific detection and quantification of multiple proteins in complex mixtures. About 50 percent of the arrayed antigens and 20 percent of the arrayed antibodies provided specific and accurate measurements of their cognate ligands at or below concentrations of 0.34 and 1.6 μg/mL, respectively. Their studies suggest that protein microarrays can provide a practical means to characterize patterns of variation in hundreds of thousands of different proteins in clinical or research

applications. Some companies have developed antibody arrays for both investigational usage as well as clinical use for monitoring allergies and small therapeutic drug monitoring. Similar to DNA chips, protein chips are able to perform thousands of reactions in parallel. Thus, by using a specific antibody, we will be able to screen for the presence of a specific protein from a specific reaction. Protein chips have become an important proteomics technology in addition to mass spectrometry and two-dimensional gel electrophoresis, both of which are however less sensitive than protein chip technology.

The original idea for protein chips followed from miniaturized immunoassay technology. In the 1980s, the development of ELISA introduced the concept of ambient analysis, which is able to quantify the antigen–antibody reaction through a specific enzyme-labeling assay. Similar to the DNA chip, development of this technology was accelerated by the genome project and improved technologies in recombinant proteins. Since most proteins used for protein arrays are made by recombination, the protein array would be able to be connected with DNA sequence and protein structural analysis. The functional analysis of the DNA-coding genes can reflect their functions.

Similar to the DNA chip, the protein chip uses covalent interactions to immobilize protein molecules onto solid surfaces by randomly conjugating the lysine residues on proteins to amine-reactive surfaces. In many cases, the recombination proteins are preferred since amino- or carboxy-terminal tags can be introduced so that the protein's functional sites can be away from the immobilization surface, which can increase the sensitivity of protein chips via the reduction of steric hindrance. The printing technologies for protein chips are similar to those for DNA chips, described in Section 3.3. However, the challenge for printing processes is how to prevent dehydration of the protein spots. Improvement in this area seems to be needed for further development [22].

Protein chips have become an important tool for biological study. Protein chips are mainly applied in micro-immunoassay, in which arrays of different capture antibodies are immobilized and subsequently exposed to a biological sample. These types of protein chips can be used for diagnostics as well as protein-profiling analysis. Specific antibodies can be immobilized on the chip to monitor the protein expression levels in a tissue or a cell. The parallel analyses would be able to monitor the protein-profiling changes for a patient and to determine the disease status or monitor the treatment or therapy through a minimum of biopsy material. In reverse immunoassays, the purified small antigens can be immobilized on the chip so that the specific antibody responses in the blood or local tissues can be evaluated. The reverse immunoassay can be used for diagnosis of various autoimmune diseases [25] or allergies [26]. These types of analyses can be used for examining binding receptor properties as well as antibody cross-reactivity and specificity.

The protein array has a very promising application for drug scanning since it directly monitors the interaction between drug and a target protein. Protein chips may be used in binding/screening assays for other small molecules, such as ligands, RNA–DNA molecules, and some artificial proteins. They can also be used for isolation of individual candidate molecules from a large pool. For instance, protein chips may be

used for studying protein–DNA interactions, especially for promoter analysis, for investigating enzyme activity with different substrates, and for epitope mapping.

3.3 DNA CHIPS AND FUNCTIONAL GENOMICS

In this section, we first discuss the details about DNA chip manufacturing technologies, focusing primarily on how the probe is printed on the slides. Then we discuss the probe design, sample labeling and hybridization, scanning and image analysis, data analysis, experimental design and data interpretation, challenges of DNA chips, and applications of DNA chips.

3.3.1 Microarray Manufacturing Technologies

Current fabrication technologies for DNA microarrays can be grouped into photolithography, mechanical microspotting, or ink-jet ejection [27]. For a photolithography array, the oligonucleotide probe is synthesized directly onto a solid surface (e.g., the Affymetrix and NimbleGene arrays) based on a combination of chemistry and photolithographic methods [28]. To produce the array, the reactive amine groups from a silane reagent are attached to a glass or fused silica surface, and then the amine groups are modified via methylnitropoperonyloxycarbonyl (MeNPOC) photoprotection. A single base can then be added to the hydroxyl groups of these MeNPOC using a standard phosphoramidite DNA synthesis method after exposure to light. The photoprotection and nucleotide insertion are repeated to obtain a desired probe [27,29]. The lengths of these probes are generally 20–25 bases. The photolithographic array can have a much higher probe density than other types of arrays. Affymetrix chips can contain about 250,000 oligonucleotides in an area of 1 cm^2 while the spotted cDNA array generally only has about 1,000 oligonucleotides in the same area. This feature offers an important advantage for the Affymetrix array over spotted arrays, which have much lower probe densities. In addition, the Affymetrix system is more stable and reproducible, since it lacks the problems associated with printing spotted arrays. However, the current price of Affymetrix arrays is still too high to be widely accepted as a biological tool.

A mechanically microspotted array is called a spotted array. This type of array utilizes pins, tweezers, or capillaries to print the molecules onto glass or other solid surfaces. The molecules can be oligonucleotides, genomic DNA, or polymerase chain reaction (PCR) fragments (DNA or cDNA). For protein chips, we can even print antibodies, small drug molecules, and other small molecules. The printing process is generally achieved by a robot monitored by a computer. Compared with the array constructed on the basis of photolithography, a spotted array is more economical as well as easily implemented. In addition, the spotted array can have many more applications than the photolithography array since the array for the latter is limited to short oligonucleotides. However, preparation of the printing material and the printing process require considerable control, as the printing quality will directly affect the analysis.

Similar to a spotted array, the ink-jet ejection array prints the molecules to the solid surface by ejecting the sample from the print head. Different from the spotted array, the print head during printing does not contact the slides, which can reduce the probability of contamination. Currently, two types of noncontact ink-jet print technologies, piezoelectric pumps and syringe-solenoid, are used for printing microarrays. Similar to the spotted array, the ink-jet ejection array can print various molecules on a slide. On the other hand, the ink-jet ejection array prints at even lower densities than the spotted array.

3.3.2 Probe Design and DNA/cDNA Synthesis

Based on the DNA molecules on the slides, DNA chips can be categorized as either DNA/cDNA microarrays or oligonucleotide microarrays. Generally, the DNA fragment on DNA/cDNA microarrays is synthesized by the polymerase chain reaction while oligonucleotides are synthesized directly by machine. To synthesize the DNA fragment or cDNA primers, we need to design unique primers, which are generally 20–28 bases long. For genes shorter than 1,000 bases, the PCR-amplified fragments should be as long as possible. For genes longer than 1,000-bases, the optimal amplified fragments should be within the range of 500–1,200 bases. Xu et al. [30] developed PRIMEGENS for primer design for cDNA amplification. PRIMEGENS finds the unique fragments from a group of gene fragments or genes in a complete genome, and then applies the Primer3 algorithm [31] to design the left and right primers for each unique fragment. The user can change the primer specification based on their PCR requirement. The biggest challenge for production of the DNA or cDNA array is that occasionally PCR amplification may not be able to generate an expected yield for a given gene. Since we generally perform the reaction in 96- or 384-well plates, one may have to amplify individual genes separately. In addition, for complete genomic analysis, it is difficult to ensure complete coverage, due to the cross-hybridization between PCR fragments on the DNA/cDNA array. Furthermore, sample contamination or mishandling during amplification may generate other problems.

Due to the laborious processes for producing DNA/cDNA arrays, many laboratories are utilizing the oligo array. It should be emphasized that oligonucleotide design is not a trivial process. A program for designing optimal probes will need: (1) to minimize hybridization free energy for the target gene and maximize hybridization energy for all other genes, yet the hybridization energy depends on the concentration of the genes, which is unknown; (2) to avoid secondary structure; (3) to consider both strands of the genome as well as the cross-hybridization of the coding region and noncoding region. Many oligo design algorithms and software packages have been developed during the last several years: ProbeSelect [32], PROBEmer [33], CommonOligo [34], Oligo Design [35], Picky [36], OligoPicker [37], OligoArray [38], ROSO [39], and GoArrays [40]. Most of these methods are for a complete genome array. ProbeSelect [32] is one of the most popular methods used for oligo design for complete genome arrays. ProbeSelect first makes a suffix array of the coding sequences from a whole genome and then builds a sequence landscape for every gene based on the sequence suffix array. Based on sequence

features and the sequence word rank values, ProbeSelect chooses probe candidates and then searches for matching sequences in the whole genome, allowing for a certain number of mismatches. After locating match sequence positions in all genes, ProbeSelect calculates the free energy and melting temperature for each valid target sequence. Finally, ProbeSelect matches sequences that have stable hybridization structures with a probe based on free-energy data and maintains high discrimination against other targets in the genome.

For environmental functional genomics, it will be even more challenging to design specific probes due to the high similarity between genes. Some algorithms have been designed for environmental community study [41,42]. The hierarchical probe design (HPD) program is an oligo design program especially suited for long oligo design, allowing for analyses of functional gene diversity in environmental samples [41]. HPD designs both sequence-specific probes and hierarchical cluster-specific probes from sequences of a conserved functional gene based on the clustering tree of the genes.

In general, DNA arrays have two advantages over oligo arrays: (1) DNA arrays have a higher sensitivity; and (2) DNA arrays do not need detailed sequence information; thus, DNA arrays are especially useful for environmental community study for which we generally do not know the exact sequence information. However, oligo arrays have two distinct advantages over DNA arrays: (1) oligo arrays have reduced cross-hybridization, thus providing higher specificity; and (2) unlike DNA arrays, oligo arrays do not require the intensive labor involved in PCR amplification and DNA purification. Oligo arrays are especially popular as the cost of custom array fabrication is steadily declining.

3.3.3 Sample Labeling and Hybridization

Sample labeling can be categorized as either direct or indirect labeling. The direct-labeling approach directly incorporates the fluorescent tags into the nucleic acid when preparing the hybridization samples. The fluorescent tags may be present in labeled nucleotides (e.g., Cy3- or Cy5-dCTP) or PCR primers. PCR and reverse transcription (RT)-PCR are common approaches to synthesize the labeled samples for hybridization. To detect the mRNA concentration, we can use RT-PCR to incorporate fluorescently labeled nucleotides into the transcribed cDNA during first-strand cDNA synthesis. Alternatively, mRNA can be amplified by 1,000–10,000-fold using T7 polymerase to obtain antisense mRNA (aRNA). The aRNA is then reverse-transcribed to obtain labeled cDNA [43]. One of the advantages of the T7 polymerase-based amplification method over other methods is that because amplification is a linear process, all mRNAs are amplified almost equally. Another advantage is that mRNA can be easily labeled with reverse transcriptase, which incorporates fluorescent tags much more readily than DNA polymerase [27].

The indirect labeling approach labels the sample with fluorescence after hybridization. To label the samples, indirect labeling requires epitope insertion into the target samples during cDNA synthesis. After hybridization, the epitopes can be bound by specific proteins to produce the signal. Biotin is one of the commonly used epitopes, which can be stained by a fluorescent streptavidin–phycoerythrin conjugate and

detected via laser [44]. Some other types of indirect hybridizations are discussed by Zhou and Thompson [27].

After labeling, the sample will be hybridized with the probes on the slide. Before hybridization, the slide requires postprocessing, which will use ultraviolet (UV) radiation or heat to cross-link probes to the slide. For example, postprocessing can be done by exposing the slides to 120 mJ/cm^2 using a UV cross-linker or by baking the printed slides for 80 min at 80°C in a drying oven. Similar to traditional membrane-based hybridization, the microarray will also need prehybridization to reduce non-specific binding. The unbound DNA on the slides can be washed away during prehybridization to reduce the competition of unbound DNA for the labeled samples.

After postprocessing and prehybridization, the microarray is hybridized with labeled samples at a certain temperature, generally 42°C to 50°C, for a period of time (overnight to several days). A key to successful hybridization is that the hybridization solution needs to evenly cover the slide. After hybridization, the slides need to be washed to eliminate unbound samples.

3.3.4 Scanning and Image Analysis

The next step after hybridization is quantification of the hybridization signal from the slides. The scanning devices are generally categorized into two types: the confocal scanning microscope and CCD camera. In general, a confocal scanner uses laser excitation of a small region of the glass slide ($\sim$100 μm^2), and the entire array image is acquired by moving the glass slide, the confocal lens, or both across the slide in two directions [45]. The fluorescence emitted from the hybridized target molecule is gathered with an objective lens and converted to an electrical signal with a photo-multiplier tube (PMT) or an equivalent detector. The confocal scanning microscope is the most common one used to scan microarray slides. The main drawback of this type of technique is that this type of device may be very expensive since each excitation wavelength must have its own laser. In addition, the confocal scanning microscope is also very sensitive to any nonuniformity of the glass slide surface [27]. The CCD camera typically utilizes broadband xenon bulb technology and spectral filtration. The CCD system allows simultaneous acquisition of relatively large images of a slide (1 cm^2), thus, it does not require moving stages and optics. On the other hand, several images need to be captured from different areas and then combined to be representative of the complete information on the slide. Since most commonly used dyes have similar excitation and emission maxima, spectral filtration processes may have difficulty separating excitation and emission wavelengths, resulting in a possible source of error.

During the scanning process, the power of the excitation light is critical since the emitted fluorescence is generally correlated with the power of the excitation light. If the power of the excitation light is too low, the scanning sensitivity will be too long and many empty spots may be generated. However, if the power of the excitation light is too high, the incoming photons can damage the dyes and reduce the fluorescent signals during successive scans. More powerful light sources and/or longer laser exposure time can lead to significant photobleaching. Generally, photobleaching should be less than 1 percent per scan.

Since different dyes have different quantum yields and photostabilities, the PMT needs to be justified for each different channel prior to scanning. The order of channel scanning may be an additional variable to gain a better image. For example, Cy5 is more sensitive to photobleaching than Cy3. To minimize photobleaching, the Cy5 channel is always scanned first, followed by the Cy3 channel [27].

After the scanning process, we need to transform the image into quantitative signals. Many software packages, such as *Imagene*, *GPC VisualGrid*, *TIGR SpotFinder*, *GenePix*, have been developed to automatically quantify the images. Most of these software packages are effective. The common challenges for image quantification include (1) irregular or non-uniform spot geometries (e.g., not round, donut shape); (2) uneven hybridization (e.g., only a portion of the scanned image is quantifiable); (3) hybridization with high background; and/or (4) weak or saturated hybridization signals. Thus, for better quantification, one generally needs to use the following parameters: (1) signal/noise ratio should be more than $\sigma + 1.96\,\mu$; (2) background area selection should be local instead of global; and (3) bad spots should be removed.

3.3.5 Data Analysis

After obtaining the quantification hybridization signals from different biological replicates, we need to perform data normalization and statistical analysis.

3.3.5.1 Data Normalization The data normalization before statistical analysis is important to obtain reliable results. Data normalization can control many of the experimental sources of variability (systematic, not random or gene specific) and bring each image to the same average intensity. Data normalization is necessary to correct for the following variabilities: (1) the use of unequal quantities of starting RNA; (2) differences in dye incorporation; (3) differences in detection efficiencies of the fluorescent dyes; (4) variations in the image saturation extent for different channels; and (5) systematic biases in the measured expression levels.

