

Systems biotechnology for strain improvement

Sang Yup Lee^{1,2}, Dong-Yup Lee^{1,2} and Tae Yong Kim¹

¹Metabolic and Biomolecular Engineering National Research Laboratory and Department of Chemical and Biomolecular Engineering, Korea Advanced Institute of Science and Technology, 373-1 Guseong-dong, Yuseong-gu, Daejeon 305-701, Korea
²Department of BioSystems and Bioinformatics Research Center, Korea Advanced Institute of Science and Technology, 373-1 Guseong-dong, Yuseong-gu, Daejeon 305-701, Korea

Various high-throughput experimental techniques are routinely used for generating large amounts of omics data. In parallel, *in silico* modelling and simulation approaches are being developed for quantitatively analyzing cellular metabolism at the systems level. Thus informative high-throughput analysis and predictive computational modelling or simulation can be combined to generate new knowledge through iterative modification of an *in silico* model and experimental design. On the basis of such global cellular information we can design cells that have improved metabolic properties for industrial applications. This article highlights the recent developments in these systems approaches, which we call systems biotechnology, and discusses future prospects.

Introduction

The indispensable role of biotechnology is increasing in nearly every industry, including the healthcare, pharmaceutical, chemical, food and agricultural industries. Biotechnological production of small-volume high-value drugs, chemicals and bioproducts is well justified economically. However, production of large-volume low-value bioproducts requires the development of lower-cost and higher-yield processes. Towards this goal, improved microorganisms have traditionally been developed through random mutagenesis followed by intelligent screening processes [1]. Rational metabolic and cellular engineering approaches have also been successful in improving strain performance in several cases; however, such attempts were limited to the manipulation of only a handful of genes encoding enzymes and regulatory proteins selected using available information and research experience.

Recent advances in high-throughput experimental techniques supported by bioinformatics have resulted in rapid accumulation of a wide range of omics data at various levels (Figure 1), thus providing a foundation for in-depth understanding of biological processes [2–4]. Even though our ability to analyze these x-omic (see Glossary) data in a truly integrated manner is currently limited, new targets for strain improvement can be identified from these global data [5,6]. More recently, several examples of

combined analysis of these x-omic data towards the development of improved strains have been reported [7]. Along with these high-throughput experimental techniques, *in silico* modelling and simulation are providing powerful solutions for deciphering the functions and characteristics of biological systems [8–11]. These *in silico* experiments would elevate our capability for understanding and predicting the cellular behaviour of microorganisms under any perturbations (e.g. genetic modifications and/or environmental changes) on a global scale [12].

Glossary

Bilevel optimization: A traditional mathematical programming problem maximizes a single objective function over a set of feasible solutions. Bilevel optimization seeks to maximize two objective functions simultaneously over a set of feasible solutions.

Flux: The production or consumption of mass (metabolite) per unit area per unit time. It is, however, often used on the basis of unit cell mass rather than unit area in metabolic flux analysis.

High cell density culture: High cell density culture is used to increase the cell concentration usually by well-controlled fed-batch culture. Because the volumetric productivity (product formed per unit volume per unit time) is a multiplication of the specific productivity (product formed per unit cell mass per unit time) by cell concentration, high cell density culture generally increases the volumetric productivity by the increase in the cell concentration.

IGF-I: IGF-I, insulin-like growth factor I, is a polypeptide produced primarily in the liver that causes hypoglycemia and a transient decrease in free serum fatty acids. It influences cellular differentiation and stimulates collagen and matrix synthesis by bone cells.

Isotopomer: An isotopomer is specified by the number of ¹³C atoms in specific positions of the molecule composed of n carbon atoms.

Leptin: A protein produced in adipocyte (fat cell) tissue, which acts as a signal to the brain. It is being considered for treating obesity.

Linear programming: A mathematical technique that finds the maximum or minimum of linear functions in many variables subject to constraints.

Metabolic control analysis: A phenomenological quantitative sensitivity analysis of fluxes and metabolite concentrations.

Metabolic flux analysis: The calculation and analysis of the flux distribution of the biochemical reaction network.

Mixed-integer optimization: Used to solve global optimization problems with piece-wise linear objective and constraints, using special ordered sets.

Quadratic programming: Used for optimization problems in which the objective function is a convex quadratic and the constraints are linear.

X-omics: Any omics studies currently being carried out, including genomics, transcriptomics, proteomics, metabolomics, fluxomics, regulomics, signalomics, physiomics and so on.

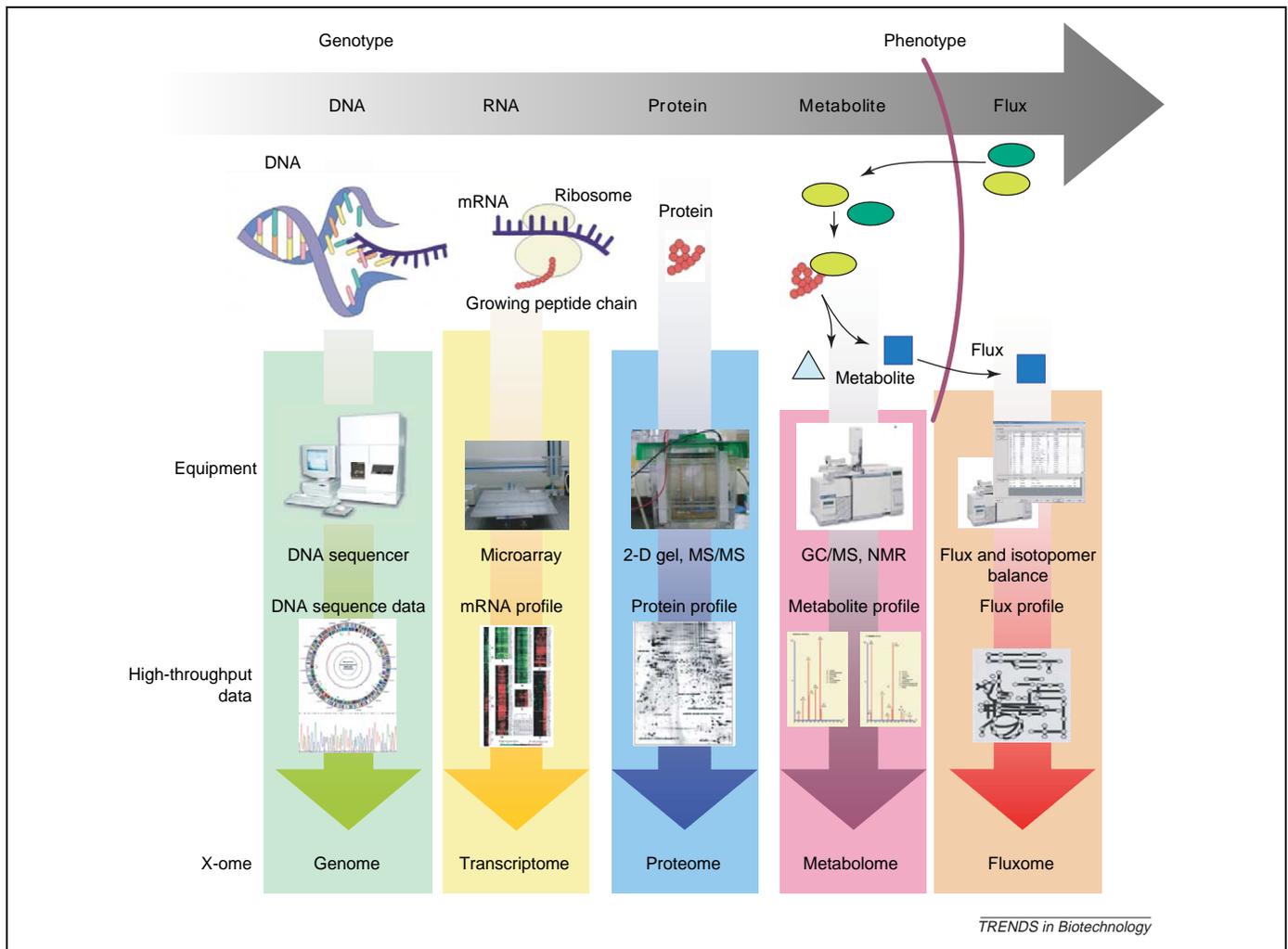
Genome: All of the genetic information and material possessed by an organism; the entire genetic complement of an organism.

