Toward genomic cell culture engineering

Katie F. Wlaschin · Gargi Seth · Wei-Shou Hu

Received: 7 March 2006 / Accepted: 24 March 2006 / Published online: 25 July 2006
© Springer Science+Business Media B.V. 2006

Abstract Genomic and proteomic based global gene expression profiling has altered the landscape of biological research in the past few years. Its potential impact on cell culture bioprocessing has only begun to emanate, partly due to the lack of genomic sequence information for the most widely used industrial cells, Chinese hamster ovary (CHO) cells. Transcriptome and proteome profiling work for species lacking extensive genomic resources must rely on information for other related species or on data obtained from expressed sequence tag (EST) sequencing projects, for which burgeoning efforts have only recently begun. This article discusses the aspects of EST sequencing in those industrially important, genomic resources-poor cell lines, articulates some of the unique features in employing microarray in the study of cultured cells, and highlights the infrastructural needs in establishing a platform for genomics based cell culture research. Recent experience has revealed that generally, most changes in culture conditions only elicit a moderate level of alteration in gene expression. Nevertheless, by broadening the conventional scope of microarray analysis to consider estimated levels of transcript abundance,

much physiological insight can be gained. Examples of the application of microarray in cell culture are discussed, and the utility of pattern identification and process diagnosis are highlighted. As genomic resources continue to expand, the power of genomic tools in cell culture processing research will be amply evident. The key to harnessing the immense benefit of these genomic resources resides in the development of physiological understanding from their application.

Keywords cDNA library · Data analysis · Mammalian cell culture · Microarray · Proteome · Transcriptome

Introduction

The completion of the human genome project in 2001 is, arguably, among the most important contributions to the biological and biomedical sciences to date (International Human Genome Sequencing Consortium 2001). The wealth of information and the development of technology that surround the human genome project propelled sequencing efforts on other important mammals including the mouse in 2002 (Mouse Genome Sequencing Consortium 2002), the rat in 2004 (Rat Genome Sequencing Project Consortium 2004) and, most recently, the dog in 2005 (Lindblad-Toh et al. 2005). With the rapid

K. F. Wlaschin · G. Seth · W.-S. Hu (✉)
Department of Chemical Engineering and Materials
Science, University of Minnesota, 421 Washington
Avenue SE, Minneapolis, MN 55455-0132, USA
e-mail: acre@cems.umn.edu

development of genomic technology and the availability of a vast amount of data in genome, transcriptome, and proteome datasets, a migration of researchers from the benchtop to the laptop is underway. Mining sequence information within and among the available genomic databases has provided insight into the relationships among species, the mechanisms and treatment of disease, and the roles biological molecules play in the creation and maintenance of life. Such a shift is pervading in all areas of biology and biotechnology, and is beginning to impact cell culture bioengineering.

Among all of the current technologies for performing global surveys of cellular physiology, DNA microarray stands out as the most comprehensive, allowing for the dynamics of a complete genome to be observed. It is a powerful means for the direct comparison of transcript levels to gain insight into the regulatory mechanisms that may occur in response to environmental factors and genetic manipulations. In conjunction with the continuing decrease in the cost of DNA sequencing and related technologies, microarray based transcriptome analysis is inducing a sea change in biomedical research. Mining microarray data from patients has revealed patterns in gene expression, epigenetic changes, or single nucleotide polymorphisms (SNPs) associated with disease, especially cancers (Guo 2003). These findings are directing the search for possible treatments. Along with discovering potential therapeutic targets, these

genomic based tools are providing opportunities to optimize and enhance bioprocessing in recombinant protein production.

Mammalian cell lines are the major workhorses for the production of biologics, especially recombinant therapeutic proteins. Although there are a large number of established cell lines from a wide range of species, only a few of them are used in producing therapeutics. For the production of human therapeutics and viral vaccines, the cells used are largely derived from human (e.g. WI38, MRC5, HK293), dog (e.g. MDCK), African green monkey (Vero), or rodents, including Chinese hamster (CHO), Syrian hamster (BHK), and mouse (NS0, Sp2/0, mouse L). For veterinary products, a wider range of species including bovine and swine are also used.

The availability of resources for genome-based bioprocess research varies widely with species. The cell lines used in bioproduction can thus be divided into two categories: either rich or lacking in the availability of genomic resources. The distinction between these two categories extends beyond simply the existence of a completed genome sequence. A wide variety of other resources are also available for many diverse species including datasets of functionally annotated genes, ESTs, differentially spliced transcripts, SNPs, two-dimensional protein gel maps, and archived DNA microarray data. A summary of available information present in the GenBank database clearly shows the disparity in available information for various cell lines (Table 1). In

Table 1 Comparison of available genomic sequence information resources for frequently used cell lines

Organism	Cell Lines	Genome sequencing status	Public data (Number of entries in GenBank)	
			Nucleotide	Protein
<i>Homo sapiens</i> (Human)	HepG2, HeLa, WI38, MRC5, HK293	Completed	10,957,434	284,936
<i>Mus musculus</i> (Mouse)	NS0/Sp2/0, 3T3, mouse hybridoma, Mouse L	Completed	8,125,107	194,322
<i>Canus familiaris</i> (Dog)	MDCK	Completed	2,582,990	37,474
<i>Cercopithecus aethiops</i> (African green monkey)	Vero, COS	In Process	13,498	471
<i>Cricetulus griseus</i> (Chinese hamster)	CHO	Some BAC clone sequence available	6,563	775
<i>Mesocricetus auratus</i> (Syrian hamster)	BHK	Some BAC clone sequence available	2,353	874

addition to entire genome resources, gene rich organisms have a large set of ESTs, providing information on tissue specific expression, alternative splicing, and regulation. Clearly, processes or projects employing cells derived from human, mouse, dog, or monkey benefit from the widely available tools for transcript and protein analysis. Conversely, for other equally important bioprocessing cell lines, notably CHO and BHK, there is very little genomic or proteomic information. In these cell lines, EST sequencing and reliance upon DNA and protein sequence homology with other information-rich species, provide an alternative and viable research approach.

The approaches for exploring the transcriptome and proteome in information rich and information poor cell lines are rather divergent, warranting separate discussion of appropriate resources and techniques for addressing biological questions. Nonetheless, the data generated from both gene rich and gene poor cell lines share many characteristics, as most biological processes are generally conserved, even among very different species. There are also many shared approaches and applications for genomic and proteomic exploration of cell lines for process diagnosis, genetic engineering, and gene discovery.

Genomic resource-rich species

Transcriptome profiling

For mammalian cell lines with comprehensive genomic resources available, global exploration of gene expression is commonly accomplished using DNA microarrays for transcriptome profiling. Proteomic analysis is certainly another powerful investigative tool; however, its applications still trail behind microarray, as will be discussed later. Whole genome microarrays are available for mouse and human cell lines, with accessibility and affordability of these technologies continually improving. Many universities, research institutions, and industrial bioprocessing companies have access to transcriptome analysis tools. It is well accepted that the potential applications of DNA microarray are numerous; however, despite the seemingly widespread effort and excitement sur-

rounding microarray technology, there remain few published examples of direct applications in cell culture bioprocessing (Korke et al. 2004; Seth et al. 2005; Shen and Sharfstein 2006; Wong et al. 2006). It is likely that there will soon be flux of publications in this field as the physiological significance of the identified changes are discerned.

