Biotechnology and Bioprocess Engineering **2005**, 10: 385-399 ©KSBB

The Role of High-throughput Transcriptome Analysis in Metabolic Engineering

Michael C. Jewett, Ana Paula Oliveira, Kiran Raosaheb Patil, and Jens Nielsen*

Center for Microbial Biotechnology, BioCentrum-DTU, Technical University of Denmark, Søltofts Plads, Building 223, DK-2800 Kgs. Lyngby, Denmark

Abstract The phenotypic response of a cell results from a well orchestrated web of complex interactions which propagate from the genetic architecture through the metabolic flux network. To rationally design cell factories which carry out specific functional objectives by controlling this hierarchical system is a challenge. Transcriptome analysis, the most mature high-throughput measurement technology, has been readily applied in strain improvement programs in an attempt to identify genes involved in expressing a given phenotype. Unfortunately, while differentially expressed genes may provide targets for metabolic engineering, phenotypic responses are often not directly linked to transcriptional patterns. This limits the application of genome-wide transcriptional analysis for the design of cell factories. However, improved tools for integrating transcriptional data with other high-throughput measurements and known biological interactions are emerging. These tools hold significant promise for providing the framework to comprehensively dissect the regulatory mechanisms that identify the cellular control mechanisms and lead to more effective strategies to rewire the cellular control elements for metabolic engineering.

Keywords: metabolic engineering, transcriptome, gene expression, bioinformatics, systems biology, data integration, cell factory

INTRODUCTION

Industrial application of microorganisms as cell factories for production of fuels, chemicals, enzymes, food ingredients, and protein therapeutics generally requires development of production strains with improved properties (*e.g.* strains with a higher product yield). However, since microorganisms have naturally evolved to perform a myriad of operations required for cellular growth and fitness within their environment, achieving desired phenotypes that are, in general, not the "Darwinian optimum", presents a formidable challenge. Given the complex nature of the catalytic inventory, metabolic pathways, signaling circuits, and regulatory networks of microorganisms, it is important to develop rational approaches to understand this complexity and optimize for preferred characteristics. Rewiring the cellular control elements of microbial metabolism and regulation through targeted genetic changes using recombinant DNA technology - termed metabolic engineering - is one rational approach to design cell factories [1-4].

Metabolic engineering seeks to design microbial strains with improved phenotypic behavior by exploiting metabolic flux networks in an iterative strategy proceeding through: (1) characterization, (2) analysis, and (3) design (Fig. 1). As a function of transcription, translation, post-translational modification, signal transduction, protein-protein interaction, and protein localization [5], metabolic fluxes represent the final outcome of cellular regulation. Therefore, understanding the mechanisms which control how carbon sources are transformed through an intricate series of biochemical reactions can provide critical insight into the origin of strain distinguishing characteristics (Fig. 2). In the metabolic engineering design cycle, mathematical modeling and analysis of the cellular flux patterns are first used to map and identify factors conferring phenotypic traits of interest. Based on this model guided assessment, improved strains with altered metabolic flux networks are designed, and subsequently constructed through the introduction of targeted genetic changes using recombinant DNA technology or environmental perturbation [6]. Targeted changes are evaluated and characterized for their ability to achieve desired phenotypic landscape and the cycle is either continued or terminated depending on the overall objective for strain performance. Insights leading to desired properties can even be transferred from one organism to another. Numerous examples of metabolic engineering, such as terpenoid production in *Escherichia coli* [7], lycopene production in *E. coli* [8], and improving galactose utilization by altering a regulatory network of *Saccharomyces cerevisiae* [9] have illustrated the power of this approach.

One of the major limitations of metabolic engineering

***Corresponding author**
Tel: +45-4525-2696 Fax: +45-4588-4148
e-mail: jn@biocentrum.dtu.dk

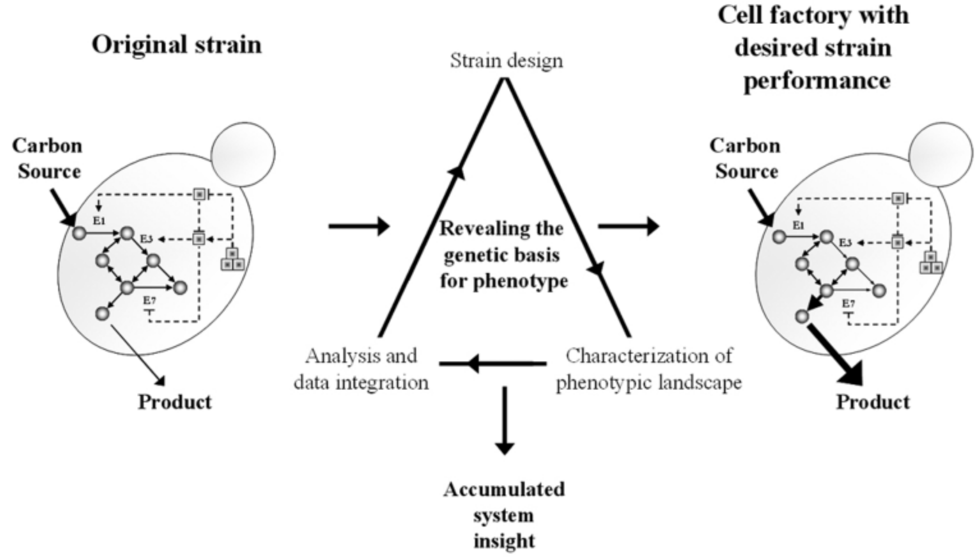

Original strain
Carbon Source
E1
E3
E7
Product
Strain design
Revealing the genetic basis for phenotype
Analysis and data integration
Characterization of phenotypic landscape
Accumulated system insight
Cell factory with desired strain performance
Carbon Source
E1
E3
E7
Product

Fig. 1. An iterative strategy for cell factory design. In this schematic, an original strain is engineered to overexpress a specific product by directing the carbohydrate substrate through metabolism as shown. Circle, metabolite. E_i, enzyme. Square, regulatory protein. Arrows indicate reactions. Dotted arrows indicate activation/up-regulation and ⊤ arrows indicate repression/down-regulation. Arrow thickness depicts the relative flux magnitude increased in the cell factory having the desired strain performance.

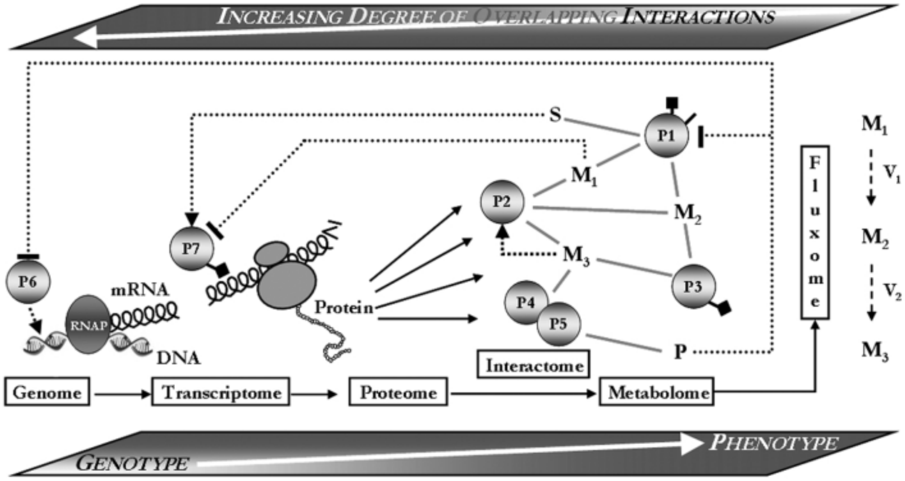


Fig. 2. Visual representation of the complex nature of the catalytic inventory, metabolic pathways, signaling circuits, and regulatory networks of microorganisms. The overall phenotypic response is characterized by numerous interactions which propagate from the genetic architecture through the metabolic flux network. DNA (the genome) is transcribed to mRNA (the transcriptome) and translated to proteins (the proteome). Subsequently, post-translational modifications expand the size of the proteome. Proteins (P_i) help determine cellular function by carrying out enzymatic functions that convert small molecule substrates (S) into metabolites (M_i) and products (P). In addition, proteins interact with other proteins (P4:P5) and DNA (P6:DNA) (the interactome). While not depicted, the location of proteins within specialized compartments also impacts the regulatory structure of the cell (the locasome). Proteins and metabolites make up a complex network of players that can both activate (↑) and inhibit (⊤) other processes in the cell. These regulatory interactions are noted with dotted lines in the figure.

is the lack of experimental and theoretical information describing regulatory and/or kinetic information upon which decisions are based. Specifically, it is often difficult to predict the overall consequences, especially secondary and non-linear effects, of a specific genetic or environmental modification to the fluxes dictating cellular metabolism [10]. To overcome this limitation, a global approach, which integrates the regulatory structure and coordinated activities of multiple cellular processes is required.