Generally, there are several assumptions underlying data normalization. (1) The average mass of each molecule is approximately the same, thus the molecule number in each sample will be the same. (2) The arrayed elements represent a random sampling of the genes in the organism; and (3) the number of molecules from each sample available for hybridization is similar, thus, the total intensity for each sample will be the same.

Data normalization includes two steps: normalization within slides and normalization between slides. Normalization within slides is generally achieved by different options [46]. First, the signals can be scaled (scale normalization) by total intensity, mean, median, or the intensity of a group of genes. Second, normalization can be achieved by linear regression normalization. The most popular method for normalization with slides is the locally weighted linear regression (Lowess) normalization. Most normalization methods correct for differences in intensities between channels and do not take into account systematic bias that may appear within the data. For instance, the $\log_2$(ratio) values can have a systematic dependence on intensity. Lowess may remove the intensity-dependent effects in the

$\log_2$(ratio) values since Lowess normalizes the value point by point and generally requires a defined percent for the local area (e.g., 20 percent). Lowess normalization requires the ratio (two dyes). While normalization adjusts the mean of the $\log_2$(ratio) measurements, stochastic processes can cause the variance of the measured $\log_2$ (ratio) values to differ from one region of an array to another or between arrays. One approach to dealing with this problem is to adjust the $\log_2$(ratio) measures so that the variance is the same. This method is called variance regularization. Interested readers are encouraged to read an excellent review on data normalization by Quackenbush [46] for more details.

Since the hybridization may vary between slides (replicates) as well as channels, normalization is extremely important. Generally, normalization between slides uses scale normalization (e.g., medium).

3.3.5.2 Statistical Analysis After normalization, we will be able to perform statistical analysis to rank results by confidence with significance metrics (e.g., *p*-value). The statistical analysis will estimate the false positive (Type I errors) and false negatives (Type II errors), achieve the desired balance of sensitivity and specificity, and result in a certain amount of flexibility (and arbitrariness) for interpreting significance metrics generated by a test.

The methods for statistical analysis depend on the experimental design. For example, for two sample statistical tests, we can utilize parametric statistical methods (*t*-test for paired and unpaired *t*-test) or nonparametric methods (Mann–Whitney test for independent samples or Wilcoxon signed-rank test for paired data). We generally assume the variations between biological replicates and technical replicates are the same to apply the two-sample statistical test. Otherwise, we can use multivariate statistics, such as one-way versus two-way analysis of variance (ANOVA) or the Kruskal–Wallis method. For multiple comparison corrections, we can use Bonferroni Correction or False Discovery Rate [47]. More details about these methods can be obtained from the book *Statistical Analysis of Gene Expression Microarray Data* by Terry Speed [48].

Many software packages, such as *GeneSpring* (Silicon Genetics), *SAM* (Stanford), and *ArrayStat* (Imaging Research), have been developed for microarray data analysis. GeneSpring is one of the most widely used microarray data analysis software tools since it has an easy-to-use interface as well as powerful normalization and statistical analysis capabilities (*t*-test, two-way ANOVA tests, one-way posthoc tests for reliably identifying differentially expressed genes, and so on). Different computational analysis tools for clustering, visual filtering, and pathway viewing have also been included. Also, the user can incorporate their own scripts/programming into GeneSpring to complete their analysis.

3.3.6 Experimental Design and Data Interpretation

Correct experimental design is the key to generation of meaningful biological results. A good experimental design will be more economic since it may save resources as well as slides. However, a corresponding statistical analysis should be proposed as well to

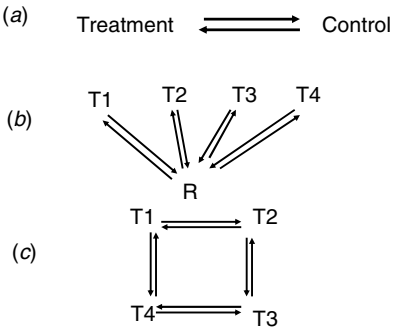

Figure 3-3 Experimental design for microarray. (a) Direct comparison. (b) Reference design. (c) Loop design.

analyze and interpret the data scientifically. Speed [48] provides a very good illustration for experimental design.

The simplest experimental design is pairwise direct comparison between treatment/experiment and control (Fig. 3-3a). For instance, we can compare the gene expression profiles for a wild type and a mutant under a certain condition to study the function of the mutated gene [49]; we can test the treatment effectiveness of a drug by comparing the gene expression profiles of the treatment group to the control group.

However, in most cases, we have to compare multiple experimental conditions. In this case, pairwise direct comparison will not meet the requirement. For example, we need to compare the gene expression profiles at different time points during bacterial growth [50]. In the drug experiments, we need to compare the effectiveness between different drugs. Obviously, it will not be wise to design all pairwise comparisons since it will be too expensive. For instance, to compare 10 conditions, one would have to design 45 pairwise experiments. In this case, we can apply common reference design (Fig. 3-3b).

More complicated designs include loop design (Fig. 3-3c) and pool design. The pool design should be very carefully used since it involves the mixture of all of the treatments and control samples as a reference sample to compare. The statistics with different experimental designs are described in the review by Yang and Terry [51].

After the statistical analysis, reconciliation between statistical results and biological functions is not a trivial matter since thousands of genes are involved in the data analysis. Generally, one overlays functional information and allows biological context to help decide what is of interest and what is not. We can use computational methods (classification, clustering, promoter prediction, and so on) to assist this analysis (Section 3.3.7). Microarray data are required to link to various public identifiers, such as Genbank, Swiss-Prot, and Gene Ontology (GO) database. GO is the most commonly used public domain sources of gene classification, and it provides controlled vocabulary hierarchies for molecular functions, biological processes, and cellular components. Other common databases include LocusLink, HomologGene, RefSeq, and UniGene.

3.3.7 Bioinformatics and Functional Genomics

The massive information that microarray profiling generates provides a great challenge for how to extract biologically meaningful information from the raw data. Thus, the discipline of bioinformatics plays an important role in microarray data analysis.

The most common approach is to deduce the coregulated genes (regulons) that have similar expression patterns. Further, the regulatory motifs are expected to be a predictor for each regulon. Many different algorithms have been used for the clustering process as well as for regulatory motif prediction. Within a single experimental condition, MotifRegressor [52] can be used to find a sequence motif. MotifRegressor first predicts all of the possible motifs and then performs regression analysis between microarray data and motif strength. MotifRegressor has an advantage in that it does not require the selection of a group of genes to predict the motif, which may generate a bias for motif prediction since some highly expressed genes are also indirectly regulated genes. For multiple experimental conditions, we can apply clustering methods, such as k-means, hierarchical clustering, self-organizing maps, and minimum spanning tree (EXCAVATOR) [53], to identify a group of potential genes with the same trend in expression pattern. EXCAVATOR [53] is based on a new framework for representing gene expression data, that is, the minimum spanning tree in graph theory. Through this data representation, an expression data-clustering problem is reduced to a tree-partitioning problem without losing information essential for the purpose of clustering. EXCAVATOR then applies an algorithm that mathematically guarantees to find globally optimal clustering efficiently, for a general objective function. After identifying the coexpressed genes in a cluster, we can apply motif prediction programs to predict the DNA binding motifs. The most commonly used *cis* regulatory motif and transcription factor DNA binding site prediction algorithms include such programs as Gibbs sampler [54], AlignACE [55,56], and BioProspector [57].

During the past several years, transcriptional regulatory networks have attracted substantial interest from both the computational and biological science communities. A number of statistical and computational methods have been applied in the modeling of gene regulation networks [58–64]. A few regulatory networks have been defined [60,65–68]. Despite this, regulatory network construction remains a great challenge due to the requirement of large experimental data sets.

The storage and management of microarray data is critical for efficient analysis. This, however, is a very challenging undertaking, since the many details of microarray analysis will affect the final results. The information about the samples hybridized, the hybridization images and their extracted data matrices, information about the physical array, and the features and reporter molecules all need to be included in the database. BioArray software environment (BASE) is a Web-based customizable bioinformatics solution for the management and analysis of all areas of microarray experimentation [69]. BASE manages biomaterial information, raw data and images, and provides integrated and "plug-in"-able normalization, data viewing and analysis tools. The organization and interface of BASE was designed to closely follow the natural workflow of the microarray biologist, and is compatible with most types of array platforms and data types (e.g., cDNA/oligos spotted on any substrate, Affymetrix, CGH on arrays, and so on).

3.3.8 Challenges of DNA Chips

Although DNA chips have advantages of high-throughput features, these technologies have several other disadvantages and challenges:

(1) Cost of diagnostic microarrays. Currently, the cheapest chips still cost the users at least $100 per experiment even for a noncustomed array. An Affymetrix array costs more than $400 per experiment. Currently, it is cost-prohibitive to apply microarrays as a routine diagnostic tool.

(2) The robustness of the microarray technologies must be improved. For SNP screening in particular, the sensitivity and specificity will need to be improved.

(3) The chip technologies need to be performed in a simplified and sturdy format without errors. A standard package includes the experimental protocol. These packages should tell the user how to justify the array quality in addition to giving the intensity of chip array data. A highly efficient quality control needs to be set up for microarray data analysis as well.

3.3.9 Development and Applications of DNA Chips

DNA chips have been widely used in many different fields. Most DNA chips focus on the protein-coding region to study the gene expression values. In addition, other types of arrays are designed to study the function of other elements in the genomes, such as small gene prediction, antisense gene study, gene alternative splicing, and so on.

The first significant application of DNA chips were serial analysis of gene expression (SAGE) for expression profiles [70]. SAGE was designed based on two principles: (1) a short nucleotide sequence tag can uniquely identify the transcript from an individual gene provided it is from a defined position within the transcript. For example, although the total number of human genes is expected to be of the order of 30,000, a sequence tag of only 9 nucleotides can, in principle, distinguish $4^9 = 262{,}144$ different transcripts. (2) Concatenation of short sequence tags allows the efficient analysis of transcripts in a serial manner. The tags from different transcripts can be covalently linked together within a single clone, and the clone can then be sequenced to identify the different tags in that clone. SAGE has been applied successfully in malarial parasite, yeast, plant, and animal systems [71].

ChIP-on-chip is a DNA array technique for isolation and identification of specific protein binding sites in genomic DNA [72]. ChIP-on-chip is useful for regulatory binding site identification, and thus, for regulatory network construction [73,74]. These regulatory binding sites can help identify the functions of the transcriptional regulatory protein during cell development and disease progression. The identified binding sites may also be used as a basis for annotating functional elements in genomes. The types of functional elements that one can identify using ChIP-on-chip include promoters, enhancers, repressor and silencing elements, insulators, boundary elements, and sequences that control DNA replication (http://www.chiponchip.org/).

Tiling array is a DNA array covering whole genome sequences using overlapped fragments, and it can be applied to examine not only upstream sequences of genes but

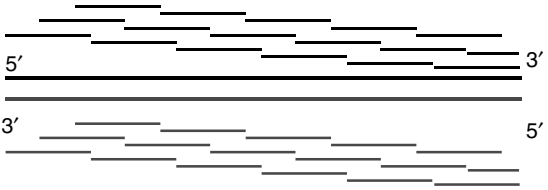

Figure 3-4 Probe design for tiling array. The oligonucleotide probes are tiled across the whole genomic sequence.

also intragenic and intergenic regions [75]. Tiling arrays use millions of DNA probes evenly spaced, or "tiled" across the genome, including coding and noncoding regions (Fig. 3-4). Tiling array has been a very useful tool for genome-wide analysis of many important biological functions, including transcription [76], antisense gene expression [77], protein binding sites [78], sites of chromatin modification [79], sites of DNA methylation [80,81], experimental genome annotation, and regulatory pathway discovery [82].

3.4 TRANSCRIPTOME PROFILING OF AN *ArcA* Mutant of *Shewanella oneidensis*

In this section, discussion will focus on cDNA microarray technology applied for the purpose of characterizing the ArcA regulon in the bacterium *S. oneidensis* MR-1. This section first introduces the background of this study (Section 3.4.1) and then describes the experimental design for this study (Section 3.4.2). The cDNA microarray and microarray hybridization procedure are followed next in Section 3.4.3. Section 3.4.4 describes the roles of bioinformatics in this study. Section 3.4.5 describes the transcriptome profiling of an *arcA* mutant. Finally, the conclusions are presented in Section 3.4.6.

3.4.1 Background

In *E. coli* and other bacteria, the Arc (anoxic redox control) two-component signal transduction system, which consists of the ArcB transmembrane sensor kinase and the cytosolic ArcA response regulator, modulates gene expression in response to changing redox conditions [83]. Under anaerobic or microaerobic respiratory conditions, ArcB autophosphorylates and then transphosphorylates the global transcriptional regulator ArcA, thereby enhancing the affinity of the latter protein for its target promoters [84–87]. ArcA is a transcriptional regulator that can act as an activator or repressor in regulating different genes in redox metabolism such as several dehydrogenases of the flavoprotein class, terminal oxidases, tricarboxylic acid cycle enzymes, enzymes of the glyoxylate shunt and enzymes in fatty acid degradation pathways [83,88]. Recently, ArcA was predicted to directly regulate 55 new genes involved in many different functional categories in *E. coli* [89].

S. hewanella oneidensis MR-1, a facultative gram-negative bacterium, is remarkable for its ability to utilize a diverse array of terminal electron acceptors during anaerobic respiration (e.g., fumarate, nitrate, nitrite, thiosulfate, elemental sulfur, trimethylamine *N*-oxide (TMAO), dimethyl sulfoxide (DMSO), Fe(III), Mn(III) and (IV), Cr(VI), and U(VI)). Because of this exceptional metabolic versatility and the potential use of this organism for bioremediation of metal/radionuclide contaminants in the environment, the approximately 5 Mb chromosome and the 0.16 Mb megaplasmid sequences comprising the *S. oneidensis* MR-1 genome were deciphered by TIGR [60]. Sequence annotation of the MR-1 genome revealed the presence of an *arcA* homologue (SO3988) but not an *arcB* homologue. In this study, whole-genome DNA microarrays for *S. oneidensis* MR-1 were used to define the *arc*A regulon under both aerobic and anaerobic batch growth conditions. Transcriptome analysis of an *arcA* null mutant and the occurrence of a predicted sequence motif for promoter recognition by ArcA suggested that ArcA functions as a global regulator in *S. oneidensis*.

3.4.2 Microarray Construction and Hybridization

3.4.2.1 Microarray Construction The *S. oneidensis* microarray contained a total of 4,761 distinct elements, representing about 99 percent of the total protein-coding capacity of the MR-1 genome [49,90] (Fig. 3-5). Of the array elements that were spotted, 4,310 constituted PCR-amplified DNA fragments corresponding to unique segments of individual MR-1 ORFs, whereas gene-specific oligonucleotide probes (50-mers) were designed and synthesized for 451 predicted genes (9 percent of the total DNA probes arrayed) that did not yield either single products or any products in PCR amplifications. PCR primers and oligonucleotide probes were designed using the program PRIMEGENS [30]. PCR products and oligonucleotides were printed in duplicate onto SuperAmine glass slides (TeleChem International, Inc.). The microarray also consisted of 32 elements corresponding to *S. oneidensis* genomic DNA (positive controls) and 42 spots representing nine genes (amplicons) from *Arabidopsis thaliana* (negative controls).