Two prevailing forms of microarray are currently utilized: DNA microarray and oligonucleotide-based microarray. DNA microarrays offer the advantage of being flexible in their construction, and fill the niche of an economical, in-house, custom-designed platform. This format allows for construction of smaller sub-arrays containing only desired sequences, reducing the sample size and easing data processing. The probes spotted on a cDNA microarray, (commonly used for complex mammalian genomes), are amplified from cloned cDNAs using universal vector primers. This differs from the construction of DNA microarrays used to study microbial species, for which primers can be designed to specifically amplify gene fragments from genomic DNA to serve as probes. Oligonucleotide-based microarrays, in contrast, utilize much shorter (20–80 bp), chemically synthesized probes. They are primarily available as commercial products, making them less flexible and more costly.

Although DNA microarrays are usually a more flexible and economical option, there are a few noteworthy limitations in using this technology. With the long length of DNA present on the array for each gene, a cDNA probe is not always specific for its intended target gene. Gene expression in mammals is highly complex, with alternatively spliced transcripts and isozymes that may cross-hybridize with probes for other related enzymes. Another limitation in the use of cDNA microarrays is the number of samples that can be compared with a reasonably small number of hybridizations. Each two-dye cDNA microarray acquires an expression ratio between two mRNA samples. If one wants to compare more than two experimental conditions, a common reference mRNA is needed to hybridize with each condition. If the number of samples to be compared is large, obtaining the required amount of reference sample can be problematic. Additionally, a reproducible reference pool is necessary to facilitate the

comparison of microarray data obtained over a long period of research.

The possibility of using fluorescent-labeled genomic DNA as a reference is becoming a more common practice, specifically for microbial DNA arrays (Kim et al. 2002; Talaat et al. 2002). The use of genomic DNA as a reference poses a challenge, because it is double stranded, as opposed to a single stranded labeled cDNA. Labeled genomic DNA fragments may hybridize to the sample cDNAs in the solution phase in addition to the intended probes on the surface. Using a mathematical model to describe competitive annealing in the solution, this effect was shown to be minor under most hybridization conditions (Gadgil et al. 2005). Genomic DNA has also been used as a universal reference in mammalian microarrays (Williams et al. 2004). The practice of using genomic DNA as a reference for mammals is fundamentally distinct from its practice in bacteria. Transcribed sequences constitute almost the entirety of a bacterial genome, whereas in mammalian cells, it is estimated that only 5% of the genome is transcribed (International Human Genome Sequencing Consortium 2001). The published results appear to demonstrate that under proper hybridization conditions, the presence of large amounts of untranscribed genomic DNA does not pose a concern. Nevertheless, caution should be taken to carefully optimize hybridization conditions if genomic DNA is used as a reference for the study of organisms with complex genomes.

cDNA microarrays and hybridization with labeled genomic DNA also offers a format for comparative genomic hybridization studies. Under optimized DNA fragmentation and hybridization conditions, it is possible to compare genomic DNA from two distinct cell lines to discern differences at the genomic level. In our laboratory, the difference in copy number of the *dhfr* gene was confidently measured by comparative hybridization of CHO cells and normal diploid hamster DNA. Since most cell lines employed for industrial use are aneuploid, such analysis among different clones may be revealing.

The other primary microarray platforms are oligonucleotide based. Since each probe is comprised of a smaller fraction of the transcript, a

higher degree of specificity for the intended target can be achieved, allowing for isozymes and alternatively spliced transcripts to be measured independently. The two most widely used types of oligonucleotide array are long oligo-DNA arrays and Affymetrix GeneChip[®] arrays. Long oligo microarrays are typically comprised of 50–70 bp probes spotted onto a glass slide. Their use is frequently similar to that of traditional two-dye cDNA microarrays for comparison of two samples. Affymetrix arrays are made by the direct synthesis of eleven sets of 25-mer probes onto the chip through photolithography-based technology (Chee et al. 1996; Lockhart et al. 1996). The 25-mers are designed to span the 600 base pair region nearest to the 3' end of the sequence. Each GeneChip[®] is used to hybridize a single sample. Multiple samples can be compared by taking ratios of their intensities for each set of data. Affymetrix has entire genome arrays available for human, mouse and rat. For some species, DNA tiling arrays and exon arrays will be available. DNA tiling arrays contain probes spanning the entire genome, including non-coding sequences. Exon arrays include probe sets designed for each known and predicted exon in a genome, allowing differential regulation by splicing to be examined.

Microarrays are most frequently used as an exploratory tool. After hybridization, the fluorescence intensities are processed to normalize data for comparison amongst arrays. A statistical analysis of replicate data is used to identify genes with varied transcript abundance. In cell culture processing (as well as in other research areas), there are rarely sufficient numbers of replicate samples to achieve a high degree of confidence in differential gene expression, especially when the degree of differential expression is small. Confirmation of microarray observations using a more sensitive method is often a necessity. Confirming microarray results using quantitative real time PCR (Q-RT-PCR), is relatively straightforward for organisms with a complete genome sequence. Several primer design programs can use the genomic sequence to test for specificity of primers over a range of PCR conditions in silico. Additionally, databases of known working primer sets are available (Pattyn et al. 2003). Functional confirmation of interesting genes is also more

rapidly achievable for resource rich species, with full-length cDNA clones available to test over-expression of a specific gene, and pre-designed siRNAs for gene suppression.

Once differentially expressed transcripts are identified in microarray experiments, they can be mapped to their genomic loci to reveal the regulatory mechanisms behind observed changes. In addition to identifying promoter and enhancer regions, the surrounding genome can be scanned to determine if epigenetic mechanisms may be involved. In mammalian cell culture, many desirable growth characteristics for bioprocessing are acquired by adaptation of a cell population. This is accomplished by gradually changing cultivation conditions over the course of a few weeks to several months. These new growth characteristics are certainly a reflection of changes in the cell's gene expression. Since these new growth characteristics are acquired in a relatively short time period, the probability is low that a mutation event will result in a growth advantage allowing a single cell to overtake an entire population. These adaptation events more likely occur through epigenetic modifications that can activate or completely silence expression of a gene in a cell population. One of the most studied mechanisms of gene silencing and activation occurs by regulating the state of DNA methylation of cytosine residues in CpG islands upstream of expressed genes. When the entire genome is available, DNA sequence surrounding a gene can be scanned for CpG islands to determine if DNA methylation might play a role in cell line adaptation (as shown in Seth et al. 2006).

Proteome profiling

Proteome profiling is another means of surveying global changes in gene expression. It is a powerful complement to microarray transcriptome profiling, serving to examine gene function directly at the protein level. In the post-genomic era, proteomic analysis allows for investigation of primary sequence, protein–protein interactions, and post-translational modifications on a large scale. Highly sensitive analytical techniques are required to study low abundance proteins and examine complex proteomes, and in this area, mass spectrometry has established itself as an

indispensable tool. A few groups have utilized proteomics to address topics relevant to bioprocessing (Alete et al. 2005; Hayduk et al. 2004; Hayduk and Lee 2005; Seth et al. 2005; Smales et al. 2004). Specifically, these have included studies in CHO cell lines (Chen et al. 2004; Kaufmann et al. 1999; Lee et al. 2003; Van Dyk et al. 2003).

2D-gel electrophoresis has been, and will likely remain a workhorse in proteome profiling for the next few years. It allows for simultaneous analysis of a large number of protein molecules. In this method, a complex protein sample is resolved in two dimensions according to charge (isoelectric point) and mass on a polyacrylamide gel, followed by staining. The stained protein spots are then observed and compared using sophisticated image analysis tools (such as PDQuest). The difference in staining intensity of observed spots gives relative quantification between protein samples. Most highly abundant proteins can be easily detected by this method, however, those present at lower levels are not easily seen in the gel image (Gygi et al. 2000). Some proteins co-migrate and cannot be resolved. Proteins with charges outside of the isoelectric range and highly hydrophobic membrane proteins are also unable to be resolved in the gel. Since it is estimated that there are likely more than 90,000 protein species expressed by cells (Harrison et al. 2002), this method only captures a fraction of those, and is, thus, limited in its ability as a true means of global protein profiling. There have been efforts to better resolve highly complex samples via sample fractionation before 2D gel analysis. More sensitive staining methods and large-format gels have also been introduced to increase the dynamic range of 2D-gel resolution and quantification (Rabilloud 2002; Unlu et al. 1997).