Recently, significant progress has been made in analyzing cellular function with quantitative high-throughput measurements of characteristic cellular components (*e.g.* genes, mRNA, proteins, metabolites) [11], linking interactions between cellular components by computational algorithms [12-14], and constructing stoichiometric models that accurately represent metabolic networks [15,16]. These developments have given researchers key tools to help understand and decipher mechanisms underlying cellular function. Coined 'systems biology', quantitatively describing properties of biological systems through integration of cell-wide measurements (*e.g.* quantifying genome-wide mRNA abundance levels) holds significant promise for *developing predictive models* that facilitate drug discovery, treatment of diseases, and improve bioprocesses [17-22]. By providing insight into the functional and regulatory behavior of the metabolic flux network, systems biology tools will enable metabolic engineers to gain a more quantitative link between genotype and phenotype than previously possible [18,20].

Of the numerous high-throughput tools that exist to probe cellular response to different perturbations (*e.g.* genetic variance or altered growth conditions), genome-wide transcriptional profiling is used most prolifically. Transcriptome analysis enables the simultaneous measurement of thousands of mRNA levels in parallel for the identification of up- and down-regulated genes. The application of microarray technologies as a diagnostic of cellular status [23-25] and tool to discover and assign function to unknown ORFs [26] has witnessed a dramatic expansion in the past few years. Moreover, use in metabolic engineering applications has also been reported [27-29]. Despite showing some promise in mapping phenotypic topography, use of DNA microarrays in metabolic engineering has been shown to be limited by three main factors. First, mRNA profiles do not explicitly reflect downstream cellular responses (*i.e.* protein activity or metabolic fluxes), making it difficult to identify the genetic basis for strain performance. Second, it is difficult to separate biological from non-biological (noise) changes in expression. Third, isolating the effects of a single variable of interest which will reveal useful information for future metabolic engineering (*i.e.* specific genetic perturbations) through appropriate experimental design is challenging. As a result, expression levels of hundreds of genes may be altered that are not a direct result or cause of the phenotype of interest.

To address these limitations, new principles are being developed with the objective to reveal the genetic architecture responsible for specific phenotypes. This review focuses on this aspect, and particularly we concentrate on the role of high-throughput transcriptome analysis within a metabolic engineering framework. We discuss examples of employing DNA microarrays in metabolic engineering, highlight the importance of rational experimental design strategies, consider several key questions dealing with conventional transcriptome data generation and analysis, and describe new approaches designed to incorporate mRNA expression data with genome-scale metabolic models to infer global regulatory patterns. We also underscore the ambitious challenge of modeling biological complexity, and comment on the future outlook of omics data integration in metabolic engineering.

mRNA Profiling to Reveal Metabolic Engineering Targets

The use of mRNA profiling in metabolic engineering to extract genotype-phenotype relationships has not experienced the same explosion of successfully documented stories as more general analysis of genetic networks [30-33]. However, in the framework of designing cell factories, identification of critical effectors linked to, or responsible for, specific phenotypes have been reported. To identify targets for metabolic engineering, genome-wide expression patterns for two or more strains with different aptitudes for a preferred trait (*e.g.* an antibiotic overproduction strain and its parent) are typically compared. While not the focus of this review, genomic disruption and plasmid based strategies (*e.g.* parallel gene trait mapping, [34]) have also been developed to screen for trait conferring effectors (see [27,35] for review).

To characterize features responsible for ethanol tolerance in *E. coli*, Gonzalez *et al*. compared a strain evolved by directed evolution to be resistant to ethanol with a strain designed for ethanol overproduction [36]. Careful investigation of six distinct growth conditions established differentially expressed genes between the strains in multiple branches of metabolism. Pathways surrounding glycine degradation, osmotic stress, the *mar* multiple drug resistance system, and gene products regulated by FNR were identified as significant for ethanol tolerance. Along with further investigations, these observations indicated a genetic basis for tolerance and suggest strategies for future metabolic engineering of ethanol overproduction strains.

In addition to investigating the difference between strains developed by molecular breeding strategies, there is also strong interest in using genome-scale mRNA profiling to characterize phenotypes obtained by classical strain improvement programs (*e.g.* chemical mutagenesis). This strategy has been applied to unravel the rationale behind desired phenotypes which are caused by unknown mutations. In one case, Lum *et al*. explored discriminatory gene expression patterns to characterize antibiotic overproducing strains [37]. First, transcriptional profiles of industrial overproducing strains for erythromycin (*Saccharopolyspora erythraea*) and tylosin (*Streptomyces fradiae*), and their respective parent (non-overproducing) strains, were obtained from batch growth experiments. Second, prominent differences between the two sets of gene expression data were shown to result in two distinct regulatory control structures. In one scheme, the genes encoding the erythromycin biosynthetic cluster were co-expressed for a longer period of time in the *S. erythraea* overproducing strain relative to the parent. Alternatively, while the expression pattern for genes encoding the tylosin biosynthetic cluster were similar in both *S. fradiae* strains, several genes, including some involved in precursor biosynthesis, had altered expression levels.

Thus, the flux of metabolite building blocks entering the tylosin biosynthetic pathway is likely increased. The two control strategies described above suggest potential metabolic engineering targets that influence the overproduction of antibiotics.

To gain insight into the xylose utilization pathway of a metabolically engineered strain of *S. cerevisiae*, mRNA expression profiles were used to characterize differences in transcription levels between the engineered parent and variant generated from chemical mutagenesis [38]. Since the mutant strain boasted a higher growth rate on xylose relative to the parent strain, this comparison was intended to identify gene targets involved in increased xylose utilization rates. Transcriptome analysis identified a number of candidate genes with altered expression levels between the two strains. From this wealth of biological information, Wahlbom *et al.* pursued only one, *PET18* (which encodes a transcriptional regulator reported to be an effector of growth on non-fermentable carbon sources), for a more thorough investigation. Unfortunately, xylose growth was unaffected by either deletion or overexpression of *PET18*. This report highlights the frustrating fact that differential gene expression alone may not indicate how the control structure of the metabolic flux network has changed.

To date, the most successful strategy to reveal targets for metabolic engineering through genome-wide expression techniques has been to analyze a collection of strains (rather than just 2) comprising of a desired phenotype to identify discriminatory genes involved in the metric for strain performance. In one example, association analysis of transcriptional and metabolite profiles from a library of unsequenced *Aspergillus terreus* fungal strains were integrated to discover the essential parameters and genes influencing the biosynthesis of lovastatin and (+)-Geodin (two commercially relevant natural products) [39]. First, a diverse collection of strains was generated with varying capacities to produce lovastatin and (+)-Geodin. Second, DNA microarrays and high-performance liquid chromatography were used to characterize transcript and metabolite profiles of these strains. Third, the genetic mechanisms that control the biosynthesis of lovastatin and (+)-Geodin were revealed by expressing transcriptional and metabolite data as ratios with respect to the parental reference strain and then correlating these ratios to product titers using Pearson correlations and principal component analysis (PCA). To design for improved lovastatin titers, Askenazi *et al.* went on to show that specific promoter sequences, which correlate to lovastatin biosynthesis, could be fused to antibiotic resistance genes in order to select for strains with improved antibiotic resistance, and consequently, lovastatin production. Subsequent engineering of desired strains led to an improvement of lovastatin biosynthesis by more than 50%.

To improve flux through the galactose utilization pathway in *S. cerevisiae*, Bro *et al.* used DNA microarrays to compare the genome-wide transcription profiles of three strains with different capacities to utilize galactose as a carbon source (Bro *et al.*, submitted). Previously, the regulatory architecture of the Leloir pathway had been engineered in two of these strains to increase the maximum specific galactose uptake rate relative to the wild type strain [9]. Initial mRNA profiling analysis did not uncover underlying mechanisms for genes associated with galactose uptake rates. However, further examination of a smaller subset of genes known to be directly involved in the galactose uptake system singled out *PGM2*, the major isoform of phosphoglucomutase, as having a high probability of being significantly changed. Consistent with this observation, overexpression of the *PGM2* gene resulted in a 70% increase in the maximum specific galactose uptake rate as compared to the wild type.

Collectively, these studies highlight three critical lessons of using transcriptional information for the design of cell factories. First, metabolic engineering strategies can directly exploit insights gained through genome-wide transcriptional analysis for generation of desired strains. Information gained through microarray analysis contributes to a better understanding of cellular organization and offers the 'possibility of elucidating global regulatory processes' [40]. Second, the underlying assumption that mRNA profiles can identify key genetic trends that may be critical for understanding cellular flux and regulation is in many cases, misleading. This is not to downplay the important role that transcriptional analysis can play in the design of cell factories, but rather to accentuate the fact that the cell has more than just one level of control to alter the flux network. Third, since many expression levels are often changed, analysis is typically guided by preconceived notions of cellular physiology, in which mRNA levels of pathways already known to be important in a particular phenotype are scrutinized while some of the entire genome-scale information is discarded in the search for new targets.

To gain a more detailed description of cellular response to specific perturbations, researchers are beginning to more readily incorporate the study of two or more omic responses simultaneously [41]. This was first reported by Ideker *et al.* as strategy to provide evidence for the explicit physiological interactions governing cellular response in the galactose utilization pathway of yeast [13]. Here, DNA microarrays, proteomics, protein-protein interactions, and protein-DNA interactions were integrated to map regulatory phenomena and suggest new hypotheses that attempt to dissect the hierarchy of interconnections which describe cellular complexity. More recently, use of high-throughput analytical approaches have been extended to metabolic engineering applications. These studies have only further solidified the disparity between mRNA levels and protein levels and/or *in vivo* metabolic fluxes and stress the importance of integrating several cell-wide measurements to garner a comprehensive view of the biological system in question.

