Journal of Biochemistry and Molecular Biology, Vol. 37, No. 1, January 2004, pp. 83-92

JOURNAL OF Biochemistry and Molecular Biology

Review

From the Sequence to Cell Modeling: Comprehensive Functional Genomics in *Escherichia coli*

Hirotada Mori*
Research and Education Center of Genetic Information, Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara 630-0101
Institute of Advanced Biosciences, Keio University, Tsuruoka, Yamagata 997-0017, Japan

Received 20 December 2003

As a result of the enormous amount of information that has been collected with *E. coli* over the past half century (e.g. genome sequence, mutant phenotypes, metabolic and regulatory networks, etc.), we now have detailed knowledge about gene regulation, protein activity, several hundred enzyme reactions, metabolic pathways, macromolecular machines, and regulatory interactions for this model organism. However, understanding how all these processes interact to form a living cell will require further characterization, quantification, data integration, and mathematical modeling, systems biology. No organism can rival *E. coli* with respect to the amount of available basic information and experimental tractability for the technologies needed for this undertaking. A focused, systematic effort to understand the *E. coli* cell will accelerate the development of new post-genomic technologies, including both experimental and computational tools. It will also lead to new technologies that will be applicable to other organisms, from microbes to plants, animals, and humans. *E. coli* is not only the best studied free-living model organism, but is also an extensively used microbe for industrial applications, especially for the production of small molecules of interest. It is an excellent representative of Gram-negative commensal bacteria. *E. coli* may represent a perfect model organism for systems biology that is aimed at elucidating both its free-living and commensal life-styles, which should open the door to whole-cell modeling and simulation.

Keywords: *E. coli*, Genome, Modeling, Systems biology

*To whom correspondence should be addressed.
Tel: 81-743-72-5660; Fax: 81-743-72-5669
E-mail: hmori@gtc.aist-nara.ac.jp

The *E. coli* Genome Sequence

The complete *E. coli* genome sequence was determined through independent efforts by American and Japanese groups using two different strains of *E. coli* K-12, MG1655 and W3110 (Aiba *et al.*, 1996; Itoh *et al.*, 1996; Oshima *et al.*, 1996; Blattner *et al.*, 1997; Yamamoto *et al.*, 1997). These strains were diverged from the same ancestral strain about 50 years ago, resulting in slight but significant differences, including the large inversion involving the ribosomal RNA genes (see below). A complete genome sequence analysis revealed precise differences between the two strains. A comparison of the genome sequence has revealed relatively infrequent, 1 per 105 base substitution mutations in the protein coding region (Itoh *et al.*, 1999). Recently, a Japanese group confirmed the differences between the genomes of MG1655 and W3110 by PCR-based direct sequencing. They found that only 9 bases in 8 different ORFs represent actual sequence differences besides large scale insertions or deletions, such as IS elements (Horiuchi, in preparation). In summary, the *E. coli* genome seems to tolerate large rearrangements such as insertion, deletion and recombination better than micro-scale changes like base substitution. In 2001, the pathogenic *E. coli* strain O157, which is very closely related to *E. coli* K-12, was subjected to genomic sequencing by two groups (Hayashi *et al.*, 2001; Perna *et al.*, 2001). The results obtained with this pathogenic strain agreed well with the above findings with strain K-12. *E. coli* O157 appears to have acquired its pathogenicity mainly by horizontal gene transfer that is mediated by a temperate bacteriophage.

Prediction of ORFs from the complete genome sequence of *E. coli* While more than 400 bacterial genomes have been sequenced, the functional assignment of all the gene products has not yet been accomplished for any microorganism. Extensive studies in a few model organisms, including

Escherichia coli and *Saccharomyces cerevisiae*, elucidated the functions of many genes and gene products, particularly by using techniques of genetics, biochemistry and physiology, etc. *E. coli* K12, the best-studied microorganism, is estimated to contain about 4,400 genes of which about 2,000 have not been characterized experimentally (Mori *et al.*, 2000). Altogether about 3,700 genes can be assigned or predicted a function with reasonable assurance, based on biochemical experiments and computational analysis in *E. coli* and other microorganisms. Of the remaining approximate 700 genes, 650 show sequence similarity to genes of unknown function in other bacteria, whereas 50 show no obvious similarity to any known genes. The assignment of function to these unknown genes is one of the major targets of functional genomics in *E. coli*. In addition to their fundamental importance for understanding *E. coli* biology, these functional assignments are significant for three other reasons: (1) complete functional assignment will result in the discovery of new physiological and biochemical pathways, (2) will facilitate functional assignment in other bacteria, and (3) will lead to the identification of new targets for antibiotic design other than for biotechnological development.

Besides, elucidation of individual gene functions, which is basic to the understanding of a cell, systematic analysis of relationships between genes or gene products is also a significant target that is just starting to be explored systematically.

The availability of numerous complete genome sequences has considerably accelerated the comparative approach. Various analyses for clustering genes (ORFs) have been performed. New concepts have been proposed such as ancient conserved regions (ACR) (Koonin *et al.*, 1995), clusters of orthologous groups (COGs) (Tatusov *et al.*, 1997) and modules (Riley and Labedan, 1997). These analyses are quite valuable, not only from an evolutionary point of view, but also from a more practical view point as well, such as the functional prediction of hitherto uncharacterized ORFs.