Transcriptome: The full complement of mRNAs transcribed from a cell's genome.

Proteome: The complete profile of proteins expressed in a given tissue, cell or biological system at a given time.

Metabolome: The whole set of metabolites in a cell, tissue, organ, organism and species.

Fluxome: The whole set of fluxes that are measured or calculated for a given metabolic reaction network.



TRENDS in Biotechnology

Figure 1. High-throughput omics research. Genomics advanced by the development of high-speed DNA sequencing is now accompanied by transcriptome profiling using DNA microarrays. Proteome profiling is joining the high-throughput race as 2D-gel electrophoresis combined with mass spectrography is advancing. Metabolome profiling is also rapidly advancing with the development of better GC/MS, LC/MS and NMR technologies. Isotopomer profiling followed by challenging with isotopically labeled substrate allows determination of flux profiles in the cell (fluxome).

Consequently, systems-level engineering of microorganisms can be achieved by integrating high-throughput experiments and *in silico* experiments (Figure 2). The results of genomic, transcriptomic, proteomic, metabolomic and fluxomic studies, the data available in databases, and those predicted by computational modelling and simulation, are considered together within the global context of the metabolic system. This gives rise to new knowledge that can facilitate development of strains that are efficient and productive enough to be suitable for industrial applications.

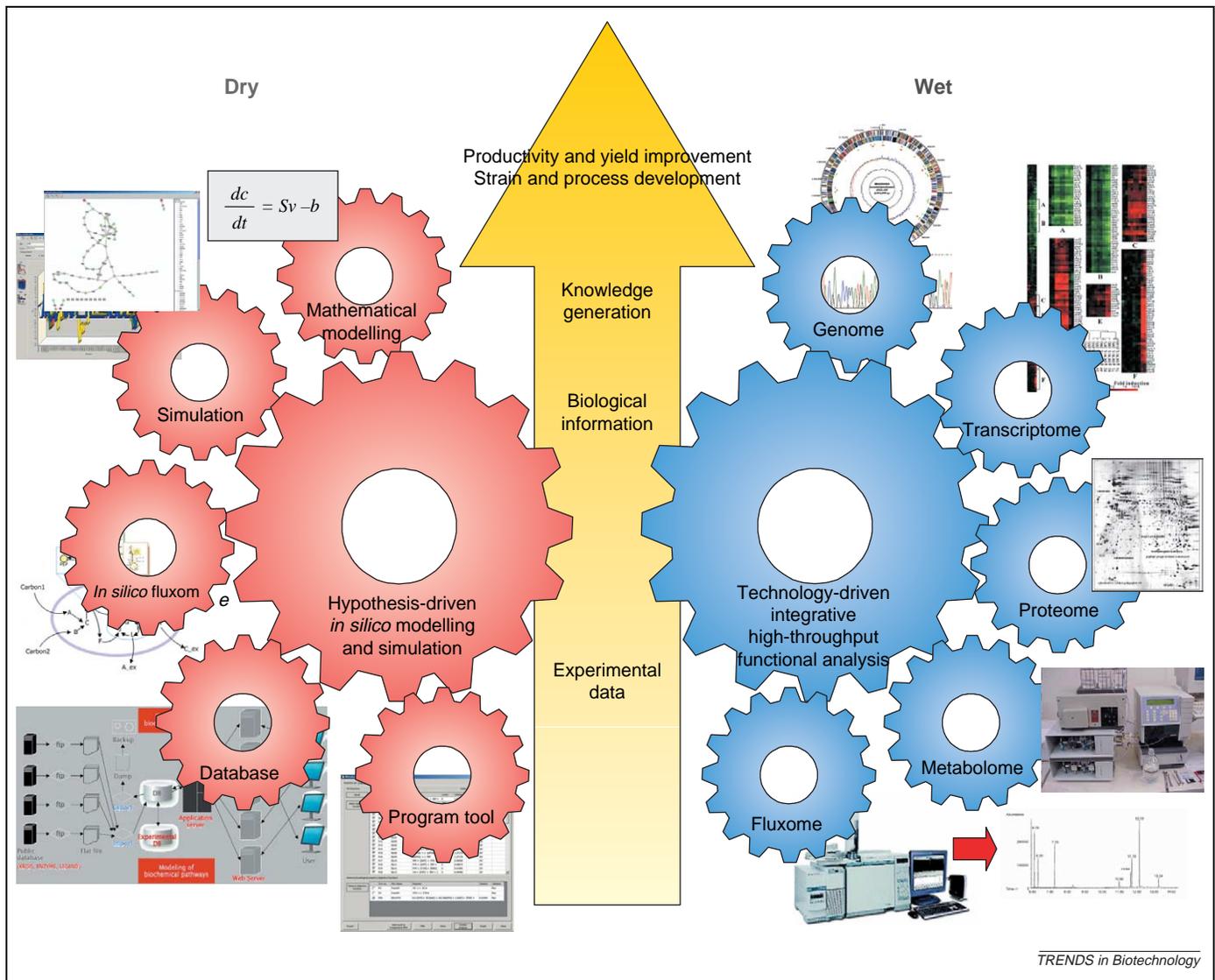
High-throughput x-omic analyses for strain improvement

As DNA sequencing has become faster and cheaper the genome sequences of many microorganisms have been completed and many more are in progress. With the complete genome sequences in our hands, post-genomic research (the 'omics' fields) is increasing rapidly. Transcriptomics allows massively parallel analysis of mRNA expression levels using DNA microarrays. Proteomics allows analysis of the protein complement of the cell or its parts by using two-dimensional gel electrophoresis

(2DGE) or chromatography coupled with various mass spectrometry methods. Metabolomics enables quantitative profiling of metabolites and metabolic intermediates using chromatography coupled with mass spectrometry or NMR. Fluxomics allows determination of metabolic fluxes based on metabolite balancing and/or isotopomer analysis. In this section, the strategies for strain improvement using these omics studies are described and representative examples are reviewed (Table 1).