3.4.2.2 RNA Isolation, cDNA Labeling, Microarray Hybridization, and Scanning Cultures of *S. oneidensis* wild-type and *arcA* mutant strains were harvested at the mid-exponential point under both aerobic and anaerobic conditions, and total cellular RNA was isolated using the TRIzol reagent (Invitrogen, Carlsbad, CA) according to the manufacturer's instructions. RNA samples were treated with RNase-free DNase I (Ambion, Inc., Austin, TX) to digest residual chromosomal DNA and then purified with the QIAGEN RNeasy Mini kit prior to spectrophotometric quantitation at 260 and 280 nm.

Fluorescein-labeled cDNA copies of total cellular RNA extracted from wild-type and mutant cells were prepared, with the exception that Cy3/Cy5-dUTP (Perkin–Elmer/NEN Life Science Products, Boston, MA) was used in the first-strand reverse transcription (RT) reaction. Two sets of duplicate reactions were carried out in

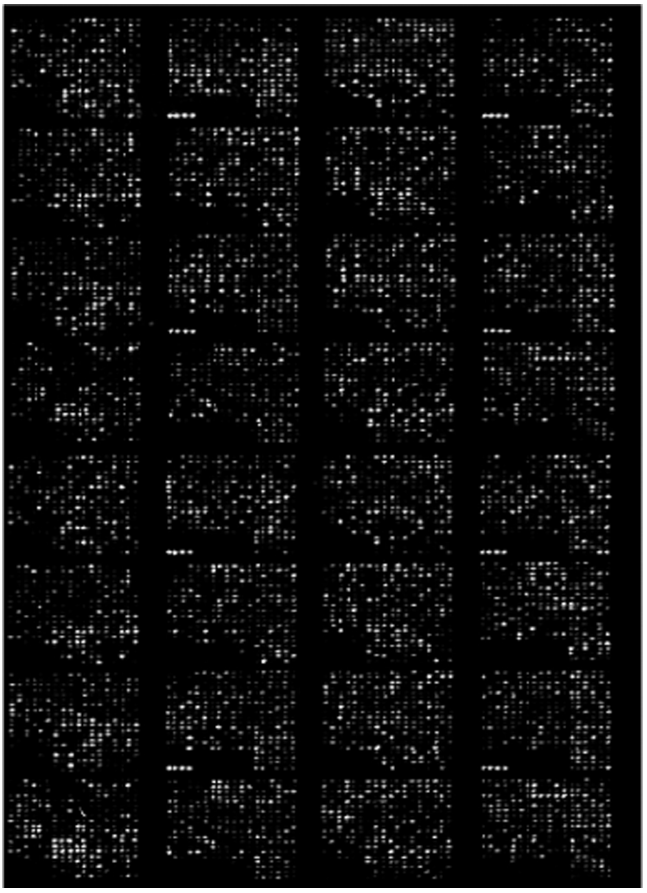

Figure 3-5 Whole genome cDNA microarray for *S. oneidensis* MR-1.

which the fluorescent dyes were reversed during cDNA synthesis to minimize gene-specific dye effects. The labeled cDNA probe was purified and concentrated by following the manufacturer's protocols.

The two labeled cDNA pools (wild type and mutant) to be compared were mixed and hybridized simultaneously to the array in a solution containing 3 × SSC (1 × SSC is 0.15 M NaCl plus 0.015 M sodium citrate), 0.3 percent sodium dodecyl sulfate, 1 μM dithiothreitol (DTT), 40 percent (v/v) formamide, 0.8 μg of unlabeled herring sperm DNA (Gibco BRL)/μL, and 8.6 percent distilled H_2O. Hybridization was carried out in a 50°C water bath for 12–15 h.

To determine the fluorescence intensity (pixel density) and background intensity, 16-bit TIFF scanned images were analyzed using the software ImaGene version 5.5 (Biodiscovery, Inc., Los Angeles, CA). Microarray outputs were first filtered to remove spots with poor signal quality by excluding those data points with a mean intensity of <2 standard deviations above the overall background for both channels.

Empty spots and spots flagged as poor were removed from subsequent analyses by using ImaGene. Data transformation and normalization were carried out using GeneSite Light (Biodiscovery, Inc.). Normalized expression ratios were imported into ArrayStat (Imaging Research, Inc., Ontario, Canada) to determine the common error and to remove outliers. Only those genes with an expression ratio of ≥ 2 were included in further analyses.

3.4.3 Experimental Design and Data Analysis

3.4.3.1 Experimental Design Figure 3-6 illustrates the experimental design for this study. To study the *arcA* gene in *S. oneidensis*, we first constructed an in-frame deletion *arcA* mutant (designated ARCA) based on the method described earlier [49] using the primers 3988-5I (5′-TGTTTAAACTTAGTGGATGGGCCTCAGTTACCACATACCC-3′), 3988-3I (5′-CCCATCCACTAAGTTTAAACACCAGATACGCCAGAAATCATCG-3′), 3988-5O (5′-GCTTCTGTCGATAAACACGGC-3′), and 3988-3O (5′-TTACCCAATACTTAGTTCAGCAAGG-3′). To monitor global changes in gene expression in response to the *arcA* deletion, we compared the ARCA strain with the DSP10 parental strain grown under aerobic and anaerobic conditions using batch cultures. *S. oneidensis* parental and mutant strains were grown in Luria-Bertani (LB) medium at 30°C under aerobic or anaerobic (with 20 mM fumarate as the electron acceptor) respiratory conditions. For the aerobic condition, cells were grown (60 mL in 250-mL flasks) with agitation (200 rpm). For the anaerobic condition, the media (80 mL in 100-mL bottle) was purged with nitrogen gas while boiling for at least 30 min prior to inoculation. To minimize differences in gene expression caused by

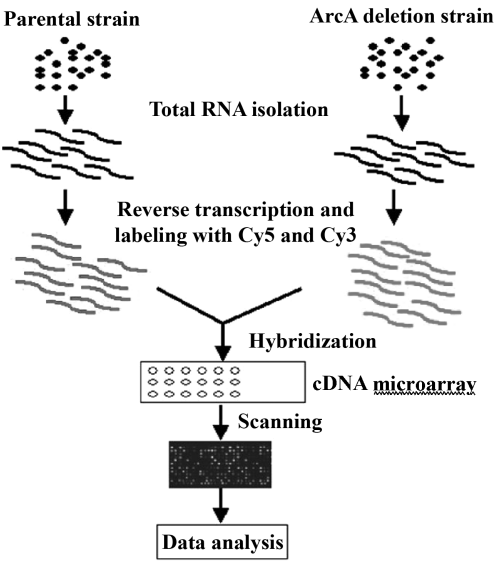

Figure 3-6 Experimental design for arcA regulon characterization.

growth-related effects, samples for transcriptome measurements were taken from exponentially growing cultures at mid-log phase.

For each growth condition tested, gene expression analysis was performed using six independent microarray experiments, including dye swapping, which yielded a total of 12 expression measurements per gene (three biological replicates, with each different mRNA preparation having four technical replicates).

3.4.3.2 Phenotype Characterization of the ARCA Mutant Strain To determine whether inactivation of the *S. oneidensis arcA* affects anaerobic metabolism, the ability of the ARCA mutant strain to grow on and/or reduce a variety of electron acceptors under anaerobic respiratory conditions was compared to that of the parental DSP10 strain [49]. The ARCA and DSP10 strains were cultured anaerobically in Luria-Bertani media with various electron acceptors, including fumarate (20 mM), colloidal Mn (5 mM), MnO_2 (2 mM), nitrite (20 mM), $MgCl_2$ (10 mM), CrO_4 (150 μM), cobalt (50 μM), FeO_2 (5 mM), ferric citrate (5 mM), $FeCl_3$ (5 mM), or Fe-NTA (10 mM). The culture turbidity was monitored spectrophotometrically at 600 nm. A growth curve was measured for the culture containing fumarate. For other electron acceptors, the growth of the culture was evaluated using end-point culture turbidity measurements.

The results indicated that the growth of ARCA in LB is slightly slower than the parent DSP10 strain (Fig. 3-7) under anaerobic conditions with fumarate (20 mM) as the electron acceptor. Based on the end-point culture turbidity, the *arcA* deletion mutant exhibits slower growth than the DSP10 parental strain under anaerobic respiratory conditions with the following electron acceptors: colloidal Mn (5 mM), MnO_2 (2 mM), nitrite (20 mM), $MgCl_2$ (10 mM), CrO_4 (150 μM), cobalt (50 μM),

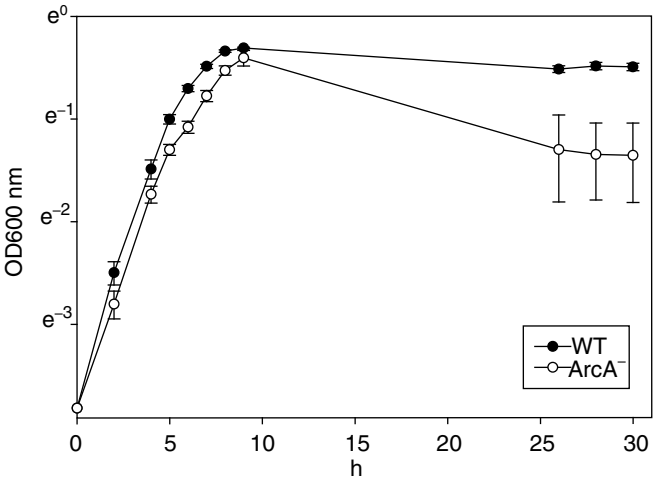

Figure 3-7 Comparison between the growth curves of ARCA (*arcA* null mutant) and the wild-type *S. oneidensis* DSP10 strain grown in Luria-Bertani medium at 30°C under anaerobic (with 20 mM fumarate as the electron acceptor) respiratory conditions.

FeO_2 (5 mM), ferric citrate (5 mM), $FeCl_3$ (5 mM). The growth of the ARCA and DSP10 strains was also evaluated in M4 minimum medium with ferric citrate (10 mM, 20 mM, and 50 mM, respectively) based on the culture turbidity at 24, 48, and 72 h, and the results demonstrated that ARCA grew slower than the parent DSP10 strain (data not shown).

The effect of hydrogen peroxide (H_2O_2) treatment at concentrations of 500 and 2500 μM on mid-exponential growth of the parental and mutant strains under aerobic conditions was also assessed. As shown in Figure 3-8, ARCA was shown to be more sensitive to H_2O_2-induced oxidative stress at different concentrations compared to the parental DSP10 strain, suggesting that ArcA might play a regulatory role in oxidative stress resistance in *S. oneidensis*. This observation agrees with the finding that *arc*A increases resistance of *Salmonella enterica* serovar Enteritidis to H_2O_2 [91].

3.4.4 Data Interpretation

3.4.4.1 Overview of Transcriptome Profiling of the ARCA Mutant under Different Respiratory Conditions A total of 654 (294 downregulated; 360 upregulated) and 504 (135; 369) genes were identified as being differentially expressed in response to the *arcA* deletion mutation under aerobic and anaerobic respiratory conditions, respectively. Comparison of the two microarray data sets indicated that the expression levels for 248 of these genes were affected under both aerobic and anaerobic growth conditions. The differentially expressed genes encode a broad variety of functions, with the majority (44–52 percent) encoding hypothetical or conserved hypothetical proteins (Fig. 3-9a and b). Genes showing changes in

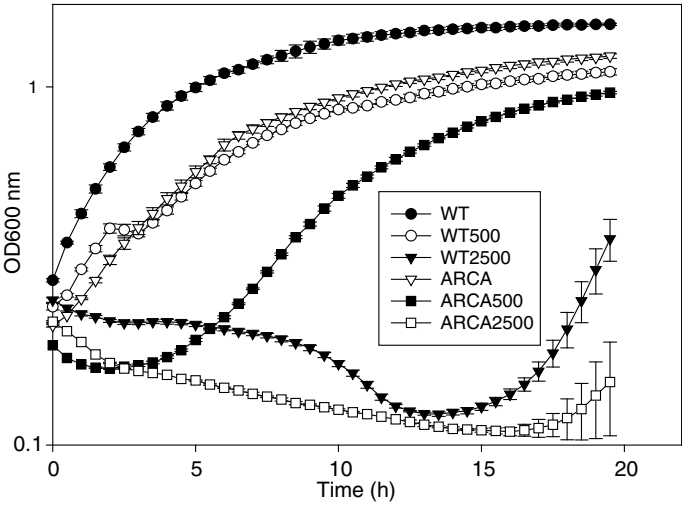

Figure 3-8 ARCA is more sensitive to oxidative stress than the wild-type *S. oneidensis* DSP10 strain. Growth was measured kinetically with a Microbiology Reader Bioscreen C (Growth Curves USA, Piscataway, NJ) [34]. WT and ArcA mutant were grown aerobically up to the mid-log phase and then treated immediately with 500 and 2500 μM H_2O_2, respectively. The cells were grown at 30°C with continuously extensive shaking. The OD_{600} nm units were read with an interval of 30 min.