Daran-Lapujade *et al.* compared genome-wide transcript levels with *in vivo* fluxes obtained using a reconstructed genome-scale model to characterize growth and regulation through central carbon metabolism in *S. cerevisiae* [42]. To isolate changes in metabolism caused by a single change in growth nutrient (glucose, maltose, ethanol, or acetate), chemostat cultivation was employed. Comparison of mRNA profiles for genes that encode enzymes corresponding to estimated *in vivo* metabolic fluxes indicated that seldom are the two precisely correlated. While there was a strong correspondence between the flux distribution and transcriptional regulation for gluconeogenesis and glyoxylate cycle (pathways known to be tightly controlled at the transcriptional level), robust association accounted for only a relatively small percentage of the assembled data. Specifically, significant variance was observed for glycolysis and the tricarboxylic acid cycle. The study underscores that transcriptional data on their own have a limited capability to discover phenotype, since they do not necessarily provide a good litmus test for *in vivo* activities downstream of transcriptional regulation. In relation to this study, the data provided suggest that post-translational modification plays an important role in the phenotypes observed.

Consistent with the results presented above, Tummala *et al*. demonstrated that the overall correlation between *in vivo* fluxes and the transcriptome is, in general, weakly positive [43]. By comparing strains with different capacities to produce butanol and acetone, they combined the use of fluxome and transcriptome analysis to investigate the genetic basis controlling organic solvent production in recombinant *Clostridium acetobutylicum* strains. Similarities and discrepancies between these omic responses were used to suggest flux-controlling steps, regulatory mechanisms, and metabolic engineering targets (such as acid uptake *via* the CoA transferase pathway or the possible control of alcohol production on the ferredoxin oxidoreductase). By assuming that protein concentrations would only need to be altered at flux-controlling nodes, transcriptional analysis can help to identify rate-determining steps.

Correlating genome-wide transcriptional levels with proteomic responses to specific perturbations has also been pursued. Yoon *et al.* sought to systematically understand the physiological and metabolic changes that occur as a result of high cell density cultivation in *E. coli* by integrating transcriptomic data with information content obtained from 2D-gel electrophoresis coupled with MALDI-TOF MS [44]. While several discrepancies existed, the patterns between the two were mostly similar when comparing protein spots with differentially expressed mRNA levels. Although these results contrast with previous reports regarding the similarity of transcriptional and translational response [13,45], it substantiates the claim that cellular responses are elegantly linked.

Strategies described in the literature for direct use of transcriptome data have been extremely profitable in demonstrating how to identify targets for metabolic engineering and for allowing researchers to recognize the limitations of these approaches which on their own, only reveal the genetic expression architecture of the cell. These, as stand alone principles, are applicable to a number of problems under investigation. However, since phenotype is not always directly translatable from transcriptomic data due to the complexity of cellular regulation, mRNA profiling alone is not the answer for metabolic engineering. Clearly, our incomplete understanding of how biochemical flux networks are connected within the cell compounds the problem. Improved tools for integrating transcriptional data with other omic responses, and extracting information from transcriptional data are emerging. They hold significant promise for revolutionizing the field of metabolic engineering. To provide the relevant background for understanding these tools, we now shift our focus towards conventional generation and analysis of transcriptome data.

Generation of Transcriptome Data

The first step in generation of transcriptome data is selection of a microarray platform. DNA technologies are based on the hybridization of labeled RNA or DNA prepared from extracted cellular mRNA to highly ordered DNA sequences attached to a solid matrix [25,33,46]. While many array technologies exist, spotted DNA microarrays and high-density oligonucleotide microarrays (commercially available from Affymetrix) are most widely used [47,48].

After selection of the appropriate array technology, experimental design must be considered. The design process should not only clearly define the biological question and/or hypothesis, but also anticipate the focal point of data analysis. Data analysis typically centers on comparative profiling of expression patterns under particular conditions (*e.g.* over a time series or during cell cycle stages), identification of relevant genes associated with a desired phenotype, or genetic responses to perturbations. In addition to anticipating data analysis, the experimental strategy must also ensure that enough biological replicates are taken to ensure statistical significance. It has been shown that the number of false negatives among the significant changing genes increases linearly when the number of biological replicates decreases [49].

When studying the effect of a genetic or environmental perturbation, it is important that all conditions, except those under investigation, are kept constant. Here, we emphasize two obstacles to achieving this objective. First, mRNA levels are known to change with the specific growth rate [50,51]. Second, mRNA profiles are influenced by the dynamic nature of batch cultures (*i.e.* different media component concentrations over time) [42]. To illustrate the latter, Daran-Lapujade *et al*. only found 117 significantly changed genes by exploring the expression level changes of yeast chemostat cultures grown on 4 different carbon sources (glucose, maltose, ethanol, or acetate). In stark contrast, previously reported data comparing batch cultures grown on glucose and ethanol alone found up to 600 significantly changed transcripts [52]. This underscores how effects inherent to batch cultures can obscure the parameter under examination. In light of these results, it is preferable to use chemostat cultures rather than batch cultures for evaluation of different environmental conditions or different mutants [53]. One cautionary note: in the case of organisms with different metabolic modes (*e.g.* both respirative and respirofermentative growth under aerobic conditions), mutants

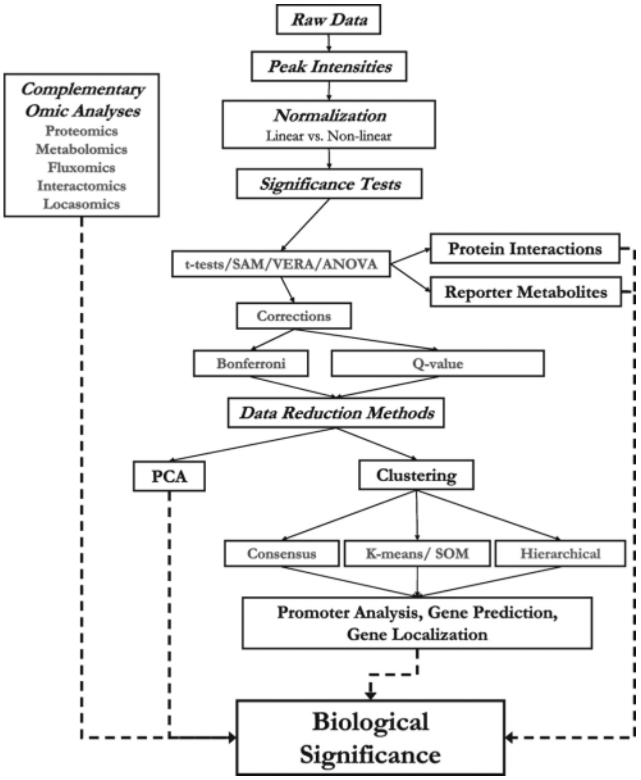


Fig. 3. Common tools used to attain biological significance and understanding from high-throughput transcription data. This roadmap serves to highlight key strategies to uncover the relationship between genotype and phenotype.

may present different metabolic modes at the same specific growth rate. This may distort evaluation of single parameters in chemostat cultures performed at the same growth rate. Again, appropriate experimental design should be taken into consideration to minimize the influence of non related factors.

Once the appropriate microarray technology and experimental design have been selected, several experimental steps need to be performed to attain an array image that can be used for further analysis. These steps include: 1) implementing the desired experiment 2) collecting cell samples 3) extracting, labeling, and hybridizing RNA to the microarray, and 4) scanning the microarray chip to obtain an image. These steps have been outlined by others [54].

From Array Image to Expression Intensities

The generation of microarray images is the first step in gene expression analysis. Before performing gene-to-gene comparisons among samples, other pre-processing steps are often necessary. In this section we describe these steps and highlight the important decisions to be made (Fig. 3). We focus our description on methods dealing with Affymetrix® microarray images. These oligonucleotide arrays provide quantification of mRNA levels through the use of perfect match (PM) and mismatch (MM) probes. Each gene has several oligonucleotide probes and the intensity of a gene is given by an "average" intensity of the respective probe-set.

Normalization

Microarray data analysis is based on the underlying hypothesis that the measured intensity for each probe-set represents the relative expression level of its corresponding gene. To ensure comparability between microarrays under investigation, it is advisable to follow the MIAME international standards [55], to use identical arrays, and to use equivalent amounts of sample per array. In addition, before comparing gene expression data between arrays, a number of transformations must be carried out to adjust the measured intensities and identify low-quality microarrays. This step is called normalization [56-60]. Microarray normalization includes both methods to correct for overall brightness and methods to remove an often observed systematic signal-dependent bias. Nonlinear methods (*e.g.* iteratively search of an invariant set [60] and qspline [58]) are generally preferred over linear methods.