Genome analysis

Comparative analysis of genomes is a relatively simple yet powerful way of identifying the genes that need to be introduced, deleted and/or modified to achieve a desired metabolic phenotype. Genomes of various organisms can be compared, as can wild-type and mutant and/or engineered strains. In one approach, a minimal strain can be designed by deleting unnecessary genes while retaining the essential genes that most effectively use metabolic functions for cell survival and production of specific bioproducts without genomic and metabolic burdens [13]. However, the concept of generating a minimal strain should be taken cautiously. The minimal



TRENDS in Biotechnology

Figure 2. Systemic integration of wet and *in silico* experiments for developing improved microorganisms for industrial applications in systems biotechnology. High-throughput experiments lead to the accumulation of large amounts of x-omes (i.e. genome, transcriptome, proteome, metabolome and fluxome data), which can be analyzed in conjunction with *in silico* modelling and simulation results, facilitating the strain improvement through the systems biotechnological research cycle (Figure 3).

strain, even after it is successfully developed, can easily become less robust owing to the deletion of many genes, some of which become important under particular culture conditions. Engineering of microorganisms based on comparative genomics has recently been successfully demonstrated. Ohnishi *et al.* [14] compared the genome sequence of a lysine-overproducing *Corynebacterium* strain with that of the wild-type strain to identify genes with point mutations that might be beneficial for the overproduction of L-lysine. Given that the genome of a particular microorganism tells us what this cell potentially can (and cannot) do, it is the starting point for engineering metabolic pathways. Even though we are currently unable to truly engineer a microorganism at the genome scale we can still benefit from engineering of local reactions and pathways that can often lead to significantly improved performance of a microorganism. Local targeted engineering based on global information is currently the most appropriate strategy exploiting the benefits of the x-omics revolution. Another important advantage

of having complete genome sequences is that genome-scale *in silico* metabolic models can now be developed that can be used to rapidly evaluate the metabolic characteristics, generate hypotheses and suggest possible engineering strategies.

Transcriptome analysis

Development of high-density DNA microarrays has changed examinations of gene transcription by allowing the simultaneous monitoring of relative mRNA abundance in multiple samples. By comparing transcriptome profiles between different strains or between the samples obtained at different time points and/or under different culture conditions, possible regulatory circuits and potential target genes to be manipulated can be identified. The new information and knowledge generated in this way can be used to engineer the local metabolic pathways for improving the performance of microorganisms.

Transcriptome profiles of recombinant *E. coli* producing human insulin-like growth factor I fusion protein (IGF-I_F)

Table 1. Examples of high-throughput x-omic analyses used for strain characterization and improvement

Omics	Strain	Product	Description	Refs
Genome	<i>Corynebacterium glutamicum</i>	Methionine	Comparative analysis of wild-type and recombinant strains for identifying gene targets	[42]
	<i>Corynebacterium glutamicum</i>	L-lysine	Comparative analysis of wild-type and mutant strains for point mutations	[14]
Transcriptome	<i>Corynebacterium glutamicum</i>	L-lysine	Comparative analysis of wild-type and mutant strains for finding production temperature	[43]
	<i>Escherichia coli</i>	IGF-I fusion protein	Comparative analysis of wild-type and recombinant strains for identifying gene targets and amplifying them	[15]
	<i>Clostridium acetobutylicum</i>	Acetone-Butanol-Ethanol	Comparative analysis of wild-type and recombinant strains for understanding regulatory mechanisms	[44]
Proteome	<i>Escherichia coli</i>	Poly(3-hydroxybutyrate)	Comparative analysis of wild-type and recombinant strains for understanding genotypic characteristics	[16]
	<i>Escherichia coli</i>	Poly(3-hydroxybutyrate)	Comparative analysis of wild-type and recombinant strains for identifying and amplifying gene targets	[17]
	<i>Escherichia coli</i>	Human leptin	Comparative analysis of wild-type and recombinant strains for understanding genotypic characteristics	[45]
Metabolome	<i>Saccharomyces cerevisiae</i>	–	Comparative analysis of wild-type and recombinant strains for understanding phenotypic behavior by aerobic shifting	[40]
	<i>Corynebacterium glutamicum</i>	L-lysine	Finding optimal condition of mutant strain	[19]
	<i>Bacillus subtilis</i>	Riboflavin	Understanding phenotypic behavior of recombinant strains by changing carbon source	[46]
Transcriptome and/or proteome	<i>Escherichia coli</i>	Cells	Identifying gene targets and understanding phenotypic behavior of wild type strain under high cell density culture	[21]
Transcriptome and/or proteome	<i>Escherichia coli</i>	L-threonine	Comparative analysis of wild-type and recombinant strains for understanding regulatory mechanisms	[22]
Transcriptome and/or metabolome	<i>Aspergillus terreus</i>	Lovastatin	Comparative analysis of wild-type and recombinant strains for identifying gene targets	[23]
Transcriptome and/or metabolome	<i>Corynebacterium glutamicum</i>	L-threonine	Understanding regulatory mechanisms of mutant strain	[24]

by high cell-density culture (HCDC) were analyzed [15]. Among the ~200 genes that were down-regulated after induction, those involved in amino acid and/or nucleotide biosynthetic pathways were selected as the first targets to be manipulated. This was because the expression of these genes is down-regulated during the HCDC of *E. coli*. Amplification of two of these genes, the *prsA* and *glpF* genes, encoding the phosphoribosyl pyrophosphate synthetase and glycerol transporter, respectively, allowed a significant increase in IFG-I_F production (from 1.8 to 4.3 g/L). This demonstrates that the strategy of 'local (targeted) engineering based on global information' allows development of a superior strain by suggesting target genes that would otherwise be difficult to identify.

Proteome analysis

Considering that most cellular metabolic activities are directly or indirectly mediated by proteins, proteome profiling takes us one step further towards understanding cellular metabolic status. However, it should be noted that not all the protein spots have been identified yet, and therefore information obtainable from the proteome is less than that from transcriptome. Nonetheless, proteome analysis can be a powerful tool when comparative profiling is carried out; one can identify protein spots that show altered intensities under two or more genetically or environmentally different conditions for further analysis and manipulation. For example, the proteome of metabolically engineered *E. coli* XL1-Blue intracellularly accumulating a biodegradable polymer poly(3-hydroxybutyrate) was compared with that of control *E. coli* strain,

thus generating new knowledge of the importance of Eda (2-keto-3-deoxy-6-phosphogluconate aldolase) in poly(3-hydroxybutyrate) production by engineered *E. coli* [16].