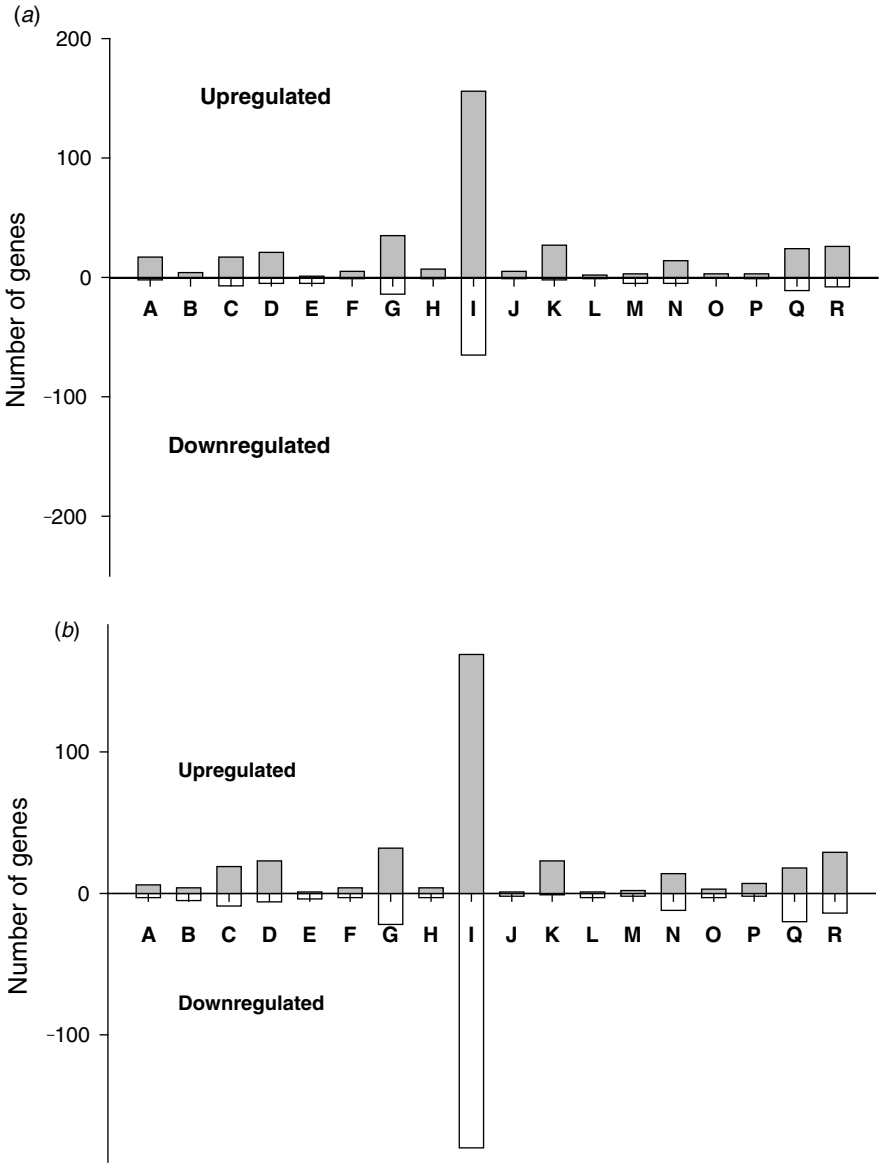

Figure 3-9 Functional distribution of differentially expressed genes in the ARCA strain under anaerobic (a) and aerobic conditions (b). Genes are grouped in their corresponding functional/homology classes (number of genes/percentage of genes) according to TIGR's annotation (http://www.tigr.org/tigr-scripts/CMR2/gene_attribute_results_org_or_role.dbi): (A) amino acid biosynthesis; (B) biosynthesis of cofactors, prosthetic groups, and carriers; (C) cell envelope; (D) cellular processes; (E) central intermediary metabolism; (F) DNA metabolism; (G) energy metabolism; (H) fatty acid and phospholipid metabolism; (I) hypothetical proteins; (J) other categories; (K) protein fate; (L) protein synthesis; (M) purines, pyrimidines, nucleosides, and nucleotides; (N) regulatory functions; (O) signal transduction; (P) transcription; (Q) transport and binding proteins; and (R) unknown function.

transcript abundance in the *arcA* deletion mutant under both growth conditions that have annotated functions are involved in a number of cellular processes including cell envelope, energy metabolism, protein fate, regulatory functions, and transport/binding proteins. Under anaerobic growth conditions, a number of genes belonging to the functional categories of protein synthesis and purines/pyrimidines/nucleosides/nucleotides were also upregulated in the *arcA* mutant (Fig. 3-9a). There were 27 and 19 predicted regulatory genes that showed significant differences in expression in a Δ*arcA* genetic background under aerobic and anaerobic conditions, respectively, and 13 genes with annotated functions in regulation were differentially expressed under both respiratory conditions. These results suggest that ArcA functions as a global regulator in *S. oneidensis*, exerting a pleiotropic effect on a number of cellular functions, and that the transcriptional effect of an *arcA* deletion was most profound under anaerobic growth conditions.

3.4.4.2 Genes with Functions in Energy Metabolism A total of 66 of 87 (∼76 percent) genes with annotated functions in energy metabolism showed altered expression profiles in the *arcA* deletion mutant (49 under anaerobic conditions and 54 under aerobic conditions). More than half of these genes (34 genes) are involved in electron transport function. Except for the *napAGHB* operon, which was upregulated under anaerobic conditions but downregulated under aerobic conditions, all other genes showed similar expression trends under anaerobic and aerobic conditions. For 19 cytochrome *b* or *c* genes, 13 were upregulated under either or both anaerobic and aerobic conditions including SO4483 (cytochrome *b*, putative), cytochrome *c* family proteins (SO1782, SO1659, SO4079–SO4078 operon, SO4142, SO4144, SO4484), cytochrome *c* oxidase *ccoPONQ*, and diheme cytochrome *c* (SO4485). However, the other six genes, cytochrome *c* (*scyA*, SO3300, SO4572, SO2727, SO0845) and decaheme cytochrome *c* (SO1427) were downregulated at either or both of these two experimental conditions. Among iron–sulfur clustering binding proteins, SO1364, *napG*, and *napH* were downregulated 2.98-, 4.44-, and 7.50-fold under aerobic conditions, respectively. The other three iron–sulfur cluster proteins (SO1519, SO1521, SO4404) were instead upregulated under aerobic conditions. Genes involved in anaerobic metabolism such as *torC* (tetraheme cytochrome *c*), c*at2* (4-hydroxybutyrate coenzyme A transferase), and *fdhB* (formate dehydrogenase) were upregulated under anaerobic conditions, but *dmaAB* (anaerobic dimethyl sulfoxide reductase), *ifcA* (fumarate reductase flavoprotein), and SO4513 (formarate dehydrogenase) were downregulated under anaerobic conditions.

Among operons/genes-encoding enzymes involved in the TCA cycle, malate synthase (*aceBA*), and aconitate hydratase 1 (*acnA*) were upregulated in ARCA under anaerobic fumarate-reducing conditions, which is similar to the situation in *E. coli* [89,92]. Up- or downregulation of the genes associated with the TCA cycle will affect redox generation, which reflects the role of ArcA in redox metabolism. However, seven other operons, including citrate synthase (*gltA*) [93], succinate dehydrogenase operon (*sdhCAB*) [94], 2-oxoglutarate dehydrogenase–succinyl–CoA synthase operon (*sucABDC*) [95,96], malate dehydrogenase (*mdh*) [94], aconitate hydratase 2 (*acnB*) [97], isocitrate dehydrogenase (*icd*) [98], and SO2222

(fumarate hydratase) were not affected in the *S. oneidensis* ARCA mutant under anaerobic fumarate-reducing conditions. Another fermentation gene, D-lactate dehydrogenase (*ldhA*), was not affected significantly under anaerobic conditions in *S. oneidensis*, but was upregulated 1.85-fold in an *E. coli arcA* mutant in MOPS-buffered LB with 20 mM D-xylose [89,99].

3.4.4.3 Expression of Genes from Other Functional Categories Here we discuss genes known to be regulated directly or indirectly by *arcA* in other microbes. About 30 operons (including the reported gene/operons discussed above), most of which are involved in respiratory metabolism, are presently known to be regulated by phosphorylated *arc*A in other organisms [99].

The glutamate synthase operon (*gltDB*) was upregulated under anaerobic conditions, which is similar to their up-regulation in *E. coli arcA* mutants under anaerobic conditions [83,89,99]. In contrast, nine other operons related to redox metabolism, including formate acetyltransferase (*pflB*), cytochrome *d* ubiquinol oxidase operon (*cydAB*) [100–102], aldehyde dehydrogenase (*aldA*), fatty acid oxidation complex (*fadBA*), NADH dehydrogenase (*nuo*) operon, the ATP binding protein operon (*cydDC*) [100,101], glycerol kinase (*glpK*), anaerobic C4-dicarboxylate membrane transporter (*dcuB*), and lipoamide dehydrogenase (*lpdA*) [83,92,99,103,104], were not affected under the growth conditions tested in this study. These nine genes were shown to be regulated by ArcA in other organisms in previous studies [83,92,99–102,104]. The transport and binding protein, C4-dicarboxylate binding periplasmic protein (*dctP*) [92], was upregulated about 2.3-fold under aerobic conditions, which is similar to the reported trend (*dctA*, up-regulated 1.58) from *E. coli arcA* mutant microarray data [89].

3.4.4.4 Resistance of S. oneidensis arcA Null Mutant to H_2O_2 Oxidative Stress As described earlier, the ARCA mutant strain is hypersensitive to H_2O_2 relative to the DSP10 parental strain under aerobic conditions. The *oxyR* gene encodes a transcriptional binding protein that regulates oxidative stress resistance in *S. enterica* serovar Typhimurium and *E. coli* [105]. In this study, the expression of the *S. oneidensis oxyR* homologue, gene SO1328, was not affected by the *arcA* deletion. Also of interest was the observation that the MR-1 counterparts for such known OxyR-controlled genes as *katG* (hydroperoxidase I), *ahpF* (alkyl hydroperoxide reductase), *gor* (glutathione reductase), *grx* (glutaredoxin, SO2745), *fur* (Fur repressor of ferric ion uptake), *dps* family protein (SO1158), and *hemH* (SO2018 and SO3348) [105,106], were not affected in the ARCA strain, even though the deletion mutant exhibited H_2O_2 hypersensitivity.

Nystrom et al. [107] also showed that an *E. coli arcA* deletion mutant was not able to decrease the synthesis of the TCA enzymes malate dehydrogenase (*mdh*), isocitrate dehydrogenase (*aceB*), lipoamide dehydrogenase E3 (*lpdA*), and succinate dehydrogenase (*sdh*). Similarly, our transcriptome profiling of ARCA under aerobic conditions demonstrated that the transcription of enzymes within the TCA cycle was not affected significantly (Table 3-2). These enzymes are encoded by *gltA*, *lpdA, sdhCAB*, *sucABCD*, *mdh*, *acnB*, *aceB*, *icd*, *acnA*, and *acnB*. The microarray data for the *arcA*

Table 3-2 Microarray expression ratio and DNA motif prediction for the genes verified to be regulated directly by ArcA in *E. coli*

Functional Category	ORF	Gene Product	Anaerobic[a] (*arcA*/WT)	Aerobic (*arcA*/WT)	Start[b]	Strand	*Z*[c]	Predicted Motif
Amino acid biosynthesis	SO1325	Glutamate synthase, large subunit (gltB)	8.31	3.21	−227	+	2.06	GTTCTTTATTTTTTA
Energy metabolism	SO0432	Aconitate hydratase 2 (acnB)	1	1.36	−266	−	2.66	GTTCATCAATTTAAA
	SO1021	NADH dehydrogenase I, A subunit (nuoA)	1.18	0.63	−1452	−	2.02	TTTAATTGAAATTTA
					−616	+	2.35	GTTCAGGAAGTGTTT
					−140	+	2.14	GTTAATCCACTTAGT
					−240	−	2.3	GTTAAGCAATACATT
	SO1483	Malate synthase A (aceB)	3.72	1.57	−712	−	2.37	GTTAACCGTTTTCCA
	SO2629	Isocitrate dehydrogenase, NADP dependent (icd)	0.79	0.75	−493	−	2.69	GTTAATTCTAATAGA
					−331	+	2.16	GTTCATGCTTTGGCT
	SO3286	Cytochrome d ubiquinol oxidase, subunit I (cydA)	1.5	1.82	−1035	+	2.67	GTTCACACAATTGAA
					−590	−	2.9	GTTAATCAAAATAAA
	SO4230	Glycerol kinase (glpK)	1.34	1.03	−255	−	2.66	GTTAACAATATCCAT
					−203	+	2.22	GTTAATTAGATTTTT
	SO0426	Pyruvate dehydrogenase complex, E3 component, lipoamide dehydrogenase (lpdA)	1.08	1.04	NF	NF	NF	NF
	SO1926	Citrate synthase (gltA)	0.86	1.33	NF	NF	NF	NF
	SO1927[e]	Operon *sdhCAB*	0.68	0.55	NF	NF	NF	NF

	SO1930[e]	Operon *sucABCD*	1.03	1.5	NF	NF	NF	NF
	SO2912[e]	Operon *pflBA*	1.18	1.05	NF	NF	NF	NF
	SO4480	Aldehyde dehydrogenase	1.34	0.95	NF	NF	NF	NF
Fatty acid and phospholipid metabolism	SO0021	Fatty oxidation complex, alpha subunit (fadB)	1.12	0.78	−384	+	2.05	GTTAATAATAAATAT
Protein fate	SO3637	Survival protein surA (surA)	1.17	1.8	−151	+	2.19	GTTAATGAAAGCCGT
Regulatory functions	SO3988	Aerobic respiration control protein ArcA (arcA)	0.11	0.07	−392	+	2.57	GTTAACAAATGCCTA
Transport and binding proteins	SO0827	L-lactate permease (lldP)	0.99	1.06	−246	−	2.57	GTTAATCAAGGTATA
	SO3134	C4-dicarboxylate binding periplasmic protein (dctP)	1.62	2.31	−151	−	2.03	GTTAATAAAGTGTAG
	SO3780	ABC transporter, ATP binding protein CydD (cydD)	1.1	1.17	−35	+	2.32	GTTAAGCCTATTTCT

[a]Relative gene expression is presented as the mean ratio of the fluorescence intensity of the *arcA* deletion mutant (ARCA) to that of the parental strain (WT), and NA means that there is not any expression value available.
[b]The start site of the predicted ArcA motif away from the gene translation start site.
[c]The statistical *Z*-score of the predicted motif from the genomic-wide scan.
[d]No ArcA-P binding motif found in *S. oneidensis* MR-1.
[e]Only the gene expression value for the first gene in the operon was shown.

deletion mutant under H_2O_2 stress also indicated that the expression of these genes was not changed significantly (T. Li and J. Zhou, unpublished data, personal communication). Therefore, these TCA cycle enzymes may still produce the reactive oxygen species (ROS) after exposure to H_2O_2. Nystrom et al. [107] demonstrated that *E. coli* was able to overproduce superoxide dismutase to scavenge superoxide radicals generated from aerobic respiration to defend against oxidative stress. The deletion of *sodB* in *Helicobacter pylori* results in hypersensitivity of the mutant to oxidative stress and a defect in host colonization [108]. Here we found that superoxide dismutase (*sodB*) was downregulated about 1.7-fold under aerobic respiratory conditions. Our experiments also demonstrated that the gene encoding periplasmic nitrate reductase (*nap*A), which has been reported to be associated with oxidative stress resistance in *H. pylori* [108,109], was downregulated 11.3-fold under aerobic respiratory conditions. This might explain the hypersensitivity of the ARCA mutant to H_2O_2 oxidative stress. In addition, we found that two heavy metal efflux pump operons (SO4597–SO4598 and SOA0154–SOA0153) were downregulated 6.7- and 25-fold, respectively. It is unknown whether the low levels of expression of these two operons will affect the ARCA strain's H_2O_2 stress resistance capability.

3.4.5 Bioinformatics Analysis

3.4.5.1 Sequence Analysis and Structural Modeling of S. oneidensis arcA The putative *arcA* gene of *S. oneidensis* MR-1 encodes a 238-amino acid protein with a predicted molecular mass of 27,220 Da and a pI of 5.43. Comparison of the deduced amino acid sequence showed that *S. oneidensis* MR-1 ArcA shares a high degree of identity to its homologues in *E. coli* (81 percent), *S. enterica* (81 percent), *Yersinia pestis* (81 percent), *V. cholerae* (81 percent), and a lower degree of sequence identity to ArcA in *Pasteurella. multocida* (75 percent) and *H. influenzae* (72 percent) (Fig. 3-10). This high level of homology at the primary sequence level strongly suggests that these proteins share similar biological functions. Moreover, analysis of the deduced amino acid sequence of *S. oneidensis* MR-1 ArcA also revealed the conservation of the Asp^{54} residue in the N-terminal receiver domain and the helix–turn–helix (HTH) DNA binding motif in the carboxy-terminal effector domain (Fig. 3-10) [85,88]. Based on structure predictions using PROSPECT-PSPP [110], ArcA shows high homology to the response regulator Drrb present in *Thermotoga maritima* (PDB id 1p2f, 30). Drrb is a multidomain response regulator of the OmpR/PhoB subfamily that may regulate gene transcription by binding as a dimer to σ^{70} promoter elements [111]. In contrast to *E. coli*, *S. oneidensis arcA* is predicted to be monocistronic, and there is no obvious cognate *arc*B encoded in the MR-1 genome based on the sequence annotation [112]. This suggests that a less conserved sensor histidine kinase might be employed by the Arc two-component signal transduction system.