Once a protein spot is documented as differentially present in 2D gels of different samples, the spot is excised and purified for further analysis, such as direct sequencing or, more commonly, mass spectrometry for subsequent protein identification. Electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI) are two techniques commonly used to volatilize and ionize the proteins or peptides for mass spectrometric analysis.

MALDI-TOF (Matrix-assisted-laser-desorption/ionization-time of flight) mass spectrometry is used to match the resulting masses of a purified, digested sample with in silico digests of protein sequences; thus, this methodology only identifies proteins with a known sequence. This method of protein identification is also known as peptide-mapping, peptide-mass mapping or peptide-mass-fingerprinting. With its high mass accuracy, resolution, and sensitivity, MALDI-TOF is used to identify more pure protein samples in conjunction with 2D gel analysis. Alternatively, liquid chromatography ESI-MS (LC-MS) is used to identify several proteins from a less pure or complex sample (Hunt et al. 1992) by similarly matching spectra to known sequences.

The mass analyzer generates ion mass spectra from peptide fragments. There are four basic types of mass analyzers currently used in proteomics research. These are the ion trap, time-of-flight (TOF), quadrupole, and Fourier transform ion cyclotron (FT-MS) analyzers. Strengths of each have been reviewed (Aebersold and Mann 2003). To enhance the resolution of complex protein samples, various peptide and protein separation schemes have been devised, including the two-dimensional (strong cation exchange/reversed phase) or three-dimensional (strong cation exchange/avidin/reversed phase) chromatographic separations of peptide mixtures generated by tryptic digestion of protein samples (Aebersold and Mann 2003).

While LC-LC/MS methodologies have the advantage of including more proteins in the sample for analysis, there was no means of quantification of different expression level among samples, until recently. Stable-isotope labeling has been applied to LC-LC/MS to facilitate quantification between different samples. This method makes use of the fact that pairs of chemically identical analytes with different isotope compositions can be differentiated in a mass spectrometer by their mass difference. The ratio of signal intensities for such an analyte pair accurately indicates an abundance ratio for the two analytes. Techniques including SILAC (Stable Isotope Labeling by Amino Acid in Cell Culture) (Ong et al. 2002), GIST (Global Internal Standard Technology) (Chakraborty and Regnier

2002), I-CAT (Isotope-Coded Affinity Tags) (Gygi et al. 1999), and iTRAQ (Isobaric Tagging for Relative and Absolute Protein Quantitation) (Ross et al. 2004; Schneider and Hall 2005) have been developed to quantify protein expression differences using LC-LC/MS techniques.

Chemical isotope tagging of proteins in vitro (after protein isolation) using approaches such as iTRAQ, GIST and ICAT, is the most versatile and widely used labeling method. ICAT labels cysteine residues, whereas methods like iTRAQ, GIST and O-18 approaches (Rose et al. 1983), are designed to tag all peptides in a digest. ICAT tags consist of a protein-reactive group (labels cysteine residues), followed by an isotope group and an affinity group. iTRAQ tags are similar, having a reporter group, a balancer group and a peptide reactive group that tag the N-terminus of every peptide. One of the unique features of iTRAQ is that up to four samples can be labeled with different tags and analyzed simultaneously (other methods only label two samples). The combined balancer and reporter groups have an identical mass of 145 for all four labeling pairs of balancers (mass 28–31) and reporters (mass 114–117). The MS spectra for each of the four, labeled samples, looks identical. The reporter-balancer fragment stays intact, giving rise to the same m/z ratio. This allows for protein identification using the combined signal of the four samples. Upon MS/MS fragmentation, the bond between the reporter and balancer group is broken. The reporter groups appear as peaks in the low mass region, and quantification of the peak area for each different reporter group gives relative abundance for a peptide between the labeled samples.

In vivo labeling methods, such as SILAC, use deuterated leucine to differentially label one of the compared protein samples by replacing leucine in the cell culture medium, allowing the isotope to incorporate into the cellular proteins. The combined labeled and unlabeled samples are analyzed by LC-LC/MS, generating two spectra for each peptide fragment, one shifted precisely by the mass of the deuterated leucine. Differences in peak height, similarly, provide the means of quantification between samples. The method allows two samples to be mixed prior to protein isolation, eliminating systematic errors due to

protein isolation efficiencies. It is very well suited when subcellular enrichment protocols are used (as for study of organelle fractions) prior to LC-LC/MS analysis. One notable limitation of this method is that the labeling time is relatively long, to allow for sufficient incorporation of the isotope into cellular proteins.

Even though all proteins in a sample are, technically, included in the analysis using non-gel based techniques, these methods are still not a true global surveying tool. Typically, only a few hundred to small number of thousands of proteins are identified in analysis of a sample due to limitations in the resolving power of liquid chromatography. The capture of a peak from a sample is probabilistic, and, it is likely that highly abundant proteins are detected over those present at lower levels. The isolation and identification of low abundance proteins can be enhanced by repetitively analyzing the same fractions in the mass spectrometer, excluding peptides that have already been identified in the previous analysis. This is done at the high expense of run time; thus, exhaustive surveys of the proteome space are not commonly practiced.

DNA microarrays come closer to possessing the power of “genome-wide” expression profiling. Indeed, the entire transcribed fraction of the genome can be probed in any readily equipped laboratory. In contrast, proteomic surveys are only performed in a small number of specialized laboratories, and the number of proteins that can be identified pale in comparison to the number of gene probes present on any microarray. Most proteomic efforts reported for cell culture bioprocessing lay out highly useful and representative datasets for the specific system under investigation, rather than provide a comprehensive exploration. As the cost of run time reduces, more exhaustive LC-LC/MS datasets will be generated. Also, the current methods only allow for a limited number (up to 4) of samples to be directly compared. This technology will likely be extended in the near future. With run time as the limiting factor, the ability to combine more samples in each run will also help create a more affordable and truly “-omic” tool.

The coordinated use of transcriptome profiling and proteome surveys will be very revealing in

terms of understanding cellular physiology. The relationship between abundance level of transcripts and their corresponding protein expression level is not always linear, and it is likely that protein level is more relevant to cell behavior than mRNA abundance. Furthermore, many proteins are localized or otherwise distributed in different cellular compartments. Their localization can also be revealed by proteomic analysis after cellular fractionation; an aspect that cannot be done in transcript analysis. In a recent comparative transcriptome and proteome study of antibody producing recombinant NS0 cells, the extent of the differences that iTRAQ and microarray have each revealed are substantial (manuscript in preparation Seth, et al. 2006). By examining the cellular samples in different dimensions using complimentary proteomic and transcriptomic analysis tools, one can truly begin to enhance the understanding of cell physiology.

Genome resource-poor species

Chinese hamster ovary (CHO) cells are one of the most important cell lines in biological research and are, arguably, the most popular host for recombinant protein production. They are also among the most widely used cell lines in bioscience research. Despite their importance, only a few researchers and industrial laboratories are known to have EST sequencing efforts. Until late 2005, less than 1000 sequences were publicly available in the Genbank database for Chinese hamster. Another important, related organism is the Syrian hamster, which is a model animal for biomedical research, and the source of another industrial cell line, BHK cells. Information for this species is currently even more limited than for CHO cells in terms of available sequences in the GenBank database.