In another example, the proteomes of recombinant *E. coli* overproducing human leptin were examined [17]. Interestingly, the expression levels of some enzymes in the serine amino acids biosynthetic pathway decreased significantly, indicating possible limitation of serine family amino acids. This was reasonable as the serine content of leptin is 11.6%, which is much higher than the average serine content of *E. coli* proteins (5.6%). Therefore, one of the down-regulated enzymes, cysteine synthase A (encoded by *cysK*), was selected for amplification. The co-expression of the *cysK* gene led to two- and fourfold increases in cell growth and leptin productivity, respectively. In addition, *cysK* co-expression could improve production of another serine-rich protein, interleukin-12 β chain (serine content of 11.1%), suggesting that this strategy might also be useful for the production of other serine-rich proteins. These examples demonstrate that even the limited information obtained by proteome profiling can successfully lead to designing new strategies for strain improvement.

Metabolome and fluxome analysis

High-throughput quantitative analysis of metabolites has become possible as increasingly sophisticated NMR, gas chromatography mass spectrometry (GC-MS), gas chromatography time-of-flight mass spectrometry (GC-TOF) and liquid chromatography-mass spectrometry (LC-MS) procedures have been developed. Comparative analysis of

metabolite profiles under genetic and environmental perturbations makes it possible to analyse the physiological states of cells. In general, the number of metabolites in the cell is far fewer than the number of genes. For example, the number of low molecular mass metabolites in *Saccharomyces cerevisiae* was estimated to be 560 [18], which is less than one-tenth of the number of genes. However, there might be many more metabolites that are still unknown to us or difficult to detect. Furthermore, some metabolites that are predicted to exist in genome-wide metabolic reaction networks might be difficult to detect owing to the lack of suitable techniques. The heterogeneous chemistry of different metabolites and availability of only a limited number of chemicals that can be used as standards are making true whole-cell metabolome profiling far from realization. Nonetheless, several good examples of using metabolome profiling for strain improvement can be highlighted. One is the integrated analysis of metabolome and transcriptome to improve the yield of lovastatin. Another is the use of metabolome profiling to determine flux distribution in *Corynebacterium glutamicum* [19]. Given that metabolome data can be analyzed together with the fluxome data, metabolome profiling will become an increasingly popular tool in systems biotechnological research [20].

Fluxome analysis – metabolic flux profiles of a cell – takes us one step further towards understanding cellular metabolic status. Because intracellular fluxes are difficult to measure, they are often obtained by computational methods. During the calculation of fluxes, some (although limited amounts of) real experimental data, such as substrate uptake and product excretion rates, are often provided as constraints to make the calculated fluxes more realistic. Isotopomer experiments provide us with additional information on the intracellular fluxes. A frequently used substrate is ^{13}C -labelled glucose, either labeled uniformly or at the specific carbon atom only. As the ^{13}C -labeled substrate is metabolized in the cell, isotopomer distribution can be obtained and used to decipher intracellular flux ratios [19].

Combined omics analysis

True integration of all x-omic data is still far from reality, but several successful examples of strain improvement by taking combined approaches are available (Table 1). High cell-density culture (HCDC) is often used to increase the concentration and productivity of a desired product such as recombinant protein. Even though the volumetric productivity ($\text{g/L}^{-\text{h}}$) of recombinant protein can be increased by HCDC, it is frequently observed that the specific productivity ($\text{g/g DCW}^{-\text{h}}$) decreases as cell density increases. The specific reasons for this phenomenon have been unknown. We recently reported the results of combined transcriptome and proteome analyses during the HCDC of *E. coli* [21]. The most important finding was that the expression of most of amino acid biosynthesis genes was down-regulated as cell density increased. This finding immediately answers why the specific productivity of recombinant protein is reduced during the HCDC. Therefore, an important metabolic engineering strategy can be

suggested for the production of recombinant proteins by monitoring the expression levels of amino acid biosynthesis genes during the HCDC, and particularly before and after induction.

Another integrated analysis of transcriptome and proteome profiles was carried out for *E. coli* W3110 and its L-threonine-overproducing mutant strain [22]. Among the 54 genes showing meaningful differential gene expression profiles, those involved in glyoxylate shunt, the tricarboxylic acid (TCA) cycle and amino acid biosynthesis were significantly up-regulated whereas ribosomal protein genes were down-regulated. In addition, mutation in the *thrA* and *ilvA* genes was suggested to have affected overproduction of L-threonine. This combined analysis provided valuable information regarding the regulatory mechanism of L-threonine production and the physiological changes in the mutant strain.

Another interesting paper describes the use of combined analysis of transcriptome and metabolome to develop an *Aspergillus* strain overproducing lovastatin, a cholesterol-lowering drug [23]. Improved lovastatin production was initiated by generating a library of strains by expressing the genes thought to be involved in lovastatin synthesis or known to broadly affect secondary metabolite production in the parental strain. These strains were characterized by metabolome and transcriptome profiling, followed by a statistical association analysis to extract potential key parameters affecting the production of lovastatin and (+)-geodin. Using this approach, the target genes were identified and manipulated to improve lovastatin production by >50%.

More recently, Krömer *et al.* performed combined transcriptome, metabolome and fluxome analysis of L-lysine producing *C. glutamicum* at different stages of batch culture [24]. A decrease in glucose uptake rate resulted in the shift of cellular activities from growth to L-lysine production, redirecting the metabolic fluxes from the TCA cycle towards anaplerotic carboxylation and lysine biosynthesis. During this shift, the intracellular metabolite pools exhibited transient dynamics, including an increase of L-lysine up to 40 mM before its excretion to the medium. The expression levels of most genes involved in L-lysine biosynthesis remained constant whereas the metabolic fluxes showed marked changes, suggesting that metabolic fluxes are strongly regulated at the metabolic level. These are good examples of associating gene expression profile with metabolite formation to enable identification of key genes to be manipulated for improving the strain.

In silico modelling and simulation

In addition to high-throughput experimental analysis, *in silico* modelling and simulation are important aspects of systems biotechnology. The effects of genetic and/or environmental perturbations on cellular metabolism can be predicted by various *in silico* modelling and simulation approaches. The results of *in silico* analysis can then be used to design strategies for strain improvement. Detailed reviews on modelling and simulation of metabolic pathways are available elsewhere [11,25].