3.4.5.2 Scanning the S. oneidensis MR-1 Genome with the ArcA-P Positional Weight Matrix Structure modeling of the deduced protein encoded by the MR-1 *arcA* gene indicated a strong degree of conservation between the DNA

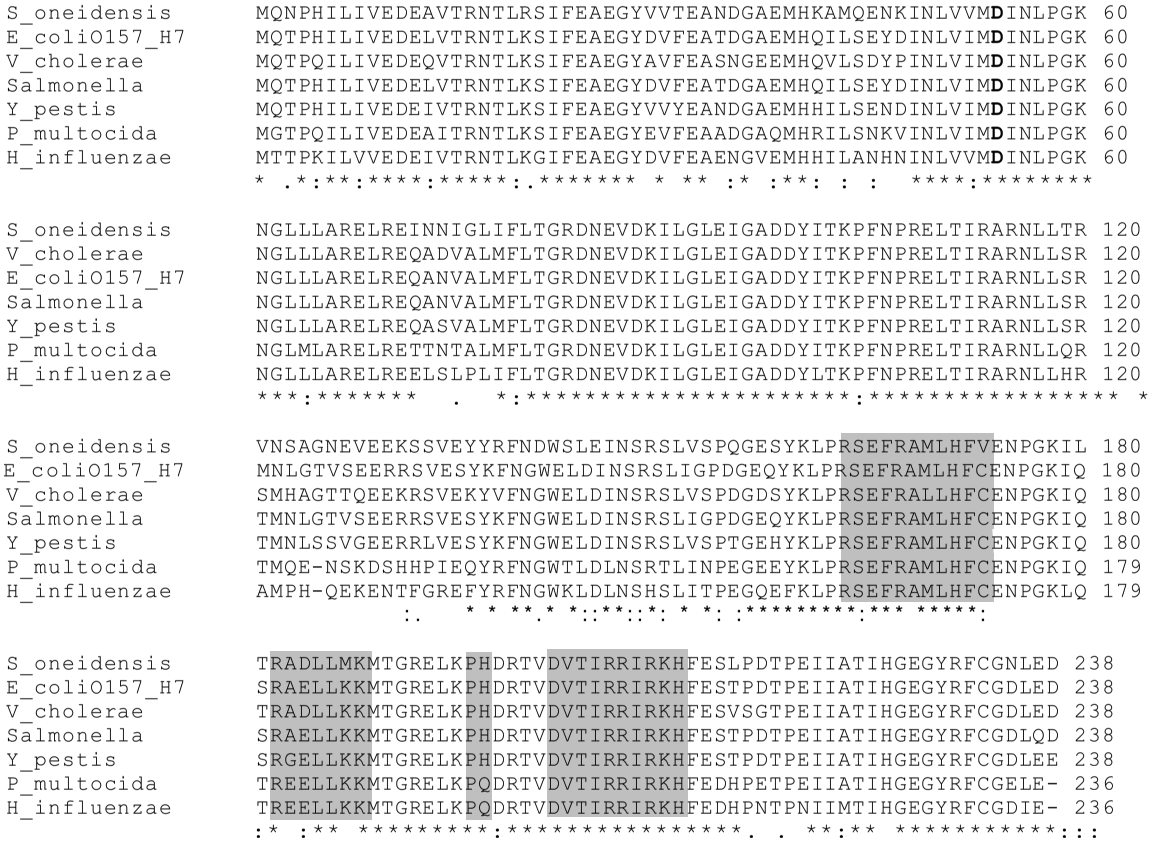

Figure 3-10 Multiple sequence alignments for ArcA from *E. coli* (GeneBank ID: 15804972), *S. enterica* (GeneBank ID: 16763386), *Y. pestis* (GeneBank ID: 16120787), *V. cholerae* (GeneBank ID: 15642365), *P. multocida* (GeneBank ID: 15602084), and *H. influenzae* (GeneBank ID: 16272824). The multiple alignments were performed using ClustalW [119]. The shading areas are the predicted helix by PROSPECT-PSPP [110] indicating the DNA binding domain, and the bold letters are the phosphorate sites.

binding domains of the *E. coli* and *S. oneidensis* ArcA proteins (Fig. 3-10). Thus, we utilized the experimentally verified ArcA-P binding sites from 10 *arc*A-regulated proteins to construct the ArcA-P positional weight matrix [89]. The score function of positional weight matrices was adapted from the log transformation method described by Berg and Von Hippel [113]. Both strands of the *S. oneidensis* MR-1 genome sequence were scanned using a sliding window size of 15 nucleotides. The motif with the highest matrix score was selected among all of the overlapping motifs from both plus and minus strands. Scores of all potential ArcA-P recognition sites were statistically analyzed using the *Z* test, and only those sites with 95 percent or greater significance are presented as potential ArcA-P binding sites in *S. oneidensis*. For each gene, only the promoters located within the upstream sequence from the gene start codons are counted in this paper.

By scanning the *S. oneidensis* genome with the ArcA-P recognition weight matrix, 13 tRNA and 668 protein-encoding genes were predicted to contain potential ArcA binding sites in their upstream regions. The predicted ArcA regulon in *S. oneidensis* includes 12 ORFs shown to be controlled by ArcA in *E. coli*: *fadB*, *acnB*, *nuoA*, *gltB*, *aceB*, *icd*, *dctP*, *cydA*, *cydD*, *arcA*, *glpK*, and *lldP* [83,89,92] (Table 3-2). The *sur*A gene, which is predicted to contain an ArcA-P binding site in *E. coli* [89], also has a strong *arc*A motif in *S. oneidensis*. However, the other six operons, *lpdA*, *gltA-sdhCAB*, *sucABDC*, *pflBA*, and *aldA*, which are regulated by *arc*A in *E. coli*, do not possess strong ArcA binding motifs based on this search. Among the 668 protein-coding genes in the *S. oneidensis* ArcA regulon, 148 genes (about 3 percent of all the predicted genes in *S. oneidensis*) exhibited significant differences in transcript levels in ARCA relative to DSP10 under aerobic and/or anaerobic conditions. Table 3-2 shows a subset of genes in *S. oneidensis* with ArcA-P binding sites, which are similar to ArcA-P binding sites in *E. coli* [83,89,92]. A sequence logo representation of the predicted conserved ArcA-P binding motif for these 148 genes is shown in Figure 3-11a. Compared with the ArcA motif in *E. coli*, the predicted motif has a weaker consensus. For example, the first, third, and fifth positions in the motif have smaller bit scores, which reflect the conservation status for each consensus position. Another genomic scanning in *S. oneidensis* using the positional weight matrix constructed from the predicted 190 binding sites for 148 genes resulted in a similar consensus (data not shown). Most (81 percent) of the motifs are located within 300 nucleotides upstream of the translation start codon (Fig. 3-11b).

Similar to the ArcA regulon in *E. coli* (20), the ArcA regulon in *S. oneidensis* is associated with 17 functional categories. Among the 148 genes with differences in expression in the *arcA* deletion mutant (Table 3-3), 46 genes were predicted to be positively regulated and 102 negatively regulated. Our results also showed that the genes controlled by ArcA are involved in functions beyond redox metabolism. These genes belong to broad functional categories and most of these genes have not been reported previously to be members of ArcA regulons from other bacterial species, such as *E. coli*. Eight of these genes with expression changes with more than 3-fold (up or down) have strong predicted ArcA binding motifs ($Z > 3.0$) and encode HoxK (SO2099), Pal/histidase family protein (SO3299), decaheme cytochrome *c* (SO1427), putative long-chain fatty acid transport protein (SO3099), TonB-dependent receptor domain

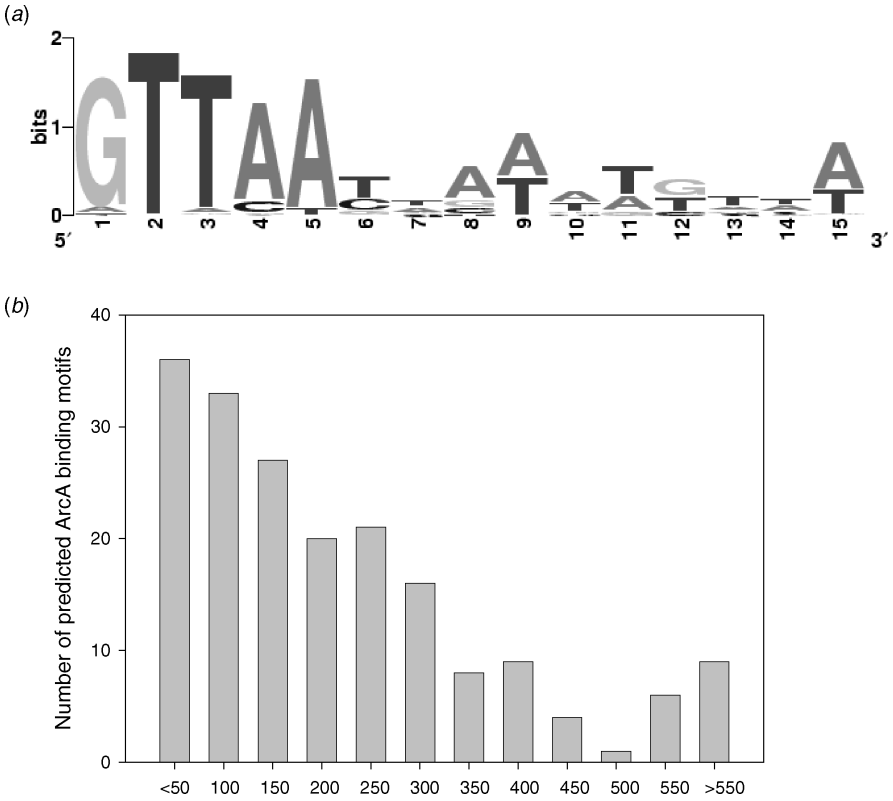

Figure 3-11 Identification of a predicted consensus ArcA binding motif in *S. oneidensis* MR-1 using computational methods. (a) Sequence logo representation of the predicted ArcA binding motif in *S. oneidensis* MR-1. (b). Position distribution of the predicted ArcA motifs.

protein (SO2907), MaoC domain protein (SO0599), PspF (SO1806), and a hypothetical protein (SO2930). Among these eight genes, *hoxK* is the first gene in the quinone-reactive Ni/Fe hydrogenase operon (*hoxK–hydB–hydC*), which catalyzes the reversible oxidation of H_2 [114]. Deletion of *hoxK* was shown to inactivate the membrane-bound hydrogenase in *Alcaligenes eutrophus* [115]. These three genes (*hoxK–hydB–hydC*) are upregulated more than 3.9-fold under aerobic conditions. Under anaerobic conditions, *hoxK* is also upregulated more than 3.5-fold. In addition, a putative undecaprenol kinase (SO4274) was also predicted to have a strong ArcA-P binding site with a *Z*-score larger than 3. The undecaprenol kinase (*so*4274) is a cell wall synthesis gene and has been associated with biofilm formation in *Mycobacterium smegmatis* [116]. Recently, *arc*A was found to be related to biofilm formation in *S. oneidensis* MR-1 [112] in which the undecaprenol kinase might be the functional gene target. These wide-ranging functions of ArcA are also supported by its requirement for virulence in *Haemophilus influenzae* [117] and *Vibrio cholerae* [118] as well as a recent genome-wide study of the ArcA regulon in *E. coli* [89].

Table 3-3 The genes in *S. oneidensis* MR-1 with expression differences that also have predicted binding sites

ORF	Gene Product	Ratio (ARCA/WT)		Start[b]	Strand	Z^c	Motif
		Anaerobic[a]	Aerobic				
Amino acid biosynthesis							
SO2483	Aspartate aminotransferase, putative	0.18	NA	−127	−	2.15	TTTAACTAAGTGTTA
SO1325	Glutamate synthase, large subunit (gltB)	8.31	3.21	−227	+	2.06	GTTCTTTATTTTTTA
SO4245	Amino acid acetyltransferase (argA)	2.01	1.16	−162	+	2.2	GTTAAAAAAATGTGA
				−62	+	2.07	GTGAATTAAAATGCA
SO2071	Imidazoleglycerol phosphate dehydratase/histidinol-phosphatase (hisB)	2.63	1.2	−60	−	2.08	GTTAATACTTGCACT
SO4349	Ketol acid reductoisomerase (ilvC)	1.21	2.41	−44	+	2.21	GTTAACAATAAGTTG
Biosynthesis of cofactors, prosthetic groups, and carriers							
SO1198	Dihydropteroate synthase (folP)	1.17	2.25	−138	+	2.75	GTTAATTGAAAGAGA
SO2445	Thiamin biosynthesis protein ThiC (thiC)	0.66	19.66	−371	−	2.13	GTTATTAAATTTTAA
				−118	−	2.18	GTTAATAGACGGCGA
Cell envelope							
SO4179	Glycosyl transferase, group 2 family protein	0.41	0.47	−55	−	2.15	GTTCAGCCAATTGAT
SO1199	Phosphoglucosamine mutase (glmM)	1.33	2.31	−47	+	2.57	GTTAAGTATTTCATT
SO2757	Membrane protein, putative	NA	0.29	−794	−	2.56	GTTAACGATCTGCCA
SO1102	TonB-dependent receptor C-terminal region domain lipoprotein	6.34	11.8	−110	−	2.22	GTTAAGCCTGATATA
				−11	+	2.02	GCTAACAAAAAGTTT
SO1673	Outer membrane protein OmpW, putative	5.42	11.04	−272	+	2.17	GTAAATGAAATGTAA

Cellular processes							
SO1278	Methyl-accepting chemotaxis protein	2.79	2	−46	−	2.01	GTTCATTTTATTTTT
				−324	+	2.01	GTTAATGTTAATGTA
SO1434	Methyl-accepting chemotaxis protein	0.47	0.37	−1012	+	2.36	GTTCACACAATCCCA
SO4557	Methyl-accepting chemotaxis protein	0.11	0.35	−155	−	2.05	ATTAATAAATATTTA
SO1961	Maltose *O*-acetyltransferase (maa)	1.56	2.77	−98	+	2.29	GTTAGCTAAATGGTA
SO0866	Minor curlin subunit CsgB, putative	61.27	7.13	−150	−	2.72	GTTAATCGTATGAAA
SO4317	RTX toxin, putative	1.72	4.18	−291	+	2.07	ATTCATAAAATTTTA
SO2389	Multidrug resistance protein D (emrD)	0.07	0.9	−217	+	2.58	GTTAACAAACAGCTT
				−65	−	2.54	GTTAATCATCTTGAT
SO4146	Toxin secretion ABC transporter protein, HlyB family	3.1	3.73	−53	+	2.76	GTTAATTAAGATGTT
				−160	−	2.13	GTTAACATTATGTTT
SO4274	Undecaprenol kinase, putative	2.39	1.11	−81	−	3.28	GTTAATTATATTTAA
SO1917	Multidrug resistance protein, putative	2.07	0.92	−10	+	2.37	GTTAATTGAGGCAAA
Central intermediary metabolism							
SO3705	5-Methylthioadenosine nucleosidase/ *S*-adenosylhomocysteine nucleosidase, putative	0.23	0.22	−300	−	2.04	GTTAAGCCTTTGAGT
SO2185	Exopolyphosphatase (ppx)	0.92	2.51	−623	−	2.37	GTTCACGATTTCATT
SO0314	Ornithine decarboxylase, inducible (speF)	0.1	0.02	−682	−	2.76	GTTAATTCATTTTGA
				−506	+	2.37	GTTAATTAAAAATAA
SO1870	Biosynthetic arginine decarboxylase (speA)	0.43	0.93	−815	+	2.39	GTTTATTATATTTTA
				−912	−	2.12	GTTCAATATTTTTTA
SO3872	Arylsulfate sulfotransferase	0.38	0.45	−167	+	2.05	GTTAATTTATATAAA