Expression Index Calculation

Once the signals of all oligo probes are normalized, the overall intensity of a given gene from the individual intensities of the oligo probes in the respective probe-set is calculated. There are a few different ways to perform this data condensation, and they may lead to different results. The easiest way, as calculated by the early version of Affymetrix® software (MAS 4.0), is by simple average of all probes signal (not including probes that deviate more than three standard deviations from the mean). More advanced methods are used to calculate weighted averages. For example, Li and Wong [61] developed an expression index calculation algorithm which multiplies each probe signal by a scaling factor obtained by fitting a statistical model to the series of experiments being analyzed. This approach takes into consideration that different probes may respond differently. Average calculation methods can be applied using either a PM-MM model (perfect-match minus mis-match) or a PM-only model. Affymetrix has included MM control probes to act as specificity controls, allowing correction for both local background and cross-hybridization [25]. Selection of a PM-MM model (difference between PM and MM intensities) is therefore recommended by Affymetrix to detect significant signals. On the other hand, some authors [61-63] have argued that subtracting MM to PM merely increases the noise of the signal and hence support the use of PM-only models. Moreover, the MM response has been shown to largely reflect interaction with the intended PM transcript [64].

Filtering the Desired Features

After normalization and expression index calculation, expression intensities of each probe-set are comparable between different arrays. In most cases a probe-set corresponds to a unique gene. However, in other cases, a gene can be represented by several different probe-sets, and it may be convenient to calculate a unique expression value for that gene. Standard procedures include: summing all, averaging all, or choosing the most representative of the probe-sets. Subsequent filtering can also be used to remove data that will not be used in further analysis (e.g. probe-sets that are absent in the experiments) or select genes that will be targeted for investigation (which may be every transcript of well identified ORFs). Biologically relevant patterns of expression can next be identified by statistical significance analysis, clustering algorithms, or advanced methods integrating gene expression with biological knowledge.

Conventional Analysis of Transcriptome Data

Significance of Change Analysis

One of the important tasks in microarray analysis is the identification of genes that are significantly changed between different experimental conditions or between different strains. Notably, a fold-change in expression level is not always an indicator of significance. For example, the mRNA level of a certain gene might change only 1.5 fold, but it may still be statistically significant.

The main idea behind most of the statistical significance tests is to assess whether two different groups of numbers (in this case-expression levels of a gene in replicates for each condition/strain) have a similar mean. The most widely used test for this purpose is called the student's t-test. Several reports have also described improved significance of change tests specifically designed for microarray data. In general, these methods attempt to capture and dilute the noise of the entire array data set in a systematic way. Examples include SAM (Significance Analysis of Microarray) [49] and VERA (Variability and ERror Assessment) [65]. Significance of change tests assign a probability value (p-value) for each gene under analysis. The p-value of a gene indicates the likelihood that the observed differential expression is by chance alone. Thus, the lower the p-value, the more significant the change. When comparing more than two data sets simultaneously, analysis of variance (ANOVA) is used.

After assigning p-values, the next and equally important step is to choose a p-value to be regarded as significant. Usually p-value cut-off of 0.05 (95% confidence) is chosen. The choice of method for deciding an appropriate p-value cut-off for microarray data is dependent on the objective and nature of analysis. Most methods attempt to account for the effect of multiple testing (*i.e.* the probability of finding significant changes when same hypothesis is tested several times). In general, if very few genes are expected to change during an experiment, a strict cut-off value based on a Bonferroni correction factor is preferred. Bonferroni correction is applied by choosing a new cut-off value obtained simply by dividing the desired p-value (*e.g.* 0.05) by the number of genes tested. This correction guarantees with 95% confidence that the number of false positives will be less than or equal to one. Although this strict measure is beneficial in certain cases, for many experimental studies a strict cut off may lead to several false negatives. A recent method presented by Storey and Tibshirani [66] offers a good alternative for choosing p-value cut-off based on false discovery rate. In the proposed method, a new measure of statistical significance called q-value is assigned to each gene and can easily be interpreted in terms of false discovery rate. The q-value approach leads to a less stringent cut-off while maintaining a good balance between false positives and false negatives.

Reduction of Dimensionality

Genome-wide analysis of gene expression generates tens of thousands of data points with at least as many variables as measured transcripts. This high-dimensionality of gene expression data makes it difficult to visualize relationships between genes and grouping of experiments by similarity of expression profiles. Several methods exist to reduce the dimensionality of array data. These approaches facilitate visualization, allow for characterization of the data structure, and separate biological meaningful information from noise. Examples of such methods are cluster analysis [23], multidimensional scaling [67], principal component analysis (PCA) [68], and singular value decomposition (SVD) [69].

Gene expression data can be represented in a matrix form, with each row of the matrix representing the expression profile on a given gene throughout the experimental conditions, and each column representing the genome-wide expression in a given experiment. Methods such as multidimensional scaling, PCA, and SVD allow projection of rows and/or columns of the data matrix in a plane such that similar rows/columns are located close to each other, therefore allowing grouping of genes and/or experiments by similarity.

Principal component analysis decomposes the original multi-dimensional space in a low-dimensional space of dimension n, where n is the number of principal components. PCA identifies the direction in space that captures most of the variance, and this direction corresponds to the first principal component (PC1). The second principal component (PC2) is determined as being the vector orthogonal to PC1 that captures most of the remaining variance (and so on to determine other principal components). Expression data can then be projected onto this low-dimensional space, whose axes are the principal components. Often, PC1 and PC2 retain most of the variance in gene expression data, making possible a two-dimension visualization of the relationships between genes and experiments. An example of a two-dimensional PCA bi-plot is depicted in Fig. 4.

A number of programs can be used to perform PCA. Both general data-mining software and gene expression analysis dedicated software are commonly used. For better interpretation, it is convenient to mean-center and scale the data (*i.e.* transform each variable vector so it has mean 0 and standard deviation 1). Moreover, it may be convenient to perform PCA only on significantly changed genes. Although PCA decomposition "filters" for genes with high variance, this variance can be either due to noise or due to biological relevant changes in gene ex-

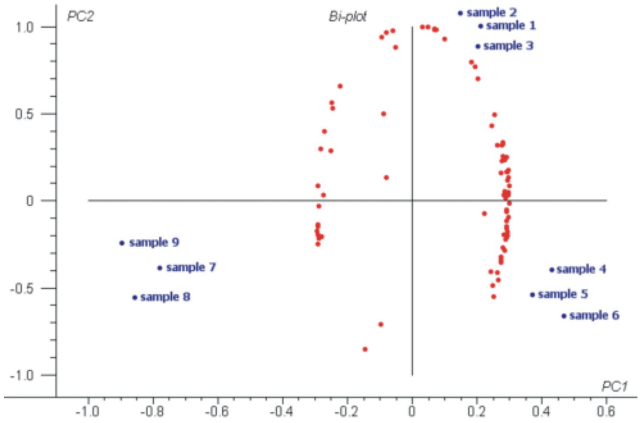


Fig. 4. Illustration of a PCA bi-plot for gene expression data. Loadings (gene transcripts) are represented by non-labeled dots; scorings (samples) are represented by labeled dots. The nine samples are distributed in three groups, suggesting the existence of three different groups of biological triplicates.

pression. Consequently, reduction in the non-biological variance greatly facilitates interpretation of PCA.

A PCA bi-plot depicts both loadings and scorings, that is, the projection of both genes and samples in the principal component space. Loadings contain information on how variables relate to each other, while scorings refer to how samples are related. Analysis of loadings tells us how the variance of a certain gene is explained by that principal component. Loading weights should be read in the principal component axis, and genes with high absolute values are the ones that contribute more for that component. The distribution of the scorings tells us how the samples can be explained by the loadings. For example, a sample standing in the upper-right quadrant of a bi-plot is positively influenced by the genes standing also in the upper-right quadrant and is negatively influenced by genes in the lower-left quadrant.

Clustering

Clustering is one of the first and widely used methods for analysis of large amounts of gene expression data [56,70,71]. The basic idea consists of grouping genes based on their similarity profile [23]. Genes sharing a common profile throughout a series of experiments cluster together, and can be further analyzed as a unique group. Since this grouping organizes data into manageable clusters, it is also considered a dimension-reduction method. Notably, clustering has shown to exhibit a certain predictive power. When a gene with unknown function is assigned to the same cluster as genes with known function, the function of the unknown gene can be hypothesized based on the 'guilt-by-association' concept. Furthermore, common regulatory motifs can be searched. Even though meaningful biological patterns are often identified, the reliability of cluster assignment should be statistically verified [72].

Metrics - Despite holding significant promise for providing an efficient description of transcriptome data, the 'ill-defined' definition for what makes expression 'ill-defined' definition for what makes expression profiles between sets of genes similar is a significant challenge [73]. Commonly used metrics to assess gene similarity are based on distances or correlation functions. Values that make up the gene expression profiles are first defined as a series of coordinates that describe a vector. Then, typically one, of two, standard metrics - the Euclidian distance and the Pearson correlation - are used to assign similarity. The Euclidian distance measures the absolute distance between two gene expression vectors, taking in consideration both the direction and the magnitude of the vectors. The Pearson correlation measures the similarity of the directions of two gene expression vectors, being insensitive to the amplitude. Since we are often interested in grouping genes with similar expression patterns (even if they differ by a factor), and abundance levels are not truly comparable between two different genes (due to microarray design), the Pearson correlation is usually preferred. While the Pearson correlation has been shown to be a superior metric to the Euclidian distance [23,74,75], this correlation may assign an artificial high-score to patterns that are not necessarily similar [70]. Other metrics, including shrinkage based similarity [76] and the jackknife correlation [77], have also been used.