Construction of an *in silico* model

As omics data are accumulating at unprecedentedly high rates, various databases are being developed, updated and improved for handling different data and information ranging from simple nucleotide sequences to complex pathways (Table 2). The availability of complete genome sequences for several microorganisms is enabling the development of genome-scale *in silico* metabolic models [12,26]. These models are mainly static stoichiometric models based on the steady state of a system by excluding the time-dependent characteristics of variables. Dynamic models, which give more accurate pictures of metabolic and regulatory behavior, are currently limited by the lack of kinetic data. Nonetheless, a large effort is being made to develop such dynamic models, as represented by the E-Cell system [27]. The latest version of the E-Cell system (version 3) allows the user to perform multi-algorithm calculations by incorporating both deterministic and stochastic models. Until more sophisticated dynamic genome-scale models are developed, we can still use static stoichiometric models for whole-cell simulations, the results of which can be used for strain improvement. Currently, genome-scale metabolic models are available for *E. coli*, *Geobacter sulfurreducens*, *Haemophilus influenzae*, *Helicobacter pylori*, *Mannheimia succiniciproducens* and *Saccharomyces cerevisiae* (Table 3), and are expected to be extended to include more organisms in the near future.

In silico metabolic simulation

When the most plausible model has been constructed, *in silico* experiments can quantify flux distribution and predict phenotypic behaviour under various conditions. Furthermore, possible targets for genetic modification to improve the strain performance can be identified through comparative studies under genetically and environmentally perturbed conditions. For example, on the basis of constraints-based flux analysis, gene knockout targets of recombinant *E. coli* can be identified by means of linear programming [26], mixed-integer optimization [28], quadratic programming [29] and bilevel optimization [30].

Several tools aiming to model and simulate metabolic reaction networks have been developed. Gepasi has been one of the most widely used programs for dynamic simulation and metabolic control analysis [31]. DBsolve can handle both ordinary differential equations and non-linear algebraic equations with improved numerical solution algorithms [32]. Jarnac/SCAMP allows dynamic simulation, steady-state analysis and metabolic control analysis, and provides an interface to the Systems Biology Workbench (SBW) [33]. BioSPICE is a more recently developed modelling framework for metabolic as well as genetic networks [34]. In addition to these tools, several program packages for metabolic flux analysis have been developed (Table 2). One of these programs, MetaFluxNet, is a stand-alone program package for managing information on metabolic reaction networks and for analyzing metabolic fluxes [35]. Systems biology markup language (SBML) is also supported in the recent version released (version 1.6.9.9).

Systemic and integrative strategy for developing improved strains

The strategic foundation of systems biotechnology is largely based on the systemic integration of high-throughput x-omic analysis and *in silico* modelling or simulation. Figure 3 outlines the conceptual procedure for systems biotechnological research. At the outset, the computational model describing the metabolic system is constructed. This model can be used to analyze and/or predict the system's behaviour for a particular experimental situation under systematic perturbation (e.g. gene deletion or addition and well-designed different culture conditions). The results of this *in silico* study suggest new experimental designs to test the hypothesis generated. The experiments include not only the genetic and metabolic engineering of strains but also high-throughput x-omic experiments to generate more global data. The resultant observations are compared with the prior *in silico* prediction to validate the working hypothetical model. In this way, computational model and experimental design are continuously modified in a cyclic manner.

Palsson and colleagues demonstrated the usefulness of constraints-based flux analysis to predict the effects of genetic modifications on cellular metabolic characteristics [36]. The ability of a constraints-based model to describe consequences of genetic modifications was examined by creating *E. coli* mutants and forcing them to undergo adaptive evolution under different growth conditions. The mutant strains evolved to computationally predicted growth phenotypes. This integrated approach of mathematical and experimental work can be applied in designing strains with improved metabolic performance.

In a recent study by Covert *et al.* [37], a simulation of an *E. coli* genome-scale model was integrated with transcriptional regulatory data and high-throughput growth profiles were obtained using multi-well plates. First, the *in silico* model of *E. coli* was fine-tuned by incorporating the regulatory circuits using the Boolean rules. The expression of 479 genes is regulated by the products of 104 regulatory genes. Then, the high-throughput growth rate data, gene expression data and the metabolic flux profiles predicted by the *in silico* model were combined to characterize the metabolic network.

We recently reported the complete genome sequence of *Mannheimia succiniciproducens*, which can produce large amounts of succinic acid, along with other acids [38]. To redesign the metabolic pathways for enhanced succinic acid production it is essential to understand the metabolic characteristics under various conditions. Based on the complete genome sequence, a genome-scale *in silico* metabolic model, composed of 373 reactions and 352 metabolites, was constructed. Metabolic flux analyses were carried out under various conditions, suggesting that CO₂ is important for cell growth as well as the carboxylation of phosphoenolpyruvate to oxaloacetate, which is then converted to succinic acid by the reductive TCA cycle using fumarate as a major electron acceptor. Based on these findings, the strategies for genome-scale metabolic engineering of *M. succiniciproducens* could be suggested. According to Galperin "This paper is probably the first example of a new approach to complete genomes, which

Table 2. Useful databases for systems biotechnological research and software tools for metabolic flux analysis