(*continued*)

Table 3.3 *(Continued)*

ORF	Gene Product	Ratio (ARCA/WT) Anaerobic[a]	Aerobic	Start[b]	Strand	Z^c	Motif
DNA metabolism							
SO1066	Extracellular nuclease	2.42	8.07	−368	−	2.05	TTTAATTATTTTGAA
				−29	+	2.16	GTTAATAAAAATTAG
SO1844	Extracellular nuclease, putative	2.2	1.23	−48	−	2.22	GTTAAGACTTTTCGA
Energy metabolism							
SO1812	Methionine gamma-lyase (mdeA)	11.36	21.69	−68	+	2.25	GTTATTTAAAAGATA
SO4674	2-Amino-3-ketobutyrate coenzyme A ligase (kbl)	0.4	0.63	−139	+	2.53	GTTAAGCATTAGTTT
SO3299	Pal/histidase family protein	0.25	0.22	−50	+	3.02	GTTAATTAATTTTGA
SO1421	Fumarate reductase flavoprotein subunit (ifcA-1)	0.46	0.72	−89	+	2.69	GTTAAGTGAATTTTT
SO2099	Quinone-reactive Ni/Fe hydrogenase, small subunit precursor (hoxK)	3.55	1.56	−307	−	2.86	GTTAATAAATTCAAA
				−131	+	3.26	GTTAATTAAATGTCA
SO2727	Cytochrome *c*3	0.4	0.47	−204	+	2.17	GTTATTCAAATGTAA
SO1427	Decaheme cytochrome c	0.07	0.17	−319	−	2.23	ATTAATTAAATGAAA
				−230	−	3.1	GTTAATAAAATGTTT
				−71	−	2.72	GTTAACGAATGTTAA
				−5	+	2.84	GTTAACCATAAGGCA
SO0344	Methylcitrate synthase (prpC)	2.07	1.07	−13	−	2.59	GTTAACGCTATGGTT
SO3496	Aldehyde dehydrogenase	2.48	1.46	−219	+	2.03	GTTAAGGGTGGCTAA
				−188	−	2.85	GTTCATTATTTCTAA
SO1006	Dienelactone hydrolase family protein	2.26	1.87	−227	+	2.17	GTTATTGAAATGTAA
SO1483	Malate synthase A (aceB)	3.72	1.57	−712	−	2.37	GTTAACCGTTTTCCA

Fatty acid and phospholipid metabolism							
SO0572	Enoyl–CoA hydratase/isomerase family protein	7.98	13.05	−113	−	2.55	GTTAAGTATTAGTGT
SO2395	Acyl–CoA dehydrogenase family protein	1.07	7.11	−198	+	2.37	GTTAATCGTGAGCTA
Hypothetical protein							
SO0563	Hypothetical protein	NA	0.21	−349	+	2.99	GTTAACAGAATTTTA
SO0912	Hypothetical protein	NA	0.23	−507	+	2.63	GTTCATGGTAAGTTA
SO1546	Hypothetical protein	NA	0.26	−261	−	2.01	GTTAACATTGTTTAA
SO0787	Hypothetical protein	NA	0.46	−225	+	2.32	GTTAACTTAAAGTAA
				−125	+	2.02	ATTAAGTAAAATTAA
SO3511	Hypothetical protein	NA	0.47	−122	+	2.07	TTTAATAATATTAAA
SO2460	Hypothetical protein	28.07	0.32	−23	+	2.16	GTAAACAAATTGTAA
SO1479	Hypothetical protein	19.47	5.62	−13	−	2.8	GTTAATAATGATTCA
SO2930	Hypothetical protein	15.88	1.57	−19	+	3.01	GTTAACGAAATTATA
SO0306	Hypothetical protein	8.12	6.45	−176	+	2.01	GTTATTAAATTGTGA
SO3395	Hypothetical protein	7.11	15.52	−534	−	2.32	GTTAATGCTTGGCTA
SO2446	Hypothetical protein	6.04	0.23	−33	+	2.29	GTTCAGAGTTTGTGA
SO1970	Hypothetical protein	5.45	6.19	−116	−	2.58	GTTAACTGTGAGGCA
SO4592	Hypothetical protein	3.68	1.29	−296	+	2.17	ATTAATTAAAACTTA
SO2199	Hypothetical protein	3.33	1.76	−85	−	2.1	GTGAATAAAATGTTT
SO2002	Hypothetical protein	2.95	1.87	−358	+	2.96	GTTAATAAAATGCCA
				−242	+	2.77	GTTAACTAACGGAAA
				−13	+	3.01	GTTAACTATATCTTT
SO1517	Hypothetical protein	2.66	0.28	−204	−	2.55	GTTAATCCATAGAAA
SO1944	Hypothetical protein	2.1	3.57	−49	−	3	GTTAAGTAATTGTAA
SO2076	Hypothetical protein	0.9	4.59	−42	−	2.88	GTTAATTAAAGCGAA
SO4018	Hypothetical protein	1.91	2.07	−99	+	2.26	GTTAAGTGTTTGAGT

(*continued*)

Table 3.3 *(Continued)*

ORF	Gene Product	Ratio (ARCA/WT)		Start[b]	Strand	Z^c	Motif
		Anaerobic[a]	Aerobic				
SO0181	Hypothetical protein	0.47	0.61	−61	−	2.49	GTTCATGAATATCAA
SO0076	Hypothetical protein	0.4	0.51	1	+	2.37	GTTAAGGGTAGTTAA
SO0712	Hypothetical protein	0.4	0.52	−240	−	2.06	ATTAATCAAAATTTA
SO0091	Hypothetical protein	0.37	0.19	−117	−	2.44	GTTAAGACTTATTCA
SO4180	Hypothetical protein	0.3	0.23	−256	+	2.04	GTTACTTATTTGTTT
SO1004	Hypothetical protein	0.21	0.53	−183	−	2.17	GTTATTGAAATGTAA
SO0120	Hypothetical protein	0.12	0.12	−233	+	2.61	GTTAACCATGTCGAA
SO0403	Hypothetical protein	0.08	0.03	−724	+	2.2	GTTATTTATTTTTAA
				−466	+	2.06	GTTAATTAAATTAGC
SO1188	Conserved hypothetical protein	NA	0.44	−302	−	2.77	GTTCATTAATTCAAA
SO0946	Conserved hypothetical protein	30.91	4.29	−150	−	2.1	GTAAATAAAATGTAT
SO3480	Hypothetical protein phosphatase	11.35	4.02	−52	−	2.95	GTTAATAATTAGATA
				−294	+	2.28	TTTAATTAAAATTTA
SO3278	Conserved hypothetical protein	9.63	3.91	−25	−	2.38	GTTAATTTTATGTAA
SO4145	Conserved hypothetical protein	7.41	5.16	−193	−	2.76	GTTAATTAAGATGTT
				−86	+	2.13	GTTAACATTATGTTT
SO3846	Conserved hypothetical protein	6.05	5.61	−180	−	2.56	GTTAACTATTGTCTT
				−163	−	2.57	GTTAACTATTGCGTT
SO3514	Hypothetical TonB-dependent receptor	3.97	6.9	−87	+	2.88	GTTAATTATTGTGTA
SO3091	Conserved hypothetical protein	3.76	4.56	−17	+	2.09	GTTAAATAAATGGCT
SO1064	Hypothetical WD domain protein	3.7	1.91	−43	+	2.71	GTTAACTGTGATTAA
SO2042	Conserved hypothetical protein	3.17	0.65	−129	+	2.03	GTTGATAAAGTGTAA
SO3280	Conserved hypothetical protein	3.06	2.1	−234	+	2.56	GTTAACGCTAATCTA
SO0440	Conserved hypothetical protein	2.96	2.94	−14	+	2.07	GTTAATTATCTAAAA
SO1267	Hypothetical glutamine amidotransferase	2.93	1.97	−263	+	2.09	GTAAATAATATTTTT
SO2711	Conserved hypothetical protein	2.81	3.34	−41	−	2.23	GTTAACAAAAAGTTG

SO4366	Conserved hypothetical protein	2.77	2.23	−179	−	3.03	GTTAATAAAATGCAA
SO0308	Conserved hypothetical protein	2.77	3.49	−133	−	3.03	GTTAATTAAAAGGGA
SO1399	Conserved hypothetical protein	2.2	3.13	−97	+	2.17	GTTATCTCAATGTTA
SO3507	Conserved hypothetical protein	2.06	1.69	−74	−	3.12	GTTAACTCAATGTTA
SO4563	Conserved hypothetical protein	2.03	1.65	−281	−	2.55	GTTCATCCATTGTAA
SO2041	Conserved hypothetical protein	0.77	3.62	−117	−	2.03	GTTGATAAAGTGTAA
SO2144	Conserved hypothetical protein	0.74	2.31	−370	−	2.57	GTTAACCAATACACA
SO1443	Conserved hypothetical protein	0.48	0.47	−173	+	2.86	GTTAATAGTATTATA
SO4196	Conserved hypothetical protein	0.43	0.61	−196	+	2.48	GTTAACACTTATGTT
				−68	−	2.55	GTTAAGTAAGTTAAT
SO1873	Hypothetical short chain dehydrogenase	0.39	0.24	−44	+	2.15	GTTAATTGTCGCAAT
SO4512	Conserved hypothetical protein	0.27	0.09	−295	−	2.84	GTTAATAATGTGTTT
				−211	+	2.42	GTTAACGCTTTTGGA
				−63	−	2.07	GTTATCAAAAAGTGA
SO3085	Conserved domain protein	2.05	2.39	−19	−	2.01	GTTGATAATATTTAT
SO2064	Conserved domain protein	0.6	3.33	−157	−	2.39	GTTAATCATCTTGGT
Other categories							
SO0643	Transposase, putative	2	1.81	−332	+	2.06	GTTAAATAATTCAAA
Protein fate							
SO3106	Cold-active serine alkaline protease (aprE)	19.93	19.72	−242	−	2.59	GTTAAGAGAATTTTT
				−111	−	2.3	GTAAATTAATTGTTA
				−24	−	2.66	GTTAATCATTATTAT
SO3942	Serine protease, HtrA/DegQ/DegS family	7.04	6.66	−133	−	2.92	GTTCATTAATATTAA
SO1915	Serine protease, subtilase family	4.84	3.8	−135	+	2.84	GTTAATAATGTGTTT
SO0491	Peptidase, M13 family	2.98	1.55	−128	−	2.52	GTTAATGCAAAGTCT
SO0867	Serine protease, subtilase family	2.72	2.68	−282	+	2.72	GTTAATCGTATGAAA
SO4537	Hypothetical Zn-dependent peptidase	2.18	1.33	−402	−	2.52	GTTAATAGATTTAAT

(*continued*)

Table 3.3 *(Continued)*

ORF	Gene Product	Ratio (ARCA/WT) Anaerobic[a]	Aerobic	Start[b]	Strand	Z^c	Motif
SO3844	Peptidase, M13 family	1.14	2.2	−524	+	2.59	GTTCAGCATAAGGTA
SO2093	Hydrogenase accessory protein HypB (hypB)	1.32	6.12	−91	−	2.32	GTTAAGTGTGATACA
SO1065	FKBP-type peptidyl–prolyl cis–trans isomerase FkpA (fkpA)	1.6	2.92	−93	−	2.71	GTTAACTGTGATTAA
SO1127	Chaperone protein DnaJ (dnaJ)	1.64	3.02	−67	−	2.06	ATTAACTATTTGAAA
SO3659	Thiol/disulfide interchange protein, putative	3.8	2.41	−103	−	2.25	GTTAACAATGGCGCT
SO1062	Polypeptide deformylase (def-2)	2.21	1	−20	−	2.16	GTTCATCGTTTGGCT
Protein synthesis							
SO2085	Phenylalanyl-tRNA synthetase, alpha subunit (pheS)	1.06	2.08	−364	−	2.99	GTTAACAATAATAAA
Purines, pyrimidines, nucleosides, and nucleotides							
SO2001	5′-Nucleotidase (ushA)	1.12	2.66	−266	+	2.82	GTTCATTATTTTTTT
SO3565	2′,3′-Cyclic-nucleotide 2′-phosphodiesterase (cpdB)	0.31	0.4	−64	−	2.22	GTTAATTCATGCGCT
				−106	−	2.42	GTTAACGAACGGGGA
				−241	−	2.27	GTTAATTAATTGTTG
				−266	−	2	GTTGATGATAATTAA
				−538	+	2.36	GTTCACAGTCATTCA

Regulatory functions							
SO1946	Transcriptional regulatory protein PhoP (phoP)	2.53	1.72	−105	−	2.19	GTTCATCAATGTCGA
SO3988	Aerobic respiration control protein ArcA (arcA)	0.11	0.07	−392	+	2.57	GTTAACAAATGCCTA
SO1661	Transcriptional regulator, LysR family	6.49	5.2	−203	+	2.36	GTTAAATAATTGTTA
SO0864	Transcriptional regulator, LuxR family	3.35	3.72	−175	+	2.1	GTTAATTTTTTGTTT
SO3516	Transcriptional regulator, LacI family	1.7	3.73	−321	−	2.88	GTTAATTATTGTGTA
SO1422	Transcriptional regulator, LysR family	0.26	0.12	−11	−	2.69	GTTAAGTGAATTTTT
SO1935	Regulator of nucleoside diphosphate kinase (rnk)	0.38	0.5	−290	+	2.24	GTTAAGTGTGAGAAT
SO1806	psp operon transcriptional activator (pspF)	5.86	3.11	−74	−	3.1	GTTAATAAAATGTTT
SO3689	Sigma-54 dependent nitrogen response regulator	2.64	3.55	−17	+	2.72	GTTAATAAATTTGCT
Signal transduction							
SO0570	Response regulator	2.59	1.47	−443	+	2.55	GTTAAGTATTAGTGT
Transcription							
SO0208	RNA binding protein	2.36	3.32	−302	−	2.45	GTTCATCAAGTGATT
				−30	+	2.76	GTTAACAGTTATTTA
SO3840	RNA polymerase sigma-70 factor, ECF subfamily	0.39	0.98	−73	+	2.04	GTGAATAATATTTTT
Transport and binding proteins							
SO3063	Sodium/alanine symporter family protein	11.32	9.27	−158	−	2.1	GTTAAATGTATGATA
				−92	+	2.18	GTTCACACAAGTCTT