Hierarchical clustering - Once the appropriate metric is selected, a distance matrix can be calculated for all pair-wise distances among all genes. Genes can then be progressively joined based on similarity (highest similarity being equivalent to shortest distance). The method initializes by finding the two most similar genes, grouping them into a new node and updating the distance matrix to account for the average distance between both genes. The process is repeated until all genes are connected, forming a single hierarchical tree. This method is simple, reproducible, and easily visualized [23]. However, accommodation of a new gene in a cluster as the tree expands may be less and less representative of the initial cluster pattern. Therefore, as hierarchical clustering goes up towards the root of the tree, the average of all profiles in a certain cluster may become less representative of the contained profiles. As a result, the number of clusters to consider should be assessed by visual inspection, based on the similar expression profiles at the desired distance cut-off. The disadvantage of this approach is that it is sensitive to outliers [78].

Partitioning algorithms - To avoid the problems associated with artifacts that arise during hierarchical clustering, partitioning (or relocation) algorithms such as *K*-means clustering [23], self-organizing maps (SOMs) [75], Mixtures of Gaussians (MoG) [79], and super-paramagnetic [80] have been applied to cluster analysis. In general these methods first partition the transcriptional profiles into relatively homogenous groups, and then organize each of these partitioned groups into more detailed clusters containing similar expression profiles. Partition methods are fast and effective, but stochastic, meaning that initial random selection of the reference vectors may result in different clustering groups. Moreover, the tendency to yield different clusters, which consequently describe different biological stories, increases

Table 1. Comparison between data-driven and integrative approaches for identification of regulatory networks

Method	Advantages	Disadvantages
Data-Driven	- Targets single gene patterns - Identifies co-regulated genes through 'guilt-by association'	- Dead-end analysis - Difficult to separate biological noise from statistical noise
Integrative	- Uncovers hidden regulatory architecture - Combines innovative thinking across scientific disciplines	- Requires reliable and global interaction models - Requires complex and expensive equipment for data generation

with noise and an increased number of transcripts. One way to manage the ambiguity associated with different clustering algorithms is to apply several methods and search for emerging patterns [81]. Another, and perhaps more powerful, concept to more robustly and reproducibly assign genes containing similar profiles is co-occurrence (Grotkjaer *et al.*, submitted). This approach consists of running the same stochastic problem several times to select consensus clusters (or co-occurrences) as the final solution.

Clustering is relatively easy to perform; however, regardless of the algorithm selected to identify similar gene expression patterns, the data will always be sorted into organized groups. Therefore, subsequent biological validation is required to confirm the observed sorting patterns. These issues have been addressed to some extent [82-84] and new tools will continue to strengthen our ability to more effectively identify co-regulated genes by their patterns of expression.

After associating genes with one another based on a specified similarity metric through clustering, several other analyses are often performed to more completely unveil the regulatory structure of the cellular response under examination. While many techniques have been reported, we highlight three schemes. Promoter analysis can be used to identify transcription factors involved in co-regulated responses [43,52,85-88]. Second, gene localization and overrepresentation within particular sections of the chromosome can be explored (Grotkjaer *et al.*, submitted). Third, the function of genes with unknown function can be predicted based on similar expression patterns of known genes [89].

Integrating Transcriptome Data with Known Biological Information

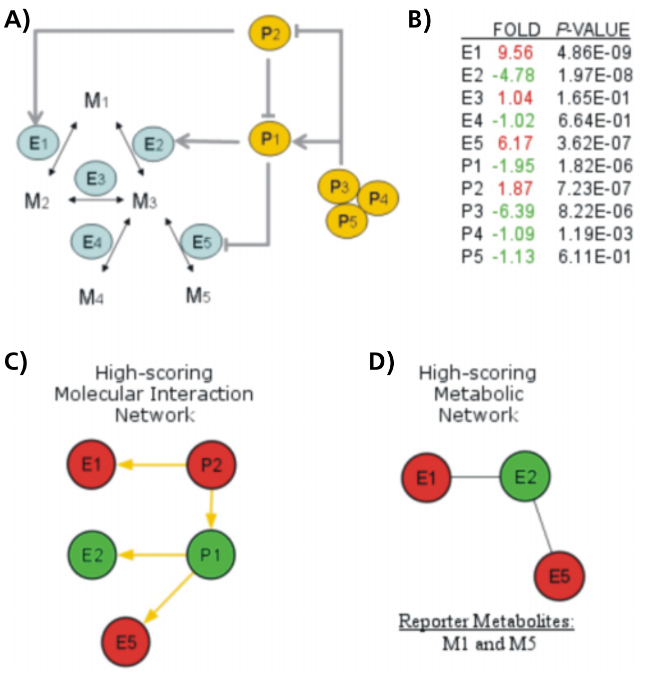
As just described, several statistical methods and clustering algorithms are available to dissect regulatory mechanisms from genome-wide expression datasets. Most of these methods are data-driven. They attempt to uncover hidden correlations in the data by using data alone. In principle, these strategies assume that transcriptionally co-regulated genes show similar expression patterns or correlation. Consequently, while searching for similar expression patterns, all mRNAs are allowed to interact with all other mRNAs in the entire solution space (transcriptome data set).

Although data-driven algorithms have been shown to enable the discovery of new and unexpected biological interactions, these methods often fail to explain the observed physiology/phenotype solely using gene-expression data [14,90,91]. From a mathematical perspective, the high degrees of freedom make these algorithms sensitive to noise in the data (either experimental or biological). Consequently, relatively weak, but biologically significant, correlations/patterns may be overshadowed by stronger but biologically insignificant and/or noisy correlations.

The natural way to overcome this problem, and facilitate identification of potentially important strain distinguishing features, is to reduce the degrees of freedom in the data-analysis by using known biological interactions, stoichiometric constraints, and/or network structure. The whole cell system can be seen as a complex web of molecular interactions (Fig. 2). Such interactions may arise from physical contact between molecules/groups of molecules (*e.g.* protein-protein, protein-DNA interactions *etc.*) or as a result of functional coupling between groups of molecules (*e.g.* genes belonging to a regulatory pathway, genes in an operon *etc.*). Here, we describe emerging strategies which seek to integrate known biological information with genome-wide transcriptional data. These approaches have not only provided insights into regulatory themes, but also facilitate a more detailed snapshot of active biomolecular pathways.

To characterize regulatory principles governing metabolic 'pathways,' several studies have explored expression patterns in the context of metabolic flux network structure. Expression patterns of genes belonging to a group of interacting components have been shown to be significantly similar relative to a set of randomly chosen genes for genes belonging to a metabolic pathway [12,32,92-95], genes belonging to a particular functional class as defined by gene ontology databases [96,97] or genes belonging to a cluster of interacting proteins [98]. Ihmels *et al.* [12] demonstrated the existence of transcriptional control at multiple levels by showing that expression is also biased towards linear pathways and that isozymes are regulated separately (preventing cross talk between different pathways). Collectively, these studies provide evidence that metabolic pathways, or at least segments associated with a specific biological function, are co-regulated. While this hypothesis is not new, it reiterates an important design principle for strain development programs.

Consistent with the concept that biological processes occur in cooperation to fulfill cellular objectives, similar studies have also pointed towards the same regulatory



	FOLD	P-VALUE
E1	9.56	4.86E-09
E2	-4.78	1.97E-08
E3	1.04	1.65E-01
E4	-1.02	6.64E-01
E5	6.17	3.62E-07
P1	-1.95	1.82E-06
P2	1.87	7.23E-07
P3	-6.39	8.22E-06
P4	-1.09	1.19E-03
P5	-1.13	6.11E-01

Fig. 5. Illustration of two methods that integrate topographic information with transcriptome data. **A)** Topography of the system, representing both the metabolic network and the molecular interactions network. M_i represents the metabolite i¸ which is converted into another metabolite through the enzyme E_j (in blue). Regulatory and sensing proteins are represented by P_k (in yellow). Black arrows indicate reactions, while grey arrows indicate activation/up-regulation and T arrows indicate repression/down-regulation. **B)** Fold changes and *P*-values calculated from a simulated experience between a reference strain and a P3 knockout mutant. Fold changes for down(up)-regulated genes are in green (red). **C)** High-scoring molecular interaction sub-network calculated for the pathway in grey (as in **A**), using Ideker *et al.* algorithm [105]. All interactions in the high-scoring sub-network do actually represent protein-DNA interactions (direction given by the yellow arrow). **D)** High-scoring metabolic sub-network calculated for the pathway in black (as in **A**), using Patil *et al.* algorithm [14], and corresponding reporter metabolites. Reporter metabolites highlight the fact that the most significant changes occurred around the metabolites *M1* and *M5*.

motifs. Zien *et al.* [93] and Pavlidis *et al.* [94] described co-regulation by characterizing pathways in the framework of gene expression patterns with a score that provides transcriptional significance of pathway under the given experimental conditions. Moreover, Shuster *et al.* [99] described flux co-regulation by combining metabolic network structure with expression data to demonstrate that sets of required enzymes which operate to satisfy mass balance constraints have similar expression profiles. Stelling *et al.* [100] extended this theory by quantitatively predicting expression patterns based on stoichiometry alone. By showing that theoretically predicted gene expression ratios in *E. coli* agree with the experimental observations, this study represents an important step towards development of the predictive metabolic models. Similar results have also been reported for *S. cerevisiae* metabolic network [101]. To aid in the development of tools to analyze transcriptome data, strategies have been developed to display or assess up/down regulation of genes in the context of metabolic networks [92,102]. Unfortunately, methods based on genome-scale metabolic models are limited to organisms where a complete and accurate roadmap of metabolism is known.