Database	Description	Availability
Primary sequence and genomics databases		
DDBJ	DNA Database of Japan	http://www.ddbj.nig.ac.jp
EMBL	Europe's primary nucleotide sequence resource	http://www.ebi.ac.uk/embl.html
Entrez Genomes	NCBI's collection of databases for the analysis of viral, prokaryotic and eukaryotic genomes	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome
GOLD	Information regarding complete and ongoing genome projects	http://www.genomesonline.org/
MBGD	Microbial genome database for comparative analysis	http://mbgd.genome.ad.jp/
TIGR	Microbial genomes and chromosomes	http://www.tigr.org/tdb/mdb/mdbcomplete.html
Microarray databases		
ArrayExpress	Public repository for microarray data by EMBL	http://www.ebi.ac.uk/arrayexpress
GEO	High-throughput gene expression, molecular abundance data repository by NCBI	http://www.ncbi.nlm.nih.gov/geo/
SMD	A microarray research database for Stanford investigators and their collaborators	http://genome-www.stanford.edu/microarray
Protein sequence and proteomics databases		
COG	Clusters of orthologous groups of proteins	http://www.ncbi.nlm.nih.gov/COG
MIPS	Protein and genomic sequence database	http://mips.gsf.de/
PAD	A database for statistical and comparative analyses of the predicted proteomes of fully sequenced organisms	http://www.ebi.ac.uk/proteome/
SWISS-PROT	Curated protein sequence database with a high level of annotation	http://www.expasy.org/sprot
STRING	Predicted functional associations between proteins	http://www.bork.embl-heidelberg.de/STRING
GELBANK	A database of two-dimensional gel electrophoresis (2DE) gel images of proteomes	http://gelbank.anl.gov
SWISS-2DPAGE	A database containing ~1200 entries in 36 reference maps	http://www.expasy.org/ch2d/
Metabolic pathways database		
BioCarta	Online maps of metabolic and signaling pathways	http://www.biocarta.com/genes/allPathways.asp
BioSilico	Integrated metabolic database system	http://biosilico.kaist.ac.kr
BRENDA	Enzyme names and properties: sequence, structure, specificity, stability, reaction parameters	http://www.brenda.uni-koeln.de/
Klotho	Biochemical Compounds Declarative Database	http://www.biocheminfo.org/klotho/
LIGAND	A composite database consisting of COMPOUND, GLYCAN, REACTION, and ENZYME databases	http://www.genome.ad.jp/ligand/
MetaCyc	A database of nonredundant, experimentally elucidated metabolic pathways from more than 240 different organisms.	http://metacyc.org/
PathDB	A rational database for metabolic information accessed through a Java client program.	http://www.ncgr.org/pathdb/
UM-BBD	A database containing information on microbial biocatalytic reactions and biodegradation pathways	http://umbbd.ahc.umn.edu/
Program	Characteristics	Availability
FBA	Flux balance analysis, phase plane analysis, robustness analysis	http://systemsbiology.ucsd.edu/downloads/fba.html
FluxAnalyzer ^a	Metabolic flux analysis, flux balance analysis, structural pathway analysis, Static graphical representation, SBML support	http://www.mpi-magdeburg.mpg.de/proejects/fluxanalyzer
Fluxor	Flux balance analysis, MOMA implementation SBML support	http://arep.med.harvard.edu/moma/biospicefluxor.html
SimPheny ^b	Flux balance analysis, phase plane analysis, structural pathway analysis, static graphical representation	http://www.genomatica.com/
INSILICO Discovery ^b	Metabolic flux analysis, flux balance analysis, structural pathway analysis, dynamic simulation, dynamic and static graphical representation, interface to database (KEGG)	http://www.insilico-biotechnology.com/
Metabologica ^a	Flux balance analysis, isotope flux analysis, structural pathway analysis, dynamic simulation, static graphical representation, SBML support	http://www.svizsystem.com/metabologica.htm
MetaFluxNet	Metabolic flux analysis, flux balance analysis, comparative flux analysis, dynamic graphical representation, SBML support, interface to database (BioSilico)	http://mbel.kaist.ac.kr/

^aFlux calculation under MATLAB environment.

^bCommercial software.

goes from genome sequence straight to chemical engineering. Such approaches will likely become common in biotechnology of the future' [39]. It is important to note that genome-scale metabolic flux analyses were carried out using the metabolite formation rates from a minimal number of cultivation experiments as additional

constraints, thus preventing exhaustive wet experiments while providing realistic simulation results for understanding metabolic characteristics of this relatively unknown bacterium. More examples and insights of strain improvement through systems biotechnological research cycle can be found in recent reviews [12,40].

Table 3. Genome-scale *in silico* metabolic models of microorganisms

Microorganism	No. of genes	No. of reactions	No. of metabolites	Year	Refs
<i>Escherichia coli</i>	660	720	436	2000	[47]
	904	931	625	2003	[48]
	952	979	814	2004	Lee, S.Y. et al. (unpublished)
<i>Geobacter sulfurreducens</i>	588	523	541	2004	[26]
<i>Haemophilus influenzae</i>	296	488	343	1999	[49]
<i>Helicobacter pylori</i>	291	388	340	2002	[50]
<i>Mannheimia succiniciproducens</i>	325	373	352	2004	[38]
<i>Saccharomyces cerevisiae</i>	708	842	584	2003	[51]
	750	1149	646	2004	[52]

Future prospects

Systems biotechnology is now in its early stages of development and presents a variety of technical challenges. The central task of systems biotechnology is to comprehensively collect global cellular information, such as omics data, and to combine these data through metabolic, signaling and regulatory networks to generate predictive computational models of the biological system. Because each x-ome alone is not enough to understand cellular physiology and regulatory mechanisms, combined

analysis will become more and more important. Most combined x-ome analyses described above did not consider the correlations among different x-omes. Therefore, new data mining approaches are needed for deciphering the correlations among these highly heterogeneous omics data, which is one of the prime targets of bioinformatics research. Because the levels of RNAs, proteins, metabolites and fluxes vary independently but in a highly orchestrated fashion using various regulatory circuits, multivariate analysis might be necessary for finding correlations among these omics data. This type of integrated analysis will become essential to better understand cellular physiology and metabolism at the systems level and to design strategies for metabolic and cellular engineering of organisms.

Computational modelling of complex biological systems is invaluable for organizing and integrating the available metabolic knowledge and designing the right experiments. Currently, however, the true predicting power of biological simulation is limited by insufficient knowledge and information on regulation and kinetics. One of the most challenging problems in the field of systems modelling and simulation is multi-scale and multi-level modelling. The question of how different biological data, ranging from DNA and protein to metabolic flux and cell growth, can be handled within one model will be answered by multiple time-scale and spatial analysis. This will lead to the more accurate prediction of phenotypic behavior from omics information. Simplifying the complex models by