(*continued*)

Table 3.3 *(Continued)*

ORF	Gene Product	Ratio (ARCA/WT) Anaerobic[a]	Aerobic	Start[b]	Strand	Z^c	Motif
SO1560	Phosphate binding protein	1.18	2.11	−17	−	2.13	GCTAACTAATTTGTA
SO3134	C4-dicarboxylate binding periplasmic protein (dctP)	1.62	2.31	−151	−	2.03	GTTAATAAAGTGTAG
SO1522	L-lactate permease, putative	2.01	1.21	−619	−	2.88	GTTAATAAAAATATT
				−227	+	2.6	GTTAAGAAAATCCCA
SO3099	Long-chain fatty acid transport protein, putative	0.07	0.16	−265	−	3.24	GTTAATTAAATTATA
				−240	−	2.68	GTTAATACTTTGTGA
				−144	−	2.86	GTTAATTAAAACATT
SO1072	Chitin binding protein, putative	2.3	0.9	−237	−	2.05	GTTAACAGGATGTAA
SO1307	Aquaporin *Z* (aqpZ)	3.13	0.75	−535	−	2.71	GTTAATCATGTGTTT
				−417	−	2.07	TTTAATTATTTGAAA
SO1100	Extracellular solute binding protein, family 7	2.86	2.36	−84	+	2.28	GTTAACAATATGTTC
SO1750	ABC transporter, ATP binding protein	2.12	2.28	−176	+	2.71	GTTAATTATTGTCAA
SO0450	Major facilitator family protein	0.33	0.79	−142	−	2.13	GATAATGAAATTTAA
Unclassified							
SO4570	Conserved domain protein	NA	0.49	−131	−	3.1	GTTAATAAAATGTTT
Unknown function							
SO0900	Oxidoreductase, aldo/keto reductase family	4.34	1.25	−397	−	2.74	GTTAACAAACATCTA
				−233	−	2	GTTAACTCTAAATCA
SO3497	Aminotransferase, class III	1.14	2.35	−45	−	2.03	GTTAAGGGTGGCTAA
				−76	+	2.85	GTTCATTATTTCTAA

SO3301	Flavocytochrome c flavin subunit	0.43	0.44	−367	−	2.75	GTTAATAGTATGCAA
SO4136	Decarboxylase, pyridoxal dependent	0.38	0.65	−204	+	2.69	GTTAACAAACATCAA
				−71	+	2.05	GTTAACTAAAAAAAT
SO4463	Prolyl 4-hydroxylase, alpha subunit domain protein	7.02	3.52	−61	+	2.15	GTAAATAAAATGTTT
SO4457	GGDEF domain protein	5.4	5.68	−58	−	2.22	GTTAACGCAATCCCT
SO3489	GGDEF domain protein	3.55	3.41	−148	+	2.8	GTTAACTATCTGTCT
SO3976	GAF domain protein	2.38	2.46	−366	+	2.14	GTGAATTAATAGTAA
SO3912	TIM-barrel protein, yjbN family	2.26	1.26	−19	+	2.08	GTTAGCAAATTTTAA
SO4324	GGDEF domain protein	0.75	2.03	−35	+	2.71	GTTAATTATAGCGTT
SO0559	MaoC domain protein	0.32	0.38	−39	+	3.22	GTTAATTAAAAGGTA
SO2907	TonB-dependent receptor domain protein	0.25	0.48	−441	+	2.03	GTTAATTCAACGAAA
				−267	−	2.1	GTCAATAAAATGTTT
				−163	−	3.1	GTTAATAAAATGTTT
SO2469	Hypothetical TonB-dependent receptor	0.1	NA	−86	+	2.69	GTTAATGATAGTTTT
				−3	+	2.54	GTTAAGGGAATGAAA

[a]Relative gene expression is presented as the mean ratio of the fluorescence intensity of the *arcA* deletion mutant (ARCA) to that of the parental strain (WT), and NA means that there is not any expression value available.

[b]The start site of the predicted ArcA motif away from the gene translation start site.

[c]The statistical *Z*-score of the predicted motif from the genomic-wide scan.

Further DNA binding experiments have confirmed the ArcA-P binding motifs predicted here, which include a transcriptional regulator (SO1661), decaheme cytochrome *c* (SO1427), and *hoxK* [120]. Further investigation will be required to verify the functionality of the other predicted ArcA binding motifs in *S. oneidensis* MR-1. We believe these 148 genes with altered expression in the *arcA* deletion strain are still a subset of the ArcA regulon in *S. oneidensis* since some ArcA-regulated genes may not show expression differences under the culture conditions tested. For example, a malate oxidoreductase (*sfcA*) was predicted with a strong binding motif, but the gene expression values were lower than 2-fold and higher than 0.5-fold under both anaerobic and aerobic conditions. The binding motif predicted in *sfcA* was also confirmed by DNA binding experiments [120]. It is also worth mentioning that *pflBA* has the ArcA-P binding site in *E. coli* but no predicted ArcA-P binding sites in *S. oneidensis* MR-1 [92]. The DNA binding experiment also demonstrated that *pflBA* does not have a strong ArcA-P binding site [120].

3.4.6 Conclusions

In summary, we used microarray-based gene expression profiling to examine the transcriptome for an *arcA* null mutant compared to the parental *S. oneidensis* DSP10 strain under both aerobic and anaerobic growth conditions. Transcriptome profiling revealed a total of 654 (294 down regulated; 360 upregulated) and 504 (135; 369) open reading frames (ORFs) that were differentially expressed in an *arcA* deletion mutant relative to the parental strain under aerobic and anaerobic respiratory conditions, respectively. By integrating computational motif prediction tools and microarray analyses, we predicted an *S. oneidensis* ArcA regulon consisting of as many as 148 *S. oneidensis* genes (46 as a positive regulator and 102 as a negative regulator), which included a number of genes shown to be under the direct control of ArcA in other bacteria. Our results also demonstrated that ArcA in *S. oneidensis* acts as both a positive and negative regulator for genes associated with various other functional categories. Both transcriptome data analysis and motif predictions suggest the Arc two-component signal transduction system in *S. oneidensis* regulates a large number of genes that are different from those regulated by ArcA in *E. coli*, although they do have overlapping regulatory functions for a small subset of genes. *S. oneidensis* is typically found at oxic–anoxic interfaces in nature such as sediments and bodies of water where oxygen is limited or absent [60] whereas *E. coli* primarily lives in the mammalian gut [83]. Different living environments for *S. oneidensis* and *E. coli* might result in the observed differences in ArcA regulon compositions during evolution for environmental adaptation. Finally, phenotype characterization indicated that ArcA enables *S. oneidensis* to resist oxidative stress.

3.5 FUTURE PROSPECTS OF CHIP TECHNOLOGY

The applications of high-throughput technologies and functional genomics have proven to be great successes in biological studies. Array technology can be considered

a milestone since it has revolutionized biological research (Fig. 3-1). Future development of economically feasible custom chips will permit functional genomics techniques to become routine lab tools. A standardized protocol for data analysis and information mining needs to be completed in the future as well.

ACKNOWLEDGMENTS

The authors acknowledge Dr. Jizhong Zhou for the construction of *S. oneidensis* MR-1 microarrays and the Miami University CFR grant.

REFERENCES

1. Sanger F. Determination of nucleotide sequences in DNA. *Science* 1981;214:1205–1210.
2. Sanger F, Coulson AR, Hong GF, Hill DF, Petersen GB. Nucleotide sequence of bacteriophage lambda DNA. *J Mol Biol* 1982;162:729–773.
3. Smith HO, Tomb JF, Dougherty BA, Fleischmann RD, Venter JC. Frequency and distribution of DNA uptake signal sequences in the *Haemophilus influenzae* Rd genome. *Science* 1995;269:538–540.
4. Jordan B. Historical background and anticipated developments. *Ann NY Acad Sci* 2002;975:24–32.
5. Gress TM, Hoheisel JD, Lennon GG, Zehetner G, Lehrach H. Hybridization fingerprinting of high-density cDNA-library arrays with cDNA pools derived from whole tissues. *Mamm Genome* 1992;3:609–619.
6. Butte A. The use and analysis of microarray data. *Nat Rev Drug Discov* 2002;1:951–960.
7. Bard F, Casano L, Mallabiabarrena A, Wallace E, Saito K, Kitayama H, et al. Functional genomics reveals genes involved in protein secretion and Golgi organization. *Nature* 2006;439:604–607.
8. Cassell GH, Mekalanos J. Development of antimicrobial agents in the era of new and reemerging infectious diseases and increasing antibiotic resistance. *JAMA* 2001;285: 601–605.
9. Aharoni A, Vorst O. DNA microarrays for functional plant genomics. *Plant Mol Biol* 2002; 48:99–118.
10. Chin KV, Kong AN. Application of DNA microarrays in pharmacogenomics and toxicogenomics. *Pharm Res* 2002;19:1773–1778.
11. Katsuma S, Tsujimoto G. Genome medicine promised by microarray technology. *Expert Rev Mol Diagn* 2001;1:377–382.
12. Macgregor PF. Gene expression in cancer: the application of microarrays. *Expert Rev Mol Diagn* 2003;3:185–200.
13. Nocito A, Kononen J, Kallioniemi OP, Sauter G. Tissue microarrays (TMAs) for high-throughput molecular pathology research. *Int J Cancer* 2001;94:1–5.
14. Smith L, Greenfield A. DNA microarrays and development. *Hum Mol Genet* 2003;12 Spec No 1: R1–R8.
15. Stenger DA, Andreadis JD, Vora GJ, Pancrazio JJ. Potential applications of DNA microarrays in biodefense-related diagnostics. *Curr Opin Biotechnol* 2002;13: 208–212.

16. Stoughton RB. Applications of DNA microarrays in biology. *Annu Rev Biochem* 2005; 74:53–82.
17. Ye RW, Wang T, Bedzyk L, Croker KM. Applications of DNA microarrays in microbial systems. *J Microbiol Methods* 2001;47:257–272.
18. Zammatteo N, Hamels S. De Longueville F, Alexandre I, Gala JL, Brasseur F, et al. New chips for molecular biology and diagnostics. *Biotechnol Annu Rev* 2002;8:85–101.
19. Nygren PA, Uhlen M. Scaffolds for engineering novel binding sites in proteins. *Curr Opin Struct Biol* 1997;7:463–469.
20. Brody EN, Willis MC, Smith JD, Jayasena S, Zichi D, Gold L. The use of aptamers in large arrays for molecular diagnostics. *Mol Diagn* 1999;4:381–388.
21. Seetharaman S, Zivarts M, Sudarsan N, Breaker RR. Immobilized RNA switches for the analysis of complex chemical and biological mixtures. *Nat Biotechnol* 2001;19: 336–341.
22. Wilson DS, Nock S. Recent developments in protein microarray technology. *Angew Chem Int Ed Engl* 2003;42:494–500.
23. Kodadek T. Protein microarrays: prospects and problems. *Chem Biol* 2001;8:105–115.
24. Haab BB, Dunham MJ, Brown PO. Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions. *Genome Biol* 2001;2:RESEARCH0004.
25. Joos TO, Bachmann J. The promise of biomarkers: research and applications. *Drug Discov Today* 2005;10:615–616.
26. Wiltshire S, O'Malley S, Lambert J, Kukanskis K, Edgar D, Kingsmore SF, et al. Detection of multiple allergen-specific IgEs on microarrays by immunoassay with rolling circle amplification. *Clin Chem* 2000;46:1990–1993.
27. Zhou JT, Dorothea K. *DNA Microarray Technology.* Hoboken, NJ: John Wiley & Sons, Inc., 2004.
28. Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. Light-directed, spatially addressable parallel chemical synthesis. *Science* 1991;251:767–773.
29. McGall GH, Fidanza JA. Photolithographic synthesis of high-density oligonucleotide arrays. In: Rampal JB, editor. *DNA Arrays—Methods and Protocols*. Totowa, NJ: Nuts & Bolts, Humana Press, 2001.
30. Xu D, Li G, Wu L, Zhou J, Xu Y. PRIMEGENS: robust and efficient design of gene-specific probes for microarray analysis. *Bioinformatics* 2002;18:1432–1437.
31. Skaletsky SRaHJ. Primer3 on the WWW for general users and for biologist programmers. In: Krawetz SMS, editor. *Bioinformatics Methods and Protocols: Methods in Molecular Biology.* Totowa, NJ: Humana Press, 2000, pp. 365–386.
32. Li F, Stormo GD. Selection of optimal DNA oligos for gene expression arrays. *Bioinformatics* 2001;17:1067–1076.
33. Emrich SJ, Lowe M, Delcher AL. PROBEmer: a Web-based software tool for selecting optimal DNA oligos. *Nucleic Acids Res* 2003;31:3746–3750.
34. Li X, He Z, Zhou J. Selection of optimal oligonucleotide probes for microarrays using multiple criteria, global alignment and parameter estimation. *Nucleic Acids Res* 2005;33: 6114–6123.
35. Herold KE, Rasooly A. Oligo design: a computer program for development of probes for oligonucleotide microarrays. *Biotechniques* 2003;35:1216–1221.

36. Chou HH, Hsia AP, Mooney DL, Schnable PS. Picky: oligo microarray design for large genomes. *Bioinformatics* 2004;20:2893–2902.
37. Wang X, Seed B. Selection of oligonucleotide probes for protein coding sequences. *Bioinformatics* 2003;19:796–802.
38. Rouillard JM, Zuker M, Gulari E. OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. *Nucleic Acids Res* 2003;31: 3057–3062.
39. Reymond N, Charles H, Duret L, Calevro F, Beslon G, Fayard JM. ROSO: optimizing oligonucleotide probes for microarrays. *Bioinformatics* 2004;20:271–273.
40. Rimour S, Hill D, Militon C, Peyret P. GoArrays: highly dynamic and efficient microarray probe design. *Bioinformatics* 2005;21:1094–1103.
41. Chung WH, Rhee SK, Wan XF, Bae JW, Quan ZX, Park YH. Design of long oligonucleotide probes for functional gene detection in a microbial community. *Bioinformatics* 2005;21:4092–4100.
42. Borneman J, Chrobak M, Della Vedova G, Figueroa A, Jiang T. Probe selection algorithms with applications in the analysis of microbial communities. *Bioinformatics* 2001;17(Suppl. 1): S39–S48.
43. Dent GW, O'Dell DM, Eberwine JH. Gene expression profiling in the amygdala: an approach to examine the molecular substrates of mammalian behavior. *Physiol Behav* 2001;73:841–847.
44. Warrington JA, Dee S, Trulson M. Large-scale genomic analysis using Affymetrix GeneChip® probe arrays. In: Schena M, editor. *Microarray Biochip Technology.* Natick, MA: Eaton Publishing, 2000.
45. Schermer MJ. Confocal scanning microscopy in microarray detection. In: Schena M, editor. *DNA Microarrays*. New York: Oxford University Press, 1999.
46. Quackenbush J. Microarray data normalization and transformation. *Nat Genet* 2002;32 (Suppl.): 496–501.
47. Li SS, Bigler J, Lampe JW, Potter JD, Feng Z. FDR-controlling testing procedures and sample size determination for microarrays. *Stat Med* 2005;24:2267–2280.
48. Speed T. *Statistical analysis of gene expression microarray data*. Chapman&Hall/CRC, 2003.
49. Wan XF, Verberkmoes NC, McCue LA, Stanek D, Connelly H, Hauser LJ, et al. Transcriptomic and proteomic characterization of the Fur modulon in the metal-reducing bacterium *Shewanella oneidensis*. *J Bacteriol* 2004;186:8385–8400.
50. Clark ME, He Q, He Z, Huang KH, Alm EJ, Wan X-F, Hazen TC, Arkin AP, Wall JD, Zhou J, Fields MW. Temporal transcriptomic analysis of *Desulfovibrio vulgaris* Hildenborough transition into stationary phase growth during electron donor depletion. *Appl Environ Microbiol* 2006;72:5578–5588.
51. Yang YH, Speed T. Design issues for cDNA microarray experiments. *Nat Rev Genet* 2002;3:579–588.
52. Conlon EM, Liu XS, Lieb JD, Liu JS. Integrating regulatory motif discovery and genome-wide expression analysis. *Proc Natl Acad Sci USA* 2003;100:3339–3344.
53. Xu D, Olman V, Wang L, Xu Y. EXCAVATOR: a computer program for efficiently mining gene expression data. *Nucleic Acids Res* 2003;31:5582–5589.