In the absence of complete genomic data, exploration of the transcriptome and proteome in these cell lines is heavily dependent upon the conserved nature of DNA and protein sequences. Mouse and rat are currently the most closely related species with complete genome information, and they typically provide the basis for molecular studies in both CHO and BHK cell

lines. Sequence identity between mouse or rat genes with CHO or BHK cell line genes is in the range of 75–97% base pair identity. The level of homology varies with each gene, making results highly variable.

One approach for transcriptome profiling in CHO and BHK cells has been to use DNA microarrays developed for mouse or rat. Similar cross-species hybridization studies have been employed in transcriptome analysis of pigs (Gladney et al. 2004; Moody et al. 2002) and primates using human microarrays. A high level of success was achieved in the primate microarray studies, which can likely be attributed to the very high degree of homology between human and primate DNA (more than 98% between humans and chimpanzees) (Gu and Gu 2003). In our laboratory, we have compared the transcript profiles of CHO cells treated with butyrate using both a mouse cDNA microarray (containing ~ 6,600 genes) and a CHO cDNA microarray (containing ~ 2,600 genes). For the CHO cells hybridized to the mouse microarray, only 4.5% of the genes were identified as differentially expressed. Using the smaller set of genes on the CHO microarray, 8.1% of the genes significantly changed using the same cut-off criteria. It was also observed that the *P*-values obtained in the cross-species (CHO: mouse) hybridization were consistently higher than the *P*-values for the orthologous sequence on the CHO microarray. This implies that the level of identity between CHO and mouse sequences affects the number of probes that satisfy the criteria for confidently identifying differentially expressed genes. This was even more evident when the time course transcript profiles obtained from both CHO:CHO and CHO:mouse hybridizations were plotted and compared side by side. A general consistent trend between the two sets of data (CHO:CHO and CHO:mouse) can be seen for many genes; however, the trend is invariably less pronounced in the cross-species hybridization. Confirmation of differential expression using methods like Q-RT-PCR is also challenging. In many cases, more than one round of primer design and optimization is required when relying upon orthologous sequences.

Other options for transcript profiling in the absence of sequence information include differential display and suppression subtractive hybridization. These techniques are designed to enrich the representation of rare transcripts and highlight those that are differentially expressed between two samples. Results are inherently dependent upon the experimental conditions, which can be varied to generate different results (Gadgil et al. 2002). Additionally, these methods are much more labor intensive than microarray based techniques.

With the cost of DNA sequencing reducing over the past few years, it is becoming more affordable to generate large batteries of sequence information by large-scale EST sequencing for these cell lines. Indeed, our laboratory in collaboration with the Bioprocessing Technology Institute of Singapore—A*STAR and a number of industrial sponsors, has embarked on such an effort for CHO cells. At the current rates, the cost of sequencing an entire mammalian genome is on the order of \$10 million U.S. dollars. With such a high price tag, a relatively small number of organisms have been selected for publicly supported genome sequencing efforts. Although Chinese and Syrian hamsters are, arguably, medically and industrially important, they are not among those currently funded. It is likely that the genome for the Syrian and Chinese hamster will be realized with the development of new sequencing technology. In 2004, the NIH initiated a large funding source titled “Revolutionary Genome Sequencing Technologies—the \$1000 genome” to reduce the cost of genome sequencing within the next 10 years (RFA-HG-05-004). With great progress being made in these research areas, it is only a matter of time before there is a concerted effort to sequence these species. In the meantime, EST sequencing provides a less expensive route for acquiring the sequence of transcribed genes and meeting the current need to establish tools for large-scale transcriptome analysis.

In general, an EST sequencing project for CHO and BHK cell lines is very similar to those conducted for other organisms; however, it should be noted that a project based on a cell line, as opposed to an entire organism, has a

number of unique hurdles. The first step in any EST sequencing project is the construction of relevant cDNA libraries. Obviously, the use of cell lines cultivated under relevant conditions in bioprocessing would be most desirable for library construction, ensuring that the transcripts are representative of conditions of interest. The choice of cell lines and conditions for isolating mRNA is not trivial. With relatively little transcript profiling data reported for production cell lines, the degree of variation in transcript representation between different cell lines is not known. Lacking this insight, it might be important to choose many cell clones exhibiting highly desirable physiological traits. Likewise, a parental cell line may also provide insight into the changes a cell undergoes to gain superior characteristics.

Neglecting cloning bias, high throughput EST sequencing can be viewed as a statistically random event. A cDNA library captures the population of mRNA species in the tissue or cell line from which it was derived. Therefore, transcripts that were present in a high copy number inside the cell line or tissue will be isolated and sequenced more frequently than mRNA species that were present as only one or two copies per cell. This redundant isolation of highly expressed species results in a decrease in the new gene discovery rate as more and more ESTs are sequenced. In most EST projects, it is common practice to construct and sequence new libraries when the gene discovery rate falls too low, and sequencing becomes economically unrewarding. When ESTs are derived from a complex organism, various tissues, physiological conditions, or diseased conditions can be used to construct a variety of libraries with very different transcript representation. For CHO and BHK cells, one resorts to using different clones, treatments, or environmental manipulation for possible alteration of transcript representation. Another approach is to construct libraries from Chinese or Syrian hamster tissues to increase the gene discovery rate, specifically for genes expressed at lower levels in the cultured cell line. It should be noted, however, that sequences isolated from tissues are not necessarily actively expressed in cultured cells. Nevertheless, these transcripts are

likely present in the cell line genome and could be actively expressed under some culture conditions.

As with every EST or genome sequencing effort, a major challenge is annotating sequenced transcripts. In a relatively small-scale CHO EST sequencing project, among nearly 7,000 unique CHO cell sequences, 75% show some level of homology with an EST sequence in the GenBank database (e-value $<10^{-10}$). Among those, only 72% can confidently be assigned a meaningful function with confidence (54% of all sequenced ESTs) (Wlaschin et al. 2005). This leaves a large number of poorly annotated sequences. For these sequences, further work to identify protein motifs within the ESTs can be done using Hidden Markov models (HMMs) to assign the ESTs as possessing a certain functional group (Durbin et al., 1998). Additionally, as more sequences become available with continued EST sequencing and work in other species, periodic BLAST comparison with multiple databases can yield more information in the future. With this in mind, an EST project for CHO or BHK cells can be implemented to accomplish the goal of constructing a DNA microarray in a relatively short time frame (less than six months); however, a sustained annotation effort will be required to provide physiological insight from the microarray data.

It goes without saying that EST sequencing in CHO and BHK cells will also aid in the application of proteomic tools for studying these two cell lines. As amino acid sequences are generally more conserved than DNA, a significant amount of work can be done using homology with mouse, human, and rat peptide sequences. Still, without complete species-specific information some proteins will not be identified. This is especially true for high-throughput mass spectrometry methodologies (e.g. SILAC, ICAT, and iTRAQ) where small peptide fragments are used to identify a mixture of proteins. EST sequencing can aid in identification of a few proteins, although a sufficiently large pool of EST data is likely required to significantly improve results. Most EST sequences only cover 600–700 bp of a transcript, which falls short of covering the full open reading frame. This is especially prevalent when 3' end sequencing is used because the reaction must span

the 3' UTR before reaching the protein coding region. Even with the limited genome and EST information available for CHO and BHK, using proteomic techniques to study these cell lines can be fruitful. A dataset of approximately 1,200 CHO cell proteins have been identified using mass spectrometry at the Bioprocessing Technology Institute—A*STAR (R. Philp and M. Yap, personal communication 2006). Within this dataset, approximately 97% of the proteins could be identified using only sequences in the UniProt database. Only 2.5–3% of the identified proteins could not have been found without including additional CHO EST sequences. It is not known how many more high-quality spectra could be identified with the availability of more complete CHO sequence information; yet, it is clear that the scarcity of information should not deter current efforts in proteomic exploration in CHO and BHK cell lines.