Parallel to the methods based on metabolic stoichiometry, many studies have reported that the supervised learning methods, like support vector machines [89] and supervised neural networks [96], improve functional classification based on genome-wide transcription profiles. These findings, again, support the need to complement gene expression data analysis with methods that account for underlying biological interactions.

Although the approaches described above to identify transcriptional regulation of metabolic flux networks attempt to put gene expression data in the perspective of metabolic pathways, there are two major limitations of these methods. First, they rely on the definition of metabolism as a group of pathways where each pathway is treated as a distinct entity, often borrowed from textbook definitions. This is a major limitation considering the recent studies on metabolic network topology [103,104] that show that metabolic networks are highly connected through various co-factors and small molecules, and consequently cannot be, in general, reduced into "pathways". Second, these methods help more in visualization or explaining known interactions/observations rather than generating new hypotheses.

We recently reported a novel algorithm to study of transcriptional regulation of metabolism through the integration of mRNA expression data and metabolic network topology [14]. This method reports two contributions that hold significant promise for metabolic engineering applications. First, the algorithm defines 'so-called' reporter metabolites, which might be functionally related to the perturbation affecting the system under investigation (*e.g.* genes that have been knocked out). In this manner, the global role of a metabolite is inferred from mRNA expression patterns and metabolic network topology without direct measurement of metabolite concentration (Fig. 5). Second, the algorithm identifies the most highly correlated metabolic subnetworks to postulate potential metabolic interactions that current methods may fail to capture. The authors also demonstrate that the proposed algorithm is robust towards the missing information in the metabolic models, a limitation of previous attempts. In a biological context, it appears that cells respond to perturbations by changing the expression pattern of several genes involved in the specific part(s) of the metabolism where a perturbation is introduced. Due to the highly connected metabolic network, these changes are then propagated through the system. The essence of these changes is reflected in and quantified by reporter metabolites and subnetworks identified. Since cellular response to genetic and environmental perturbations is often reflected and/or mediated through changes in the metabolism, the reporter algorithm can be effectively used not only to identify regulatory modules but also for functional genomics and cell factory design.

The pioneering work of Ideker *et al.* [105] demonstrated a significant advance for integration of molecular interaction networks with gene expression data analysis to discover transcriptional regulatory subnetworks (Fig. 5). The study showed that existing knowledge about protein-protein and protein-DNA interactions could be used to extract regulatory modules using gene expression data. Similar ideas were presented in another study by the author [13] where the yeast galactose utilization pathway was analyzed from a systems biology perspective.

In addition to methods which rely on experimental data, flux balance analysis (FBA) [106] and related approaches are being increasingly used for in silico prediction of fluxes and identifying metabolic engineering targets [107-110]. One of the major challenges for improving the applicability of such mass-balance based approaches is to account for the metabolic regulation. Availability of genome-wide gene expression data has opened new opportunities to integrate transcriptional level regulation into these models. Akesson *et al.* [111] used genome-wide transcription data to identify the genes that are not expressed in the study and used this information to impose additional on-off constraints into FBA. Interestingly, this simple approach improved the flux balance analysis predictions significantly, illustrating the need to integrate transcriptional regulation information into FBA models. Covert *et al.* [112-114] have also proposed Boolean-logic based incorporation of regulatory constraints into FBA and thus improving the quality of predictions. Such regulatory rules can be, *e.g.*, formulated/curated based on existing gene expression datasets.

Perspective

High-throughput transcription analysis offers much prospect in the field of metabolic engineering as it enables rapid screening of which genes have altered expression at different growth conditions or in different mutant strains. However, we have seen relatively few applications of genome-wide transcription analysis for identification of new targets for metabolic engineering. Clearly this technology provides valuable insight into the physiology of the cells under study and hence is on the “nice to have” list of technologies applied in the field of metabolic engineering. Since it has shown to be of limited use in directly guiding metabolic engineering, it has not penetrated the field as strongly as one could have anticipated 5~6 years ago. One major reason for this is the relatively poor correlation between gene transcription and metabolic fluxes - the latter being the focal point of most metabolic engineering exercises. Perhaps of even more importance is that even small changes in expression of a couple of genes may have significant impact on the operation of the complete metabolic network, and identification of small changes in expression of a few genes is not compatible with the high-throughput nature of the analysis. Hence, to move forward there is a requirement for development of novel methods that enable mapping of even small changes in the transcriptional level and particularly linking these to the different parts of the metabolic networks. As discussed here, there are some developments in this area and the use of model guided data analysis may be a step in the right direction, but approaches where analysis of different omes are combined and the data are integrated into one analysis is surely also a way forward. Once we start to have such integrated data and novel methods for integration of these data we are confident that ome analysis, and particularly, high-throughput transcription analysis will move from a “nice to have” tool to a “need to have” tool in metabolic engineering.

Acknowledgements We thank Thomas Grotkjaer for critical reading of the manuscript. Supported by grants from the Danish Research Agency. M.C.J. is a recipient of a NSF International Research Postdoctoral Fellowship.

REFERENCES

[1] Patil, K. R., M. Akesson, and J. Nielsen (2004) Use of genome-scale microbial models for metabolic engineering. *Curr. Opin. Biotechnol*. 15: 64-69.

[2] Nielsen, J. (2001) Metabolic engineering. *Appl. Microbiol. Biotechnol*. 55: 263-283.

[3] Stephanopoulos, G., A. Aristidou, and J. Nielsen, (1998) *Metabolic Engineering*. Academic Press, San Diego, USA.

[4] Bulter, T., J. R. Bernstein, and J. C. Liao (2003) A perspective of metabolic engineering strategies: Moving up the systems hierarchy. *Biotechnol. Bioeng*. 84: 815-821.

[5] Nielsen, J. (2003) It is all about metabolic fluxes. *J. Bacteriol.* 185: 7031-7035.

[6] Bailey, J. E., A. Sburlati, V. Hatzimanikatis, K. Lee, W. A. Renner, and P. S. Tsai (1996) Inverse metabolic engineering: A strategy for directed genetic engineering of useful phenotypes. *Biotechnol. Bioeng*. 52: 109-121.

[7] Martin, V. J., D. J. Pitera, S. T. Withers, J. D. Newman, and J. D. Keasling (2003) Engineering a mevalonate pathway in Escherichia coli for production of terpenoids. *Nat. Biotechnol*. 21: 796-802.

[8] Farmer, W. R. and J. C. Liao (2000) Improving lycopene production in *Escherichia coli* by engineering metabolic control. *Nat. Biotechnol*. 18: 533-537.

[9] Ostergaard, S., L. Olsson, M. Johnston, and J. Nielsen (2000) Increasing galactose consumption by *Saccharomyces cerevisiae* through metabolic engineering of the GAL gene regulatory network. *Nat. Biotechnol.* 18: 1283-1286.

[10] Bailey, J. E. (1999) Lessons from metabolic engineering for functional genomics and drug discovery. *Nat. Biotechnol*. 17: 616-618.

[11] Bro, C. and J. Nielsen (2004) Impact of ‘ome’ analyses on inverse metabolic engineering. *Metab. Eng*. 6: 204-211.

[12] Ihmels, J., R. Levy, and N. Barkai (2004) Principles of

transcriptional control in the metabolic network of *Saccharomyces cerevisiae*. *Nat. Biotechnol.* 22: 86-92.

[13] Ideker, T., V. Thorsson, J. A. Ranish, R. Christmas, J. Buhler, J. K. Eng, R. Bumgarner, D. R. Goodlett, R. Aebersold, and L. Hood (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* 292: 929-934.

[14] Patil, K. R. and J. Nielsen (2005) Uncovering transcriptional regulation of metabolism by using metabolic network topology. *Proc. Natl. Acad. Sci. USA* 102: 2685-2689.

[15] Price, N. D., J. A. Papin, C. H. Schilling, and B. O. Palsson (2003) Genome-scale microbial in silico models: The constraints-based approach. *Trends Biotechnol*. 21: 162-169.

[16] Burgard, A. P., P. Pharkya, and C. D. Maranas (2003) OptKnock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol. Bioeng*. 84: 647-657.

[17] Ideker, T., T. Galitski, and L. Hood (2001) A new approach to decoding life: Systems biology. *Annu. Rev. Genomics Hum. Genet*. 2: 343-372.

[18] Nielsen, J. and L. Olsson (2002) An expanded role for microbial physiology in metabolic engineering and functional genomics: Moving towards systems biology. *FEMS Yeast Res*. 2: 175-181.

[19] Weston, A. D. and L. Hood (2004) Systems biology, proteomics, and the future of health care: Toward predictive, preventative, and personalized medicine. *J. Proteome Res*. 3: 179-196.

[20] Stephanopoulos, G., H. Alper, and J. Moxley (2004) Exploiting biological complexity for strain improvement through systems biology. *Nat. Biotechnol*. 22: 1261-1267.