54. Thompson W, Rouchka EC, Lawrence CE. Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Res* 2003;31:3580–3585.

55. Hughes JD, Estep PW, Tavazoie S, Church GM. Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J Mol Biol* 2000;296:1205–1214.

56. Roth FP, Hughes JD, Estep PW, Church GM. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat Biotechnol* 1998;16:939–945.

57. Liu X, Brutlag DL, Liu JS. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput* 2001:127–138.

58. Wang T, Stormo GD. Identifying the conserved network of cis-regulatory sites of a eukaryotic genome. *Proc Natl Acad Sci USA* 2005;102:17400–17405.

59. Zou M, Conzen SD. A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics* 2005;21: 71–79.

60. Heidelberg JF, Paulsen IT, Nelson KE, Gaidos EJ, Nelson WC, Read TD,et al. Genome sequence of the dissimilatory metal ion-reducing bacterium *Shewanella oneidensis*. *Nat Biotechnol* 2002;20:1118–1123.

61. Xing B, van der Laan MJ. A statistical method for constructing transcriptional regulatory networks using gene expression and sequence data. *J Comput Biol* 2005;12:229–246.

62. Chang WC, Li CW, Chen BS. Quantitative inference of dynamic regulatory pathways via microarray data. *BMC Bioinformatics* 2005;6:44.

63. Zhou X, Wang X, Pal R, Ivanov I, Bittner M, Dougherty ER. A Bayesian connectivity-based approach to constructing probabilistic gene regulatory networks. *Bioinformatics* 2004;20:2918–2927.

64. Mehra S, Hu WS, Karypis G. A Boolean algorithm for reconstructing the structure of regulatory networks. *Metab Eng* 2004;6:326–339.

65. Missal K, Cross MA, Drasdo D. Gene network inference from incomplete expression data: transcriptional control of hematopoietic commitment. *Bioinformatics* 2006;22:731–738.

66. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N,et al. The transcriptional landscape of the mammalian genome. *Science* 2005;309:1559–1563.

67. Aracena J, Gonzalez M, Zuniga A, Mendez MA, Cambiazo V. Regulatory network for cell shape changes during Drosophila ventral furrow formation. *J Theor Biol* 2006;239:49–62.

68. Gutierrez-Rios RM, Rosenblueth DA, Loza JA, Huerta AM, Glasner JD, Blattner FR,et al. Regulatory network of *Escherichia coli*: consistency between literature knowledge and microarray profiles. *Genome Res* 2003;13:2435–2443.

69. Saal LH, Troein C, Vallon-Christersson J, Gruvberger S, Borg A, Peterson C. BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data. *Genome Biol* 2002;3:SOFTWARE0003.

70. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. *Science* 1995;270:484–487.

71. Tuteja R, Tuteja N. Serial analysis of gene expression: applications in malaria parasite, yeast, plant, and animal studies. *J Biomed Biotechnol* 2004;2004:106–112.

72. Horak CE, Snyder M. ChIP–chip: a genomic approach for identifying transcription factor binding sites. *Methods Enzymol* 2002;350:469–483.
73. Qian J, Lin J, Luscombe NM, Yu H, Gerstein M. Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data. *Bioinformatics* 2003;19:1917–1926.
74. Hanlon SE, Lieb JD. Progress and challenges in profiling the dynamics of chromatin and transcription factor binding with DNA microarrays. *Curr Opin Genet Dev* 2004;14: 697–705.
75. Mockler TC, Chan S, Sundaresan A, Chen H, Jacobsen SE, Ecker JR. Applications of DNA tiling arrays for whole-genome analysis. *Genomics* 2005;85:1–15.
76. Li L, Wang X, Stolc V, Li X, Zhang D, Su N,et al. Genome-wide transcription analyses in rice using tiling microarrays. *Nat Genet* 2006;38:124–129.
77. Stolc V, Samanta MP, Tongprasit W, Sethi H, Liang S, Nelson DC,et al. Identification of transcribed sequences in *Arabidopsis thaliana* by using high-resolution genome tiling arrays. *Proc Natl Acad Sci USA* 2005;102:4453–4458.
78. Sun LV, Chen L, Greil F, Negre N, Li TR, Cavalli G,et al. Protein-DNA interaction mapping using genomic tiling path microarrays in Drosophila. *Proc Natl Acad Sci USA* 2003;100:9428–9433.
79. Selzer RR, Richmond TA, Pofahl NJ, Green RD, Eis PS, Nair P,et al. Analysis of chromosome breakpoints in neuroblastoma at sub-kilobase resolution using fine-tiling oligonucleotide array CGH. *Genes Chromosomes Cancer* 2005;44:305–319.
80. Schumacher A, Kapranov P, Kaminsky Z, Flanagan J, Assadzadeh A, Yau P,et al. Microarray-based DNA methylation profiling: technology and applications. *Nucleic Acids Res* 2006;34:528–542.
81. Lippman Z, Gendrel AV, Colot V, Martienssen R. Profiling DNA methylation patterns using genomic tiling microarrays. *Nat Methods* 2005;2:219–224.
82. Bertone P, Gerstein M, Snyder M. Applications of DNA tiling arrays to experimental genome annotation and regulatory pathway discovery. *Chromosome Res* 2005;13: 259–274.
83. Lynch AS, Lin ECC. Responses to molecular oxygen. In: Neidhardt FC, Curtiss R III, Ingraham JL, Lin ECC, Low KB, Magasanik B, Reznikoff WS, Riley M, Schaechter M, Umbarger HE, editors. *Escherichia coli* and Salmonella: cellular and molecular biology. Washington, DC: American Society for Microbiology, 1996; pp. 1526–1538.
84. Georgellis D, Lynch AS, Lin EC. *In vitro* phosphorylation study of the arc two-component signal transduction system of *Escherichia coli*. *J Bacteriol* 1997;179:5429–5435.
85. Iuchi S, Lin EC. Mutational analysis of signal transduction by ArcB, a membrane sensor protein responsible for anaerobic repression of operons involved in the central aerobic pathways in *Escherichia coli*. *J Bacteriol* 1992;174:3972–3980.
86. Kwon O, Georgellis D, Lin EC. Phosphorelay as the sole physiological route of signal transmission by the arc two-component system of *Escherichia coli*. *J Bacteriol* 2000;182:3858–3862.
87. Kwon O, Georgellis D, Lynch AS, Boyd D, Lin EC. The ArcB sensor kinase of *Escherichia coli*: genetic exploration of the transmembrane region. *J Bacteriol* 2000; 182:2960–2966.
88. Bauer CE, Elsen S, Bird TH. Mechanisms for redox control of gene expression. *Annu Rev Microbiol* 1999;53:495–523.

89. Liu X, De Wulf P. Probing the ArcA-P modulon of *Escherichia coli* by whole genome transcriptional analysis and sequence recognition profiling. *J Biol Chem* 2004;279: 12588–12597.

90. Gao H, Wang Y, Liu X, Yan T, Wu L, Alm E, et al. Global transcriptome analysis of the heat shock response of *Shewanella oneidensis*. *J Bacteriol* 2004;186:7796–7803.

91. Lu S, Killoran PB, Fang FC, Riley LW. The global regulator ArcA controls resistance to reactive nitrogen and oxygen intermediates in *Salmonella enterica* serovar Enteritidis. *Infect Immun* 2002;70:451–461.

92. Lynch AS, Lin EC. Transcriptional control mediated by the ArcA two-component response regulator protein of *Escherichia coli*: characterization of DNA binding at target promoters. *J Bacteriol* 1996;178:6238–6249.

93. Park SJ, McCabe J, Turna J, Gunsalus RP. Regulation of the citrate synthase (gltA) gene of *Escherichia coli* in response to anaerobiosis and carbon supply: role of the *arcA* gene product. *J Bacteriol* 1994;176:5086–5092.

94. Park SJ, Tseng CP, Gunsalus RP. Regulation of succinate dehydrogenase (sdhCDAB) operon expression in *Escherichia coli* in response to carbon supply and anaerobiosis: role of ArcA and Fnr. *Mol Microbiol* 1995;15:473–482.

95. Wilde RJ, Guest JR. Transcript analysis of the citrate synthase and succinate dehydrogenase genes of *Escherichia coli* K12. *J Gen Microbiol* 1986;132: 3239–3251.

96. Wood D, Darlison MG, Wilde RJ, Guest JR. Nucleotide sequence encoding the flavoprotein and hydrophobic subunits of the succinate dehydrogenase of *Escherichia coli*. *Biochem J* 1984;222:519–534.

97. Cunningham L, Gruer MJ, Guest JR. Transcriptional regulation of the aconitase genes (acnA and acnB) of *Escherichia coli*. *Microbiology* 1997;143(Pt 12): 3795–3805.

98. Chao G, Shen J, Tseng CP, Park SJ, Gunsalus RP. Aerobic regulation of isocitrate dehydrogenase gene (icd) expression in *Escherichia coli* by the *arcA* and *fnr* gene products. *J Bacteriol* 1997;179:4299–4304.

99. Lynch AS, Lin ECC. Regulation of aerobic and anaerobic metabolism by the Arc system. In: Lynch AS, Lin ECC, editors. *Regulation of gene expression in Escherichia coli*. Austin, TX: Landes Co., 1996; pp. 361–373.

100. Cotter PA, Chepuri V, Gennis RB, Gunsalus RP. Cytochrome *o* (cyoABCDE) and *d* (cydAB) oxidase gene expression in *Escherichia coli* is regulated by oxygen, pH, and the fnr gene product. *J Bacteriol* 1990;172:6333–6338.

101. Cotter PA, Gunsalus RP. Contribution of the fnr and arcA gene products in coordinate regulation of cytochrome *o* and *d* oxidase (cyoABCDE and cydAB) genes in *Escherichia coli*. *FEMS Microbiol Lett* 1992;70:31–36.

102. Cotter PA, Melville SB, Albrecht JA, Gunsalus RP. Aerobic regulation of cytochrome d oxidase (cydAB) operon expression in *Escherichia coli*: roles of Fnr and ArcA in repression and activation. *Mol Microbiol* 1997;25:605–615.

103. Kuritzkes DR, Zhang XY, Lin EC. Use of phi(glp-lac) in studies of respiratory regulation of the *Escherichia coli* anaerobic sn-glycerol-3-phosphate dehydrogenase genes (glpAB). *J Bacteriol* 1984;157:591–598.

104. Quail MA, Guest JR. Purification, characterization and mode of action of PdhR, the transcriptional repressor of the pdhR-aceEF-lpd operon of *Escherichia coli*. *Mol Microbiol* 1995;15:519–529.

105. Zheng M, Wang X, Doan B, Lewis KA, Schneider TD, Storz G. Computation-directed identification of OxyR DNA binding sites in *Escherichia coli*. *J Bacteriol* 2001;183: 4571–4579.

106. Zheng M, Wang X, Templeton LJ, Smulski DR, LaRossa RA, Storz G. DNA microarray-mediated transcriptional profiling of the *Escherichia coli* response to hydrogen peroxide. *J Bacteriol* 2001;183:4562–4570.

107. Nystrom T, Larsson C, Gustafsson L. Bacterial defense against aging: role of the *Escherichia coli* ArcA regulator in gene expression, readjusted energy flux and survival during stasis. *EMBO J* 1996;15:3219–3228.

108. Seyler RW Jr, Olson JW, Maier RJ. Superoxide dismutase-deficient mutants of *Helicobacter pylori* are hypersensitive to oxidative stress and defective in host colonization. *Infect Immun* 2001;69:4034–4040.

109. Cooksley C, Jenks PJ, Green A, Cockayne A, Logan RP, Hardie KR. NapA protects *Helicobacter pylori* from oxidative stress damage, and its production is influenced by the ferric uptake regulator. *J Med Microbiol* 2003;52:461–469.

110. Guo JT, Ellrott K, Chung WJ, Xu D, Passovets S, Xu Y. PROSPECT-PSPP: an automatic computational pipeline for protein structure prediction. *Nucleic Acids Res* 2004;32: W522–W525.

111. Robinson VL, Wu T, Stock AM. Structural analysis of the domain interface in DrrB, a response regulator of the OmpR/PhoB subfamily. *J Bacteriol* 2003;185:4186–4194.

112. Thormann KM, Saville RM, Shukla S, Spormann AM. Induction of rapid detachment in *Shewanella oneidensis* MR-1 biofilms. *J Bacteriol* 2005;187:1014–1021.

113. Berg OG. Selection of DNA binding sites by regulatory proteins. Functional specificity and pseudosite competition. *J Biomol Struct Dyn* 1988;6:275–297.

114. Menon AL, Mortenson LE, Robson RL. Nucleotide sequences and genetic analysis of hydrogen oxidation (hox) genes in *Azotobacter vinelandii*. *J Bacteriol* 1992;174: 4549–4557.

115. Kortluke C, Friedrich B. Maturation of membrane-bound hydrogenase of *Alcaligenes eutrophus* H16. *J Bacteriol* 1992;174:6290–6293.

116. Rose L, Kaufmann SH, Daugelat S. Involvement of *Mycobacterium smegmatis* undecaprenyl phosphokinase in biofilm and smegma formation. *Microbes Infect* 2004;6: 965–971.

117. De Souza-Hart JA, Blackstock W, Di Modugno V, Holland IB, Kok M. Two-component systems in *Haemophilus influenzae*: a regulatory role for ArcA in serum resistance. *Infect Immun* 2003;71:163–172.

118. Sengupta N, Paul K, Chowdhury R. The global regulator ArcA modulates expression of virulence factors in *Vibrio cholerae*. *Infect Immun* 2003;71:5583–5589.

119. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 1994;22:4673–4680.

120. Gao H, Wang X, Yang ZK, Palzkill, Zhou J. Probing regulon of ArcA in Shewanella oneidensis MR-1 by integrated genomic analyses. *BMC Genomics* 2008;9:42.