Data pipeline and connection to physiology

As the use of microarrays becomes more frequent or even routine, an infrastructure for storing, processing, and analyzing data becomes critical. A common pipeline can be developed to handle both cDNA and Affymetrix microarray data for a variety of species. A more advanced database is able to incorporate proteomic data from 2D gel electrophoresis and 2D LC mass spectrometry in addition to microarray data management. For those using a genome resource-poor species (e.g. CHO or BHK) with a need to develop EST based microarray, the infrastructure will need to include an EST management database. The incorporation of an EST project requires a few extra elements in a pipeline to handle clones, chromatograms, sequence processing, sequence assembly, and annotation. This is best accomplished using databases and modules that share common identifiers with the main microarray data pipeline, but are maintained externally. EST datasets will require more frequent updating and flexibility as information is expanded, re-analyzed and re-annotated. DNA sequences, BLAST search results, and contig assembly information are basic elements in most EST databases, but most also

include ontological information and other identifiers that allow commercial analysis software to be used within the primary data processing pipeline. A schematic representation a functional infrastructure that incorporates EST sequencing, microarray construction, data analysis, and data archiving is shown in Fig. 1.

For all microarray experiments, the first element in a processing pipeline is likely raw data treatment and storage of experimental parameters, conditions and results. After obtaining the microarray images, intensity values are extracted for each spot or gene on the array. Following the extraction of raw intensity values, the data is normalized to compensate for systematic biases and to allow for comparison of data obtained in different microarray experiments. The images, raw intensity data and normalized data are typically archived for future data mining. In addition to data archiving, it is equally important to link appropriate experiment and sample information. This is critical to allow microarray data to be analyzed at a later date when annotations are further enhanced. It is generally expected that the greatest discoveries using microarray data will be realized in years to come, through mining the thousands of datasets generated today.

Once raw microarray data is extracted, normalized and archived, the next stage is implementation of statistical analysis to determine expression ratios and the level of confidence in identification of the differentially expressed transcripts. For static data, as is the case for direct comparison of samples, the statistics employed can be relatively straightforward. Ratios are directly calculated and a standard *t*-test can be used to obtain *P*-values across technical and biological replicates. For cultured cells, more complex analysis may be required as experiments are commonly multi-dimensional, with samples taken over a time course. For this analysis, experimental design usually involves the use of a common reference sample to which all time points are compared.

Proper comparison of time course data is reliant on comparison of samples from the equivalent growth stage of the culture. With the difficulty and seemingly inevitable variability in clones and inoculum preparation, kinetic variables like

Fig. 1 Data management infrastructure for EST sequencing and microarray-based transcriptome analysis

growth rates, nutrient consumption, and metabolite accumulation need to be examined to determine if re-alignment of datasets is required to compare appropriate conditions. Even replicate cultures can behave differently. One culture may develop a significant lag phase, or reach its stationary phase sooner than the other cultures. A properly identified “equivalent” stage or time point for compared cultures is important for the identification of genes exhibiting expression changes. Time alignment of growth and gene expression profiles can be even more important when comparing mutants or different culture conditions, where altered growth rates might be expected.

After alignment of the data, one needs to identify the genes that are changing during the course of a culture. Fold change data for each gene at each time point needs to be evaluated simultaneously to capture dynamics of changes in gene expression. This type of analysis can be accomplished by applying distance calculations, like Euclidean distance methods, to help identify genes with different gene expression profiles. As is discussed further in later sections, fold changes

in gene expression are often small in magnitude for mammalian cell culture processes, making it difficult to detect time profile changes using distance calculations. Additionally, few genes change significantly under most conditions in mammalian cell culture. To save the effort of implementing distance calculations for every gene, it may be sufficient to identify kinetically changing genes by looking at the degree of deviation from the initial time at all subsequent time points. For those genes exhibiting a dynamic response, threshold values for both the magnitude of change between the cultures, and the number of time points differing in the series can be used to find differentially regulated transcripts. As an example, a gene in two compared cultures may increase its expression over time to two-fold of the original time point. In both cultures, the gene expression is dynamically changing; however, there is no difference between the two datasets. Conversely, if gene expression increases to two-fold in one culture and four-fold in the other, the gene should be identified as kinetically differentially expressed with a two-fold difference between the two cultures.

Up until this point, data has been processed in numerical form, and genes are identified as differentially expressed by applying cutoff values to reduce the dataset to a much smaller, relevant set of genes. With the large amount of data, this work involves scanning and filtering large spreadsheets of numerical data. After filtering, one should carefully verify that the calculations, algorithms, and filters applied accurately capture the important features of the data. Our innate ability to comprehend large amounts of complex data is greatly facilitated by visualization. Regardless of whether the experimental design is static (comparison of multiple samples) or dynamic (time course comparisons of two or more cultures), microarray data are inherently multidimensional. Multidimensional visualization software, such as Spotfire[®], GeneData Expressionist[®], or GeneSpring[®], can greatly assist in the critical tasks of determining appropriate statistical parameters, diagnosis of bias, verification of results, and other troubleshooting issues. Depending upon the criteria and confidence values used in statistical analysis, a wide range of results can be derived from the same set of microarray data. Data interpretation is, thus, highly subjective, and it is useful to treat microarray data analysis as an exploratory adventure, with the aim of surveying multiple terrains to generate a hypothesis or to direct research down unexplored avenues. When initially diagnosing a dataset, it is advisable to vary statistical criterion and examine the differences in the resulting gene sets.

In many cases, expression changes with lower confidence levels are excluded by applying more stringent criteria in statistical analysis. It is well known that high statistical significance in gene expression correlation or differential expression is no guarantee of physiological relevance. Conversely, a low statistical significance does not necessarily warrant immediate exclusion from consideration. Inclusion or exclusion of some gene sets from the final list of changed genes may alter physiological interpretation of the data. It might be advisable to vary the statistical stringency in the course of data analysis if the inclusion of less confidently changing genes affects the resulting interpretation drastically. Their differ-

ential expression can be validated using other more sensitive methods like Q-RT PCR. A good visualization software with the ability to quickly select data by filtering, sorting, and plotting in multiple dimensions can help to scan a range of statistical criteria and determine the most appropriate means of analysis. Grouping data in different ways, such as by functional class, or by the profile of their response, allows for consistency checking among conclusions drawn using different criteria or variables.

Another type of analysis tool assists in finding relationships among differentially expressed genes by network visualization, gene function analysis, and metabolic pathway visualization. Ingenuity[®], PathwayAssist[®], and GenMAPP (Dahlquist et al. 2002) are three widely used and publicly or commercially available examples of such tools. Given a list of differentially expressed genes, any individual researcher will likely be knowledgeable of only a portion of them, making it very difficult to discern the relevant relationships among the genes. A network or pathway visualization tool can help by gathering and displaying information about each gene in a list by searching databases of well-curated physiological and molecular knowledge. For example, they can be used to group genes having similar functions, genes acting in a specific signaling cascade, or genes participating in the same metabolic or regulatory pathway. Results are displayed on a network or pathway map to highlight potential relationships. This analysis is particularly useful in identifying indirect relationships between genes by showing interactions with elements that may be missing from the list of genes. For instance, gene A, a cell cycle gene, and gene D, which is involved in apoptosis, are both differentially expressed, but are not related according to the current knowledge of the researcher. By using a program that identifies and displays relationships among genes, it could be discovered that gene A and D are related through their interactions with two other elements, B and C. In addition to stimulating the development of hypotheses, network visualization can also provide a terse means of storing interpreted results. The visual formats utilized clearly define the relationships among genes and

can be interpreted by other researchers at a later date.