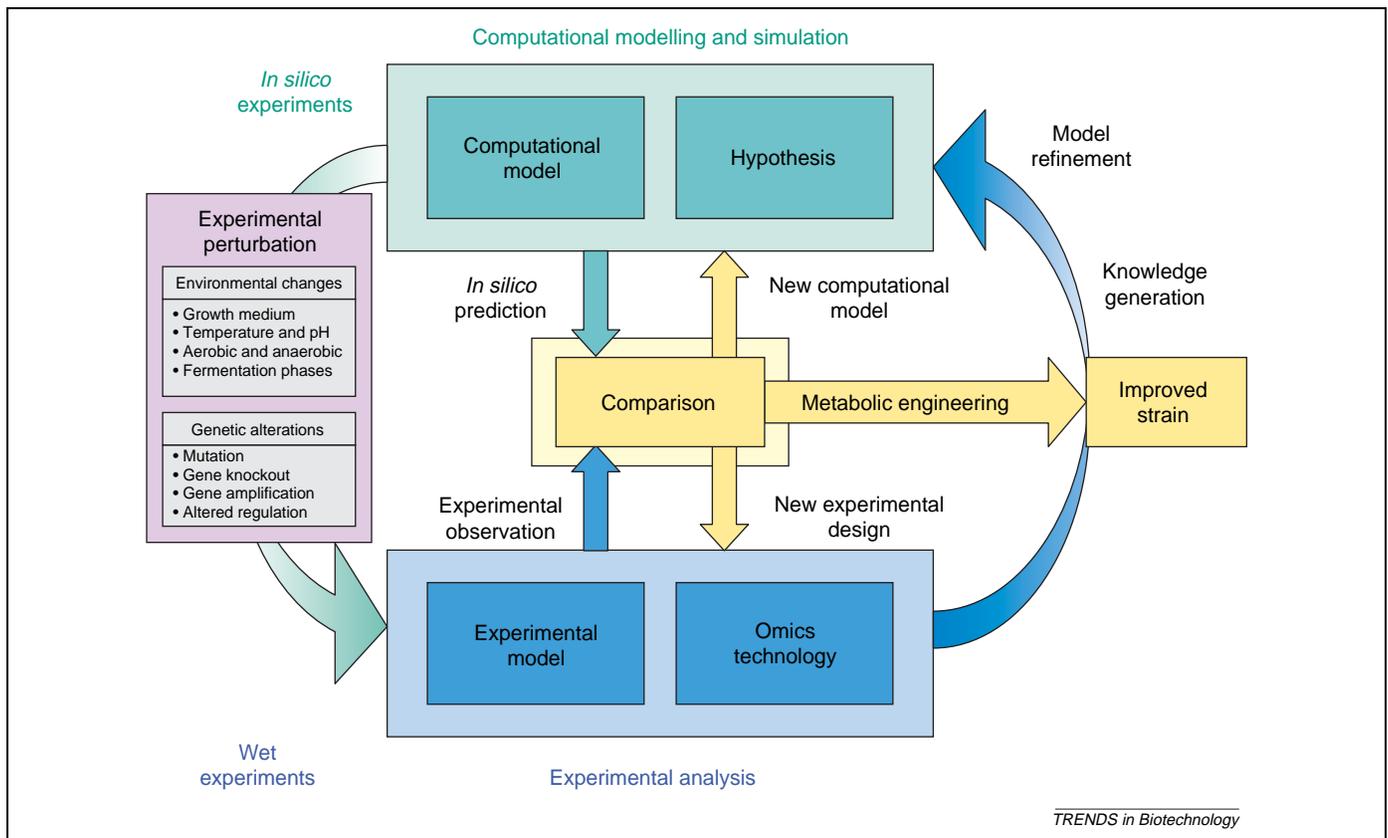


Figure 3. Overview of the systems biotechnological research cycle during the process of knowledge generation. Hypothesis-driven computational modelling or simulation and technology-driven high-throughput experimental analyses are combined to generate new knowledge via *in silico* and wet experiments in which *in silico* model and experimental design are continuously evolved during the iterations.

model reduction methods (e.g. lumping, sensitivity analysis and time-scale-based model reduction) is one option available at this point [41].

Systems biotechnology will have an increasing impact on industrial biotechnology in the future. All steps of biotechnological development, from up-stream (strain, cell and organism development) and mid-stream (fermentation and other unit operations) to down-stream processes, will benefit significantly by taking systems biotechnological approaches.

Acknowledgements

Our work described in this paper was supported by the Korean Systems Biology Research Program (M10309020000-03B5002-00000) of the Ministry of Science and Technology and by the BK21 project. Further support from the LG Chem Chair Professorship, KOSEF and the IBM-SUR program are greatly appreciated.