[21] Brent, R. (2004) A partnership between biology and engineering. *Nat. Biotechnol*. 22: 1211-1214.

[22] Hood, L. and R. M. Perlmutter (2004) The impact of systems approaches on biological problems in drug discovery. *Nat. Biotechnol*. 22: 1215-1217.

[23] Eisen, M. B., P. T. Spellman, P. O. Brown, and D. Botstein (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 95: 14863-14868.

[24] Schena, M., R. A. Heller, T. P. Theriault, K. Konrad, E. Lachenmeier, and R. W. Davis (1998) Microarrays: Biotechnology's discovery platform for functional genomics. *Trends Biotechnol*. 16: 301-306.

[25] Lipshutz, R. J., S. P. A. Fodor, T. R. Gingeras, and D. J. Lockhart (1999) High density synthetic oligonucleotide arrays. *Nat. Genetics* 21: 20-24.

[26] Hughes, T. R., M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Dai, Y. D. He, M. J. Kidd, A. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepaniants, D. D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S. H. Friend (2000) Functional discovery *via* a compendium of expression profiles. *Cell* 102: 109-126.

[27] Lynch, M. D., R. T. Gill, and G. Stephanopoulos (2004) Mapping phenotypic landscapes using DNA micro-arrays. *Metab. Eng*. 6: 177-185.

[28] Stafford, D. E. and G. Stephanopoulos (2001) Metabolic engineering as an integrating platform for strain development. *Curr. Opin. Microbiol*. 4: 336-340.

[29] Kao, C. M. (1999) Functional genomic technologies: Creating new paradigms for fundamental and applied biology. *Biotechnol. Prog*. 15: 304-311.

[30] de Lichtenberg, U., L. J. Jensen, S. Brunak, and P. Bork (2005) Dynamic complex formation during the yeast cell cycle. *Science* 307: 724-727.

[31] Laub, M. T., H. H. McAdams, T. Feldblyum, C. M. Fraser, and L. Shapiro (2000) Global analysis of the genetic network controlling a bacterial cell cycle. *Science* 290: 2144-2148.

[32] Spellman, P. T., G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* 9: 3273-3297.

[33] DeRisi, J. L., V. R. Iyer, and P. O. Brown (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278: 680-686.

[34] Gill, R. T., S. Wildt, Y. T. Yang, S. Ziesman, and G. Stephanopoulos (2002) Genome-wide screening for trait conferring genes using DNA microarrays. *Proc. Natl. Acad. Sci. USA* 99: 7033-7038.

[35] Gill, R. T. (2003) Enabling inverse metabolic engineering through genomics. *Curr. Opin. Biotechnol*. 14: 484-490.

[36] Gonzalez, R., H. Tao, J. E. Purvis, S. W. York, K. T. Shanmugam, and L. O. Ingram (2003) Gene array-based identification of changes that contribute to ethanol tolerance in ethanologenic *Escherichia coli*: Comparison of KO11 (parent) to LY01 (resistant mutant). *Biotechnol. Prog*. 19: 612-623.

[37] Lum, A. M., J. Huang, C. R. Hutchinson, and C. M. Kao (2004) Reverse engineering of industrial pharmaceutical-producing actinomycete strains using DNA microarrays. *Metab. Eng*. 6: 186-196.

[38] Wahlbom, C. F., R. R. Cordero Otero, W. H. van Zyl, B. Hahn-Hagerdal, and L. J. Jonsson (2003) Molecular analysis of a *Saccharomyces cerevisiae* mutant with improved ability to utilize xylose shows enhanced expression of proteins involved in transport, initial xylose metabolism, and the pentose phosphate pathway. *Appl. Environ. Microbiol*. 69: 740-746.

[39] Askenazi, M., E. M. Driggers, D. A. Holtzman, T. C. Norman, S. Iverson, D. P. Zimmer, M. E. Boers, P. R. Blomquist, E. J. Martinez, A. W. Monreal, T. P. Feibelman, M. E. Mayorga, M. E. Maxon, K. Sykes, J. V. Tobin, E. Cordero, S. R. Salama, J. Trueheart, J. C. Royer, and K. T. Madden (2003) Integrating transcriptional and metabolite profiles to direct the engineering of lovastatin-producing fungal strains. *Nat. Biotechnol*. 21: 150-156.

[40] Oh, M. K. and J. C. Liao (2000) DNA microarray detection of metabolic responses to protein overproduction in *Escherichia coli*. *Metab. Eng*. 2: 201-209.

[41] Sanford, K., P. Soucaille, G. Whited, and G. Chotani (2002) Genomics to fluxomics and physiomics - pathway engineering. *Curr. Opin. Microbiol*. 5: 318-322.

[42] Daran-Lapujade, P., M. L. Jansen, J. M. Daran, W. van

Gulik, J. H. de Winde, and J. T. Pronk (2004) Role of transcriptional regulation in controlling fluxes in central carbon metabolism of *Saccharomyces cerevisiae*. A chemostat culture study. *J. Biol. Chem*. 279: 9125-9138.

[43] Tummala, S. B., S. G. Junne, and E. T. Papoutsakis (2003) Antisense RNA downregulation of coenzyme A transferase combined with alcohol aldehyde dehydrogenase overexpression leads to predominantly alcohologenic *Clostridium acetobutylicum* fermentations. *J. Bacteriol.* 185: 3644-3653.

[44] Yoon, S. H., M. J. Han, S. Y. Lee, K. J. Jeong, and J. S. Yoo (2003) Combined transcriptome and proteome analysis of *Escherichia coli* during high cell density culture. *Biotechnol. Bioeng*. 81: 753-767.

[45] Griffin, T. J., S. P. Gygi, T. Ideker, B. Rist, J. Eng, L. Hood, and R. Aebersold (2002) Complementary profiling of gene expression at the transcriptome and proteome levels in *Saccharomyces cerevisiae*. *Mol. Cell. Proteomics* 1: 323-333.

[46] Schena, M., D. Shalon, R. W. Davis, and P. O. Brown (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270: 467-470.

[47] Harrington, C. A., C. Rosenow, and J. Retief (2000) Monitoring gene expression using DNA microarrays. *Curr. Opin. Microbiol*. 3: 285-291.

[48] Lockhart, D. J. and E. A. Winzeler (2000) Genomics, gene expression and DNA arrays. *Nature* 405: 827-836.

[49] Knudsen, S. (2004) *Guide to Analysis of DNA Microarray Data.* John Wiley & Sons, Inc., Hoboken, NJ, USA.

[50] Parada, G. and F. Acevedo (1983) On the relation of temperature and RNA content to the specific growth rate in *Saccharomyces cerevisiae*. *Biotechnol. Bioeng.* 25: 2785-2788.

[51] Waldron, C. and F. Lacroute (1975) Effect of growth rate on the amounts of ribosomal and transfer ribonucleic acids in yeast. *J. Bacteriol.* 122: 855-865.

[52] Gasch, A. P., P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein, and P. O. Brown (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* 11: 4241-4257.

[53] Hayes, A., N. Zhang, J. Wu, P. R. Butler, N. C. Hauser, J. D. Hoheisel, F. L. Lim, A. D. Sharrocks, and S. G. Oliver (2002) Hybridization array technology coupled with chemostat culture: Tools to interrogate gene expression in *Saccharomyces cerevisiae*. *Methods* 26: 281-290.

[54] Leung, Y. F. and D. Cavalieri (2003) Fundamentals of cDNA microarray data analysis. *Trends Genet*. 19: 649-659.

[55] Brazma, A., P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, T. Gaasterland, P. Glenisson, F. C. Holstege, I. F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet*. 29: 365-371.

[56] Quackenbush, J. (2001) Computational analysis of microarray data. *Nat. Rev. Genetics* 2: 418-427.

[57] Schadt, E. E., C. Li, C. Su, and W. H. Wong (2000) Analyzing high-density oligonucleotide gene expression array data. *J. Cell. Biochem.* 80: 192-202.

[58] Workman, C., L. Jensen, H. Jarmer, R. Berka, L. Gautier, H. Nielser, H. H. Saxild, C. Nielsen, S. Brunak, and S. Knudsen (2002) A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biology* 3: research0048.

[59] Irizarry, R. A., B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4: 249-264.

[60] Li, C. and W. H. Wong (2001) Model-based analysis of oligonucleotide arrays: Model validation, design issues and standard error application. *Genome Biology* 2: research0032.

[61] Li, C. and W. H. Wong (2001) Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl. Acad. Sci. USA* 98: 31-36.

[62] Zhou, Y. and R. Abagyan (2002) Match-only integral distribution (MOID) algorithm for high-density oligonucleotide array analysis. *BMC Bioinformatics* 3: 3.

[63] Naef, F., D. A. Lim, N. Patil, and M. Magnasco (2002) DNA hybridization to mismatched templates: A chip study. *Phys. Rev. E. Stat. Nonlin. Soft Matter Phys*. 65: 040902.

[64] Chudin, E., R. Walker, A. Kosaka, S. X. Wu, D. Rabert, T. K. Chang, and D. E. Kreder (2002) Assessment of the relationship between signal intensities and transcript concentration for Affymetrix GeneChip arrays. *Genome Biol*. 3: Research0005.