Among the tasks involved in microarray analysis, the most difficult is usually making the connection between the gene expression and an observed behavior. Visualization and network analysis are an invaluable aid to this effort; however, in some cases, the level of understanding required for correct interpretation is beyond the scope of the software. Both Ingenuity[®] and PathwayAssist[®] provide a small number of links to relevant literature; however, it is not always complete. To aid in finding the most relevant literature for a long list of candidate genes, batch search engines and automated information processing algorithms are useful. Recently, a search engine was developed for our laboratory to perform a batch search of the Medline database for a list of genes. The literature found for each gene in the list is pooled and clustered according to the content of the abstracts, generating groups of related articles. The clusters and articles within the clusters are ranked by their relevance to the group of genes submitted to the search. User-defined weighting factors can also be employed to further customize the result display (e.g. year of publication, journal, etc.). With the large amount of unknown genes needing further study, more sophisticated literature search engines will be a great assistance.

Unique features and utility of gene expression profiling in cell culture processes

With DNA microarray becoming a relatively “mature technology” in microbial and biomedical communities, the amount of effort in applying microarrays to cell culture processes is greatly eased. For many researchers, the resources, equipment, and methodologies are readily available. Even with this strong foundation, there are some applications and considerations that are unique to cell culture studies, which should be considered in the development of an analysis pipeline. Some common applications, observations and challenges in using microarray technology for bioprocess development are addressed in the following section.

Characteristics of gene expression in cultured mammalian cells

One of the unique features of transcript analysis in cultured mammalian cells is the small number of genes differentially expressed, and the lower magnitude in the observed fold-changes as compared to those observed in other biological systems. For microbial cells, it is not unusual to observe more than 20% of the genes as differentially expressed by more than two-fold within 10 min of changing culture conditions. For cultured mammalian cells, it is rare to see more than two-fold expression changes occurring in more than a small percentage of the genes surveyed. Figure 2 shows the fluorescence intensity and fold-change data for cultured mouse hybridoma cells (Fig. 2a) treated with sodium butyrate and for mouse adult stem cells (MAPC) being induced for differentiation into a liver lineage (Fig. 2b) using the same Affymetrix mouse GeneChip[®]. For each dataset, average intensities are plotted along the y-axis, and the $\log_2$ of the expression ratios are plotted on the x-axis. Each marker represents a gene on the array. The vertical lines mark the bounds of a two-fold expression change. Markers lying outside of these lines are more than two-fold up or down regulated between the two samples compared. In comparing Fig. 2a and b, the number of genes that are more than two fold differentially expressed is substantially higher for the stem cell differentiation study than for the butyrate treated hybridoma cells. This trend has been consistently observed in the overwhelming majority of cultured cell microarray experiments. In fact, among cultured cell datasets, sodium butyrate treatment was observed to elicit a more dramatic response than many other treatments, including studies of stressed cells, temperature shift cultures, fedbatch cultures, and metabolic shift (unpublished results). It is possible that these observations are common with the terminally differentiated phenotype of the cell lines used in bioprocessing. A logical interpretation is that culture condition changes or genetic manipulations that might cause profound changes in physiological conditions for unicellular microorganisms are relatively minor for individual mammalian cells. Most

Fig. 2 Comparative transcriptome dynamics between (a) mouse hybridoma cells treated with butyrate and (b) differentiating mouse stem cells (MAPC). MAPC cells were grown on liver lineage differentiation medium for 6 days. Mouse hybridoma cells were treated with 1 mM butyrate for 27 h. Both MAPC and hybridoma cell samples were referenced to a corresponding, untreated time sample. The transcript levels were probed with the Affymetrix MOE430A array

environmental perturbations are dealt with by a coordinated response of tissues and organs in vivo, dampening the effects on an individual cell. An individual cell might not be programmed to respond to the condition change. Conversely, stem cells, when induced to differentiate by exposure to cytokines and growth factors, begin to undergo drastic changes in many ways. The natural course of stem cell differentiation elicits a transformation in cell morphology, biological function and gene expression. The cellular gene expression machinery in these stem cells is programmed to undergo these major changes

with the treatment utilized, while it may otherwise be unresponsive to different changes in culture conditions.

Use of absolute intensity data in transcriptome analysis and physiological interpretation

While the intended use of microarray is for comparison of transcript levels among different conditions, the data acquired can also be used to assess the abundance level for expressed genes. This information could be useful for the analysis of differential expression, particularly for cultured cells where few genes are highly differentially expressed. In microarray data analysis, it is customary that differentially expressed genes are identified, and that not much consideration is given to genes that are not differentially expressed. Furthermore, genes undergoing larger fold changes are generally considered more worthwhile for subsequent analysis. Genes with a similar fold change value may differ in their transcript levels. The transcript abundance level (or the intracellular concentration) of rare and abundant genes spans over three orders of magnitude. For an abundant gene expressed at a high level, a low level of differential expression represents a large net increase of transcript copies, which might also substantially increase protein levels. In the analysis of differentially expressed transcripts, a smaller change in differential expression for a highly expressed transcript may, thus, be very significant, and should not be casually excluded from analysis by applying a rigid fold-change cutoff.

Analysis of genes that are not differentially expressed can also greatly affect the physiological interpretation of microarray data. In mammalian cells, isozymes and alternative pathways are common. One could falsely conclude that a pathway has altered by observing that a single gene encoding for a key enzyme is differentially expressed. It is possible that an isozyme that is expressed at a much higher abundance level, is actually the dominant enzyme in this reaction, and has not changed its expression. Certainly this casts doubt on the original conclusion, and warrants further investigation. Thorough data analysis should, minimally, examine the

expression of genes with multiple isoforms, especially when the degree of differential expression is low.

With sufficiently large archived microarray data, one can establish a dynamic range of gene expression for different classes of genes or a given gene. Figure 3, shows the distribution of fluorescence intensities for different gene classes obtained with Affymetrix microarrays in NS0 cells. Many protein processing and cytoskeleton genes are among those with the most abundant transcript levels. Most metabolic enzymes in carbohydrate metabolism are expressed at moderately high levels. The varying range in expression between gene classes certainly provides useful

insight on expression levels and regulation. This type of information can guide work in cell engineering and in the development of mathematical descriptions of gene regulation.

For use of absolute intensity for the described applications, the extraction and calibration of intensity data is highly important. In the case of Affymetrix arrays, direct comparison is possible when all arrays are normalized to the same level. For spotted cDNA arrays, the extraction of representative intensity data is inherently more difficult. For semi-quantitative results, a common standard for all genes is needed to provide a baseline for the spotted probes. Genomic DNA is well suited for this purpose, as each gene is

Fig. 3 Representative fluorescence intensity distribution for important biological processes and gene functions in mouse NS0 cells. The transcript levels were probed with the Affymetrix MOE430A array

equally present in the sample. This is important because fluorescence intensities are dependent upon both the amount of labeled transcript in the sample, and the amount of probe DNA bound to the surface of the array. The concentration and quality of the spotted DNA probe (for printed cDNA arrays) varies for each gene, making the relationship of intensity to the true expression level nonlinear. Additionally, both data normalization and scanning affect intensity values. With the presence of these sources of variability, intensity data can, at best, provide an “order of magnitude” estimate of the actual gene expression level. With more work in analysis and optimization of conditions, such data has the potential to yield more precise estimates of relative transcript levels in the future.

Utility of clustering and time alignment algorithms for cultured cells

A well-known, visually appealing analysis tool for microarray is the heat map. Heat maps are typically generated using clustering algorithms to group transcripts with “similarities” in their dynamic behavior or in their patterns of gene expression. Again, for this application, the small number and low amplitude of changing genes is problematic. When the fold changes of transcript data fed into clustering algorithms are small, noisy, and statistically uncertain, clustering can lead to wrong conclusions. Results of clustering algorithms should be carefully reviewed for physiological relevance and blatant contradictions. For many of the same reasons, caution should be exercised in the use of time alignment algorithms. False results could lead to the comparison of the wrong stages of a growth profile.