References

- 1 Parekh, S. *et al.* (2000) Improvement of microbial strains and fermentation processes. *Appl. Microbiol. Biotechnol.* 54, 287–301
- 2 Patterson, S.D. and Aebersold, R.H. (2003) Proteomics: the first decade and beyond. *Nat. Genet.* 33, 311–323
- 3 Stephanopoulos, G. (2002) Metabolic engineering by genome shuffling. *Nat. Biotechnol.* 20, 707–712
- 4 Oliver, D.J. *et al.* (2002) Functional genomics: high-throughput mRNA, protein, and metabolite analyses. *Metab. Eng.* 4, 98–106
- 5 Bro, C. and Nielsen, J. (2004) Impact of 'ome' analyses on inverse metabolic engineering. *Metab. Eng.* 6, 204–211
- 6 Han, M.-J. and Lee, S.Y. (2003) Proteome profiling and its use in metabolic and cellular engineering. *Proteomics* 3, 2317–2324
- 7 Hermann, T. (2004) Using functional genomics to improve productivity in the manufacture of industrial biochemicals. *Curr. Opin. Biotechnol.* 15, 444–448
- 8 Endy, D. and Brent, R. (2001) Modelling cellular behaviour. *Nature* 409, 391–395
- 9 Selinger, D.W. *et al.* (2003) On the complete determination of biological systems. *Trends Biotechnol.* 21, 251–254
- 10 Stelling, J. (2004) Mathematical models in microbial systems biology. *Curr. Opin. Microbiol.* 7, 513–518
- 11 Wiechert, W. (2002) Modeling and simulation: tools for metabolic engineering. *J. Biotechnol.* 94, 37–63
- 12 Patil, K.R. *et al.* (2004) Use of genome-scale microbial models for metabolic engineering. *Curr. Opin. Biotechnol.* 15, 64–69
- 13 Kolisnychenko, V. *et al.* (2002) Engineering a reduced *Escherichia coli* genome. *Genome Res.* 12, 640–647
- 14 Ohnishi, J. *et al.* (2002) A novel methodology employing *Corynebacterium glutamicum* genome information to generate a new L-lysine-producing mutant. *Appl. Microbiol. Biotechnol.* 58, 217–223
- 15 Choi, J.H. *et al.* (2003) Enhanced production of insulin-like growth factor I fusion protein in *Escherichia coli* by coexpression of the down-regulated genes identified by transcriptome profiling. *Appl. Environ. Microbiol.* 69, 4737–4742
- 16 Han, M.J. *et al.* (2001) Proteome analysis of metabolically engineered *Escherichia coli* cells producing poly(3-hydroxybutyrate). *J. Bacteriol.* 183, 301–308
- 17 Han, M.J. *et al.* (2003) Engineering *Escherichia coli* for increased production of serine-rich proteins based on proteome profiling. *Appl. Environ. Microbiol.* 69, 5772–5781
- 18 Oliver, S.G. (1998) Introduction to functional analysis of the yeast genome. In *Methods in Microbiology* (Brown, A.J.P. and Tuite, M., eds), pp. 1–13, Academic Press
- 19 Wittmann, C. and Heinzle, E. (2001) Modeling and experimental design for metabolic flux analysis of lysine-producing *Corynebacteria* by mass spectrometry. *Metab. Eng.* 3, 173–191
- 20 Sanford, K. *et al.* (2002) Genomics to fluxomics and physiomics - pathway engineering. *Curr. Opin. Microbiol.* 5, 318–322
- 21 Yoon, S.H. *et al.* (2003) Combined transcriptome and proteome analysis of *Escherichia coli* during high *Cell* density culture. *Biotechnol. Bioeng.* 81, 753–767
- 22 Lee, J.H. *et al.* (2003) Global analyses of transcriptomes and proteomes of a parent strain and an L-threonine-overproducing mutant strain. *J. Bacteriol.* 185, 5442–5451
- 23 Askenazi, M. *et al.* (2003) Integrating transcriptional and metabolite profiles to direct the engineering of lovastatin-producing fungal strains. *Nat. Biotechnol.* 21, 150–156
- 24 Krömer, J.O. *et al.* (2004) In-depth profiling of lysine-producing *Corynebacterium glutamicum* by combined analysis of the transcriptome, metabolome, and fluxome. *J. Bacteriol.* 186, 1769–1784
- 25 Ishii, N. *et al.* (2004) Toward large-scale modeling of the microbial cell for computer simulation. *J. Biotechnol.* 113, 281–294
- 26 Price, N.D. *et al.* (2004) Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat. Rev. Microbiol.* 2, 886–897
- 27 Tomita, M. *et al.* (1999) E-CELL: software environment for whole-cell simulation. *Bioinformatics* 15, 72–84
- 28 Burgard, A.P. and Maranas, C.D. (2001) Probing the performance limits of the *Escherichia coli* metabolic network subject to gene additions or deletions. *Biotechnol. Bioeng.* 74, 364–375
- 29 Segre, D. *et al.* (2002) Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci. U. S. A.* 99, 15112–15117
- 30 Burgard, A.P. *et al.* (2003) Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol. Bioeng.* 84, 647–657
- 31 Mendes, P. (1993) GEPASI: a software package for modeling the dynamics, steady states and control of biochemical and other systems. *Comput. Appl. Biosci.* 9, 563–571
- 32 Goryanin, I. *et al.* (1999) Mathematical simulation and analysis of cellular metabolism and regulation. *Bioinformatics* 15, 749–758
- 33 Sauro, H.M. *et al.* (2003) Next generation simulation tools: the Systems Biology Workbench and BioSPICE integration. *OMICS* 7, 355–372
- 34 Garvey, T.D. *et al.* (2003) BioSPICE: access to the most current computational tools for biologists. *OMICS* 7, 411–420
- 35 Lee, D.-Y. *et al.* (2003) MetaFluxNet: the management of metabolic reaction information and quantitative metabolic flux analysis. *Bioinformatics* 19, 2144–2146
- 36 Fong, S.S. and Palsson, B.O. (2004) Metabolic gene-deletion strains of *Escherichia coli* evolve to computationally predicted growth phenotypes. *Nat. Genet.* 36, 1056–1058
- 37 Covert, M.W. *et al.* (2004) Integrating high-throughput and computational data elucidates bacterial networks. *Nature* 429, 92–96
- 38 Hong, S.H. *et al.* (2004) The genome sequence of the capnophilic rumen bacterium *Mannheimia succiniciproducens*. *Nat. Biotechnol.* 22, 1275–1281
- 39 Galperin, M.Y. (2004) Genomes back-to-back: when sequencing race is a good thing. *Environ. Microbiol.* 6, 1205–1209
- 40 Stephanopoulos, G. (2004) Exploiting biological complexity for strain improvement through systems biology. *Nat. Biotechnol.* 22, 1261–1267
- 41 Okino, M.S. and Mavrouniotis, M.L. (1998) Simplification of mathematical models of chemical reaction systems. *Chem. Rev.* 98, 391–408
- 42 Ruckert, C. *et al.* (2003) Genome-wide analysis of the L-methionine biosynthetic pathway in *Corynebacterium glutamicum* by targeted gene deletion and homologous complementation. *J. Biotechnol.* 104, 213–228
- 43 Ohnishi, J. *et al.* (2003) Efficient 40 degrees C fermentation of L-lysine by a new *Corynebacterium glutamicum* mutant developed by genome breeding. *Appl. Microbiol. Biotechnol.* 62, 69–75
- 44 Tummala, S.B. *et al.* (2003) Transcriptional analysis of product-concentration driven changes in cellular programs of recombinant *Clostridium acetobutylicum* strains. *Biotechnol. Bioeng.* 84, 842–854
- 45 Kabir, M.M. and Shimizu, K. (2003) Fermentation characteristics and protein expression patterns in a recombinant *Escherichia coli* mutant lacking phosphoglucose isomerase for poly(3-hydroxybutyrate) production. *Appl. Microbiol. Biotechnol.* 62, 244–255
- 46 Dauner, M. *et al.* (2002) Intracellular carbon fluxes in riboflavin-producing *Bacillus subtilis* during growth on two-carbon substrate mixtures. *Appl. Environ. Microbiol.* 68, 1760–1771
- 47 Edwards, J.S. and Palsson, B.O. (2000) The *Escherichia coli* MG1655 *in silico* metabolic genotype: its definition, characteristics, and capabilities. *Proc. Natl. Acad. Sci. U. S. A.* 97, 5528–5533

- 48 Reed, J.L. *et al.* (2003) An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biol.* 4, R54
- 49 Edwards, J.S. and Palsson, B.Ø. (1999) Systems properties of the *Haemophilus influenzae* Rd metabolic genotype. *J. Biol. Chem.* 274, 17410–17416
- 50 Schilling, C.H. *et al.* (2002) Genome-scale metabolic model of *Helicobacter pylori* 26695. *J. Bacteriol.* 184, 4582–4593
- 51 Forster, J. *et al.* (2003) Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Res.* 13, 244–253
- 52 Duarte, N.C. *et al.* (2004) Reconstruction and validation of *Saccharomyces scerevisiae* iND750, a fully compartmentalized genome-scale metabolic model. *Genome Res.* 14, 1298–1309

Five things you might not know about Elsevier

1.

Elsevier is a founder member of the WHO's HINARI and AGORA initiatives, which enable the world's poorest countries to gain free access to scientific literature. More than 1000 journals, including the *Trends* and *Current Opinion* collections, will be available for free or at significantly reduced prices.

2.

The online archive of Elsevier's premier Cell Press journal collection will become freely available from January 2005. Free access to the recent archive, including *Cell*, *Neuron*, *Immunity* and *Current Biology*, will be available on both ScienceDirect and the Cell Press journal sites 12 months after articles are first published.

3.

Have you contributed to an Elsevier journal, book or series? Did you know that all our authors are entitled to a 30% discount on books and stand-alone CDs when ordered directly from us? For more information, call our sales offices:

+1 800 782 4927 (US) or +1 800 460 3110 (Canada, South & Central America)
or +44 1865 474 010 (rest of the world)

4.

Elsevier has a long tradition of liberal copyright policies and for many years has permitted both the posting of preprints on public servers and the posting of final papers on internal servers. Now, Elsevier has extended its author posting policy to allow authors to freely post the final text version of their papers on both their personal websites and institutional repositories or websites.

5.

The Elsevier Foundation is a knowledge-centered foundation making grants and contributions throughout the world. A reflection of our culturally rich global organization, the Foundation has funded, for example, the setting up of a video library to educate for children in Philadelphia, provided storybooks to children in Cape Town, sponsored the creation of the Stanley L. Robbins Visiting Professorship at Brigham and Women's Hospital and given funding to the 3rd International Conference on Children's Health and the Environment.