[65] Ideker, T., V. Thorsson, A. F. Siegel, and L. E. Hood (2000) Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *J. Comput. Biol.* 7: 805-817.

[66] Storey, J. D. and R. Tibshirani (2003) Statistical significance for genome wide studies. *Proc. Natl. Acad. Sci. USA* 100: 9440-9445.

[67] Taguchi, Y. H. and Y. Oono (2005) Relational patterns of gene expression *via* non-metric multidimensional scaling analysis. *Bioinformatics* 21: 730-740.

[68] Yeung, K. Y. and W. L. Ruzzo (2001) Principal component analysis for clustering gene expression data. *Bioinformatics* 17: 763-774.

[69] Alter, O., P. O. Brown, and D. Botstein (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA* 97: 10101-10106.

[70] Sherlock, G. (2000) Analysis of large-scale gene expression data. *Curr. Opin. Immunol*. 12: 201-205.

[71] Valafar, F. (2002) Pattern recognition techniques in microarray data analysis: A survey. *Ann. NY Acad. Sci*. 980: 41-64.

[72] Dharmadi, Y. and R. Gonzalez (2004) DNA microarrays: Experimental issues, data analysis, and application to bacterial systems. *Biotechnol. Prog*. 20: 1309-1324.

[73] Grotkjaer, T. and J. Nielsen (2004) Enhancing yeast

transcription analysis through integration of heterogeneous data. *Curr. Genomics* 5: 673-686.

[74] Gibbons, F. D. and F. P. Roth (2002) Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Res.* 12: 1574-1581.

[75] Tamayo, P., D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub (1999) Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA* 96: 2907-2912.

[76] Cherepinsky, V., J. Feng, M. Rejali, and B. Mishra (2003) Shrinkage-based similarity metric for cluster analysis of microarray data. *Proc. Natl. Acad. Sci. USA* 100: 9668-9673.

[77] Heyer, L. J., S. Kruglyak, and S. Yooseph (1999) Exploring expression data: Identification and analysis of coexpressed genes. *Genome Res*. 9: 1106-1115.

[78] Hastie, T., R. Tibshirani, and J. Friedman, (2001) *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*. Springer-Verlag, New York, NY, USA.

[79] MacKay, D. J. C. (2003) *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, Cambridge, UK.

[80] Blatt, M., S. Wiseman, and E. Domany (1996) Superparamagnetic clustering of data. *Phys. Rev. Lett.* 76: 3251-3254.

[81] Kaminski, N. and N. Friedman (2002) Practical approaches to analyzing results of microarray experiments. *Am. J. Respir. Cell Mol. Biol*. 27: 125-132.

[82] Kerr, M. K. and G. A. Churchill (2001) Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments. *Proc. Natl. Acad. Sci. USA* 98: 8961-8965.

[83] McShane, L. M., M. D. Radmacher, B. Freidlin, R. Yu, M. C. Li, and R. Simon (2002) Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. *Bioinformatics* 18: 1462-1469.

[84] Zhang, K. and H. Zhao (2000) Assessing reliability of gene clusters from gene expression data. *Funct. Integr. Genomics* 1: 156-173.

[85] Zhu, J. and M. Q. Zhang (2000) Cluster, function and promoter: Analysis of yeast expression array. *Pac. Symp. Biocomput.* 479-490.

[86] Wei, G. H., D. P. Liu, and C. C. Liang (2004) Charting gene regulatory networks: Strategies, challenges and perspectives. *Biochem. J.* 381: 1-12.

[87] Pilpel, Y., P. Sudarsanam, and G. M. Church (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet*. 29: 153-159.

[88] Banerjee, N. and M. Q. Zhang (2002) Functional genomics as applied to mapping transcription regulatory networks. *Curr. Opin. Microbiol.* 5: 313-317.

[89] Brown, M. P., W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares, Jr., and D. Haussler (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. USA* 97: 262-267.

[90] Miki, R., K. Kadota, H. Bono, Y. Mizuno, Y. Tomaru, P. Carninci, M. Itoh, K. Shibata, J. Kawai, H. Konno, S. Watanabe, K. Sato, Y. Tokusumi, N. Kikuchi, Y. Ishii, Y. Hamaguchi, I. Nishizuka, H. Goto, H. Nitanda, S. Satomi, A. Yoshiki, M. Kusakabe, J. L. DeRisi, M. B. Eisen, V. R. Iyer, P. O. Brown, M. Muramatsu, H. Shimada, Y. Okazaki, and Y. Hayashizaki (2001) Delineating developmental and metabolic pathways *in vivo* by expression profiling using the RIKEN set of 18,816 full-length enriched mouse cDNA arrays. *Proc. Natl. Acad. Sci.* 98: 2199-2204.

[91] Bro, C., B. Regenberg, and J. Nielsen (2004) Genome-wide transcriptional response of a Saccharomyces cerevisiae strain with an altered redox metabolism. *Biotechnol. Bioeng*. 85: 269-276.

[92] Grosu, P., J. P. Townsend, D. L. Hartl, and D. Cavalieri (2002) Pathway Processor: A tool for integrating whole-genome expression results into metabolic networks. *Genome Res*. 12: 1121-1126.

[93] Zien, A., R. Kuffner, R. Zimmer, and T. Lengauer (2000) Analysis of gene expression data with pathway scores. *Proc. Int. Conf. Intell. Syst. Mol. Biol*. 8: 407-417.

[94] Pavlidis, P., D. P. Lewis, and W. S. Noble (2002) Exploring gene expression data with class scores. *Pac. Symp. Biocomput*. 474-485.

[95] Nakao, M., H. Bono, S. Kawashima, T. Kamiya, K. Sato, S. Goto, and M. Kanehisa (1999) Genome-scale gene expression analysis and pathway reconstruction in KEGG. *Genome Inform. Ser. Workshop Genome Inform*. 10: 94-103.

[96] Mateos, A., J. Dopazo, R. Jansen, Y. Tu, M. Gerstein, and G. Stolovitzky (2002) Systematic learning of gene functional classes from DNA array expression data by using multilayer perceptrons. *Genome Res*. 12: 1703-1715.

[97] Breitling, R., A. Amtmann, and P. Herzyk (2004) Graph-based iterative Group Analysis enhances microarray interpretation. *BMC Bioinformatics*. 5: 100.

[98] Jansen, R., D. Greenbaum, and M. Gerstein (2002) Relating whole-genome expression data with protein-protein interactions. *Genome Res.* 12: 37-46.

[99] Schuster, S., D. A. Fell, and T. Dandekar (2000) A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat. Biotechnol*. 18: 326-332.

[100] Stelling, J., S. Klamt, K. Bettenbrock, S. Schuster, and E. D. Gilles (2002) Metabolic network structure determines key aspects of functionality and regulation. *Nature* 420: 190-193.

[101] Cakir, T., B. Kirdar, and K. O. Ulgen (2004) Metabolic pathway analysis of yeast strengthens the bridge between transcriptomics and metabolic networks. *Biotechnol. Bioeng.* 86: 251-260.

[102] Pandey, R., R. K. Guru, and D. W. Mount (2004) Pathway miner: Extracting gene association networks from molecular pathways for predicting the biological significance of gene expression microarray data. *Bioinformatics* 20: 2156-2158.

[103] Jeong, H., B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabasi (2000) The large-scale organization of metabolic networks. *Nature* 407: 651-654.

[104] Fell, D. A. and A. Wagner (2000) The small world of metabolism. *Nat. Biotechnol*. 18: 1121-1122.

[105] Ideker, T., O. Ozier, B. Schwikowski, and A. F. Siegel (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* 18 S233-S240.

[106] Majewski, R. A. and M. M. Domach (1990) Simple constrained-optimization view of acetate overflow in *E. coli*. *Biotechnol. Bioeng*. 35: 732-738.

[107] Burgard, A. P. and C. D. Maranas (2003) Optimization-based framework for inferring and testing hypothesized metabolic objective functions. *Biotechnol. Bioeng*. 82: 670-677.

[108] Burgard, A. P., E. V. Nikolaev, C. H. Schilling, and C. D. Maranas (2004) Flux coupling analysis of genome-scale metabolic network reconstructions. *Genome Res*. 14: 301-312.

[109] Pharkya, P., A. P. Burgard, and C. D. Maranas (2003) Exploring the overproduction of amino acids using the bilevel optimization framework OptKnock. *Biotechnol. Bioeng*. 84: 887-899.

[110] Segre, D., D. Vitkup, and G. M. Church (2002) Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci. USA* 99: 15112-15117.

[111] Akesson, M., J. Forster, and J. Nielsen (2004) Integration of gene expression data into genome-scale metabolic models. *Metab. Eng.* 6: 285-293.

[112] Covert, M. W., E. M. Knight, J. L. Reed, M. J. Herrgard, and B. O. Palsson (2004) Integrating high-throughput and computational data elucidates bacterial networks. *Nature* 429: 92-96.

[113] Covert, M. W. and B. O. Palsson (2003) Constraints-based models: regulation of gene expression reduces the steady-state solution space. *J. Theor. Biol*. 221: 309-325.

[114] Covert, M. W., C. H. Schilling, and B. Palsson (2001) Regulation of gene expression in flux balance models of metabolism. *J. Theor. Biol*. 213: 73-88.

[Received April 15, 2005; accepted August 12, 2005]