Applications of genomic tools in cell culture processing

Global survey tools like DNA microarray and high-throughput proteomic analysis are versatile technologies with their primary strength in performing exploratory work. The results have aided in formulation of hypotheses and in directing future research and technology development. Microarrays also hold the promise of being

utilized as diagnostic tools, by detecting patterns that are associated with certain traits. Such applications are attractive because there is little need to develop a physiologically relevant interpretation of relationships between gene expression and the observed trait.

Direct application of microarray and proteomic technology in cell culture bioprocessing is rather recent and still relatively infrequent. In an early example, microarrays and 2D gel electrophoresis were used to find relationships between gene expression and the observed metabolic behavior where cultured mammalian cells change their metabolism (without affecting growth rate) to a state where glucose is consumed slowly and metabolized more aerobically, producing very little lactate (Seow et al., 2001). This phenomenon is referred to as metabolic shift. A similar alteration in metabolism can be observed in yeast, where cells change from an ethanol producing state to an ethanol consuming state (DeRisi et al. 1997). Microarray results from the yeast experiment showed a large number of genes in the central metabolic pathways dramatically changing in their expression profile. Changes in the same genes in the metabolically shifted hybridoma cell work were comparatively subtle (Korke et al., 2004). These data were an early and, at the time, surprising indication that even seemingly drastic physiological alterations can arise from very small transcriptome changes. The results of this analysis, while not directly providing a target for metabolic engineering, directed research toward engineering nutrient transport for controlling metabolic properties, as opposed to altering genes in central metabolism.

In another nutrition related study, microarrays were used to examine cholesterol auxotrophy in NS0 cells. Comparison of transcript profiles of cholesterol requiring NS0 cells with cells that have been adapted for growth in the absence of cholesterol led to the discovery of the mechanism of cholesterol auxotrophy in NS0 cells. The activation of a single gene, *Hsd17b7*, was found to alleviate dependency on exogenous cholesterol (Seth et al. 2005). In another example, an in-house developed CHO cell microarray was used to successfully identify anti-apoptotic genes that are expressed in the late stages of fedbatch

culture. Cloning of these newly discovered transcripts into CHO cells was shown to be an effective means for delaying the onset of stationary and death phase in CHO fedbatch cultures (manuscript accepted for publication, Wong, C.F.D, 2006).

These few examples illustrate the use of microarrays as global survey tools. The results helped to identify target genes for successful metabolic engineering of cells with enhanced properties. For the described examples, this success can, in part, be attributed to the fact that knowledge of the pathways involved is relatively comprehensive. For other traits that are not well studied, physiological interpretation is markedly more complex. Microarrays, 2D gel electrophoresis, and iTRAQ have recently been used to examine the more complex trait of high productivity in IgG producing NS0 cells. For this analysis, the relevant pathways are not well defined, and the changes incurred as a cell gains the qualities of an efficient protein producer likely involve complex interactions among several pathways. Furthermore, many combinations of traits could give rise to similar productivities in different clones. The complexity of the trait makes such analysis a daunting task. Deriving a conclusive causal relationship between expression patterns and the hyperproductive trait will be a Holy Grail in the years to come.

Unlike hypothesis driven research, for which variables are carefully designed and controlled, bioprocess variability is usually multidimensional and seemingly unpredictable. Rarely can a single set of variables be assigned as the controlling factor. When a sound physiological interpretation cannot be discerned, microarray can still be a powerful tool for analysis of complex traits. Identification of shared gene expression patterns can allow for process diagnosis and the identification of key variables. By comparing global gene expression for related conditions, one can easily assess the level of similarity and dissimilarity in the response of the compared samples. An example of such an application of microarray is illustrated in Fig. 4. The purpose of this cDNA microarray experiment was to examine the changes in gene expression that gave rise to different process outcomes for three cultures.

Fig. 4 Comparison of transcriptome profiles for diagnosis of expression similarity between cultures. The degree of similarity between cultures (a) S_2 and S_1 , or (b) S_3 and S_1 , can be seen by the degree of alignment of spots along the positive diagonal

Each of the three cultures (S_1 – S_3) was hybridized to a two-dye cDNA microarray using a common reference sample (S_{ref}). The $\log_2$ value of the transcript ratio from one of the three cultures is plotted versus the $\log_2$ of expression ratio for another of the cultures. The horizontal and vertical lines drawn represent a ± 1.4 fold change cutoff used to arbitrarily determine differential expression. Spots lying outside of these lines along either axis are differentially expressed in the respective culture. The spots enclosed within the center square are not differentially expressed in either culture. The markers lying in the regions labeled C and C' are

differentially expressed in the same direction (in C upregulated, in C' down-regulated) in both cultures. Conversely, the spots that fall into the regions labeled B and B', are differentially expressed in only the culture whose data is plotted on the y-axis, while those in the regions labeled E and E', are only differentially expressed in the culture plotted along the x-axis. The remaining markers in the areas labeled D and D' are genes that have changed in opposite directions in the two cultures compared. This type of visualization allows for a quick survey to identify both similarities and differences in the response of the cultures. The observed phenotypes can then be interpreted in terms of their similarities and differences. In this example, Fig. 4a shows the majority of spots fall along the diagonal, with a high level of agreement in gene expression for cultures S₁ and S₂. Cultures S₁ and S₃, compared in Fig. 4b, are less similar, having a wider spread in the data, and having more spots with inconsistent values (in regions D and D'). Further study of the consistent and inconsistent gene sets could provide clues that help explain the physiological cause of observed behaviors. Furthermore, desirable patterns in gene expression can be identified and used to diagnose culture data according to their correlation with "normal" and "abnormal" standards, using similar techniques.

Future outlook and concluding remarks

In the decade following the emergence of microarray technology and within the short 5 years since the completion of the first mammalian genome, many powerful and mature research platforms have been developed and made widely available. The direct application of these tools specifically for cell culture bioprocessing is still in an early stage. In some ways, this is related to the lack of the development of such technology for the for the most commonly used cell line (CHO). In this article, we review the application of these tools in the context of bioprocessing, with an emphasis on microarray analysis. For global gene expression profiling to become more widely practiced there is a need to build up an infrastructure for data archiving and

future data mining. In these efforts, it is important to keep in mind that the ultimate goal of genomic exploration is to develop physiological insights to better direct process and cell line development efforts. Cell physiology is a key element in this effort, and it will be important to better develop and integrate cell physiology in data analysis pipelines, as opposed to focusing on determining statistical significance. The power of these global survey technologies will be realized in the context of the cumulative magnitude of information derived to build new knowledge. The immense data generated will provide more opportunity to derive new insight through pattern recognition. We believe these efforts will eventually lead to a better physiological understanding of cell culture processes. We believe that we are facing the dawn of a new era where bioprocess research will greatly benefit from genomic technologies.

Acknowledgements The support from Pfizer, Inc. and Bayer Healthcare for the cell culture research work in W-S. Hu's laboratory are gratefully acknowledged. KFW was supported by the NIH Biotechnology Training Grant (GM08347). The bioinformatic support was provided by The Minnesota Supercomputing Institute.

References

- Aebersold R, Mann M (2003) Mass spectrometry-based proteomics. *Nature* 422:198–207
- Alete DE, Racher AJ, Birch JR, Stansfield SH, James DC, Smales CM (2005) Proteomic analysis of enriched microsomal fractions from GS-NS0 murine myeloma cells with varying secreted recombinant monoclonal antibody productivities. *Proteomics* 5:4689–4704
- Chakraborty A, Regnier FE (2002) Global internal standard technology for comparative proteomics. *J Chromatogr A* 949:173–184
- Chee M, Yang R, Hubbell E, Berno A, Huang XC, Stern D, Winkler J, Lockhart DJ, Morris MS, Fodor SP (1996). Accessing genetic information with high-density DNA arrays. *Science* 274:610–614
- Chen Z, Southwick K, Thulin CD (2004) Initial analysis of the phosphoproteome of Chinese hamster ovary cells using electrophoresis. *J Biomol Tech: JBT* 15:249–256
- Dahlquist KD, Salomonis N, Vranizan K, Lawlor SC, Conklin BR (2002) GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat Genet* 1:19–20
- DeRisi JL, Iyer VR, Brown PO (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278:681–685

- Durbin R, Eddy S, Krogh A, Mitchison G (1998) Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge UK, Cambridge University Press
- Gadgil C, Rink A, Beattie CW, Hu WS (2002) A Mathematical Model for Suppression Subtractive Hybridization. *Comp Funct Genomics* 3:405–422
- Gadgil M, Lian W, Gadgil C, Kapur V, Hu W-S (2005) An analysis of the use of genomic DNA as a universal reference in two channel DNA microarrays. *BMC Genomics* 6
- Gladney CD, Bertani GR, Johnson RK, Pomp D (2004) Evaluation of gene expression in pigs selected for enhanced reproduction using differential display PCR and human microarrays: I. Ovarian follicles. *J Anim Sci* 82:17–31
- Gu J, Gu X (2003) Induced gene expression in human brain after the split from chimpanzee. *Trends Genet* 19:63–65
- Guo QM (2003) DNA microarray and cancer. *Curr Opin Oncol* 15:36–43
- Gygi S, Rist B, Gerber S, Turecek F, Gelb M, Aebersold R (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* 17:994–999
- Gygi SP, Corthals GL, Zhang Y, Rochon Y, Aebersold R (2000) Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc Natl Acad Sci USA* 97:9390–9395
- Harrison PM, Kumar A, Lang N, Snyder M, Gerstein M (2002) A question of size: the eukaryotic proteome and the problems in defining it. *Nucleic Acids Res* 30:1083–1090
- Hayduk EJ, Choe LH, Lee KH (2004) A two-dimensional electrophoresis map of Chinese hamster ovary cell proteins based on fluorescence staining. *Electrophoresis* 25:2545–2556
- Hayduk EJ, Lee KH (2005) Cytochalasin D can improve heterologous protein productivity in adherent Chinese hamster ovary cells. *Biotechnology & Bioengineering* 90:354–364
- Hunt DF, Henderson RA, Shabanowitz J, Sakaguchi K, Michel H, Sevilir N, Cox AL, Appella E, Engelhard VH (1992) Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry. *Science* 255:1261–1263
- International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* 409:860–921
- Kaufmann H, Mazur X, Fussenegger M, Bailey JE (1999) Influence of low temperature on productivity, proteome and protein phosphorylation of CHO cells. *Biotechnol Bioeng* 63:573–582
- Kim H, Zhao B, Snesrud E, Haas B, Town C, Quackenbush J (2002) Use of RNA and genomic DNA references for inferred comparisons in DNA microarray analysis. *Biotechniques* 33:924–930
- Korke R, De Leon Gatti M, Lau AL, Lim JW, Seow TK, Chung MC, Hu WS (2004) Large scale gene expression profiling of metabolic shift of mammalian cells in culture. *J Biotechnol* 107:1–17
- Lee MS, Kim KW, Kim YH, Lee GM (2003) Proteome analysis of antibody-expressing CHO cells in response to hyperosmotic pressure. *Biotechnol Prog* 19:1734–1741
- Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, Clamp M, Chang JL, Kulbokas EJ, Zody MC, et al (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. 438:803–819
- Lockhart DJ, Dong H, Byrne MC, Follett MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* 14:1675–1680
- Moody DE, Zou Z, McIntyre L (2002) Cross-species hybridisation of pig RNA to human nylon microarrays. *BMC Genomics* 3:27
- Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420:520–562
- Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M (2002). Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* 1:376–386
- Pattyn F, Speleman F, De Paepe A, Vandesompele J (2003) RTPrimerDB: the real-time PCR primer and probe database. *Nucleic Acids Research* 31:122–123
- Rabilloud T (2002) Two-dimensional gel electrophoresis in proteomics: old, old fashioned, but it still climbs up the mountains. *Proteomics* 2:3–10
- Rat Genome Sequencing Project Consortium (2004). Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* 428:493–521
- Rose K, Simona MG, Offord RE, Prior CP, Otto B, Thatcher DR (1983) A new mass-spectrometric C-terminal sequencing technique finds a similarity between gamma-interferon and alpha 2-interferon and identifies a proteolytically clipped gamma-interferon that retains full antiviral activity. *Biochem J* 215:273–277
- Ross P, Huang Y, Marchese J, Williamson B, Parker K, Hattan S, Khainovski N, Pillai S, Dey S, Daniels S, et al (2004) Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Mol Cellular Proteomics* 3:1154–1169
- Schneider LV, Hall MP (2005) Stable isotope methods for high-precision proteomics. *Drug Discovery Today* 10:353–363
- Seow TK, Korke R, Liang R, Ong S-E, Ou K, Wong K, Hu WS, Chung M (2001) Proteomic investigation of metabolic shift in mammalian cell culture. *Biotechnol Progr* 17:1137–1144
- Seth G, Philp RJ, Denoya CD, McGrath K, Stutzman-Engwall KJ, Yap MG, Hu W-S (2005) Large scale gene expression analysis of cholesterol dependence in NS0 cells. *Biotech Bioeng* 90:552–567
- Seth G, Ozturk M, Hu W-S (2006) Reverting cholesterol dependence of NS0 cells by altering epigenetic gene silencing. *Biotech Bioeng* 93:820–827

- Shen D, Sharfstein ST (2006) Genome-wide analysis of the transcriptional response of murine hybridomas to osmotic shock. *Biotechnol Bioeng* 93:132–145
- Smales CM, Dinnis DM, Stansfield SH, Alete D, Sage EA, Birch JR, Racher AJ, Marshall CT, James DC (2004) Comparative proteomic analysis of GS-NS0 murine myeloma cell lines with varying recombinant monoclonal antibody production rate. *Biotechnol Bioeng* 88:474–488
- Talaat A, Howard S, Hale W, Lyons R, Garner H, Johnston S (2002) Genomic DNA standards for gene expression profiling in *Mycobacterium tuberculosis*. *Nucleic Acids Res* 30
- Unlu M, Morgan ME, Minden JS (1997) Difference gel electrophoresis: a single gel method for detecting changes in protein extracts. *Electrophoresis* 18:2071–2077
- Van Dyk DD, Misztal DR, Wilkins MR, Mackintosh JA, Poljak A, Varnai JC, Teber E, Walsh BJ, Gray PP (2003) Identification of cellular changes associated with increased production of human growth hormone in a recombinant Chinese hamster ovary cell line. *Proteomics* 3:147–156
- Williams B, Gwirtz R, Wold B (2004) Genomic DNA as a cohybridization standard for mammalian microarray measurements. *Nucleic Acids Res* 32
- Wlaschin K, Nissom PM, de Leon Gatti M, Fern PFO, Arleen S, Tan KS, Rink A, Cham B, Wong K, Yap M, Hu W-S (2005) EST sequencing for gene discovery in Chinese hamster ovary cells. *Biotech Bioeng* 91:592–606
- Wong VVT, Nissom PM, Sim S-L, Yeo JHM, Chuah S-H, Yap MGS (2006) Zinc as an insulin replacement in hybridoma cultures. *Biotechnol Bioeng* 93:553–563