A workshop on the Annotation of *Escherichia coli* K-12 2003, chaired by Monica Riley, was recently held at Woods Hole, MA, on November 14-18. Fourteen scientists from the US, Europe and Japan gathered and took parts in two tasks. One group addressed the annotation of known and predicted gene products. The other group focused on gene boundaries and sequences. The intensively curated and coordinated data will be submitted to GenBank in early 2004, and made available on the Internet for public access (M. Riley, personal communication).

Repetitive sequences, sites, RNA genes etc. Many kinds of repetitive sequences are found within the genome, some of which have important physiological functions. The distribution of repeats along the chromosome is not random and seems to be related to some feature of DNA replication. Repetitive sequences in the *E. coli* genome are encountered in different contexts. Various classes of repeats are present in diverse prokaryotes, including *E. coli*. Coding sequences such as ribosomal RNA genes, transfer RNA genes, and insertion sequences are usually present in multiple copies, but their copy numbers are relatively low. Other interspersed repetitive DNA sequences are relatively short but abundant and located within intercistronic non-coding sequences. *E. coli* has 7 copies of rRNA coding genes (*rrn*) and an additional copy of the 5S rRNA gene (Blattner *et al*., 1997). Other rRNA-related DNA sequences (TRIP) showing significant similarity to 5S rRNA were recently discovered. These may have important functions that are related to 5S rRNAs (Rudd, 1999). The rRNA genes are located within half of the chromosome that contains the origin of DNA replication (*oriC*), whereas many of the TRIP sequences are located in the other half of the chromosome. Moreover, *rrn* and some of the related TRIP sequences are distributed symmetrically on the leading strands on both sides of *oriC*. Consistent with this location, the transcription of *rrn* operons generally proceeds away from the replication origin. Repetitive sequences, such as rRNA genes, provide a driving force for genome rearrangement. Large inversions of the genome between rRNA genes (between *rrnD* and *rrnE*) have been documented in *E. coli*, and one such inversion was found in the W3110 strain (Hill and Harnish, 1981). Although the direction of *rrn* gene transcription is strictly oriented away from the replication origin, the inversion in *E. coli* appears to be stable because the geometric relationship between *rrn* operons and the replication origin are preserved. Insertion Sequences (IS) represent another family of repetitive elements and cause genetic variation among different strains of *E. coli*: both the abundance and distribution of insertion sequences can vary in different strains (Mahillon and Chandler, 1998; Mahillon *et al.*, 1999). The W3110 strain contains about 60 ISs that belong to at least 10 distinct families. At least 10 of these ISs differ in their location when compared to MG1655 strain. These results testify to the great variability in both number and family of ISs in closely related strains. Interspersed repetitive sequences represent relatively short (usually less than 500 bp), non-coding, intercistronic and dispersed elements that are found in bacterial genomes. Six classes of highly repetitive sequences, BIME, IRU (Sharples and Lloyd, 1990), Box C (Bergler *et al.*, 1992), RSA (Mizobuchi, personal communication), iap (Ishino *et al.*, 1987; Nakata *et al.*, 1989), and Ter Sequences (Hill, 1996) have been identified (Table 1). These sequences were

Table 1. Extragenic highly repetitive sequence families

Sequence	Size (bp)	Copy number
BIME	40~400	~800
Ter	30	63
BoxC	56	36
Iap	29	23
IRU	127	19
RSA	152	6

primarily discovered by computer analysis of sequence data. None of these sequences encode proteins. They are dispersed throughout the chromosome.

Finally, a number of RNAs that do not function as either mRNAs, tRNAs, or rRNAs have been discovered, mostly fortuitously. The non-mRNAs have been predominantly termed as small RNAs in bacteria. The potentially important function of some of them was recently documented (Storz, 2002). Systematic analysis of non-coding RNA in *E. coli* has just started (G. Storz, personal communication).

Post Genome Sequencing Project

As previously described, the function of nearly half of the total ORFs in *E. coli* is not unknown, of which 20% remain difficult even to predict their function. The latest estimates reveal that about 700 to 800 of total ORFs have no attributable function. Therefore, a high priority will be placed on the development of novel, high-throughput technologies to identify their function. A new comprehensive "molecular tool kit" would be required to define unknown gene functions and to assign potentially new roles to known genes. A large number of valuable plasmid constructs, *E. coli* strains, assay tools, and other biological materials that are useful for analyzing the gene function have been constructed, developed, collected, and tested. These will form key resources for getting basic and comprehensive knowledge on *E. coli* biology and for exploiting these data with *E. coli* for other organisms. Furthermore, methods for genome-wide analysis of transcriptional regulation (transcriptome), protein dynamics (proteome) and flow of metabolites (metabolome) have rapidly evolved to make the best use of the rich DNA sequence information ("-ome" is a Greek suffix for "whole").

Experimental resources A DNA sequence analysis has identified about 4,400 protein-coding genes in *E. coli*. The power of genetically- tractable model organisms resides in the fact that they can facilitate the global and systematic analysis of physiological gene function *in vivo*. Precise genetic manipulation is particularly important for functional genomics. Genome sequence data has permitted the design of oligo DNA primers for the precise amplification of entire ORFs and generation of a complete set of histidine-tagged ORF clones (with or without fusion to the GFP gene) (Mori *et al.*, 2000).

One approach for systematic functional analysis is to make use of gene deletion (replacement) that is obtained by homologous recombination. Targeted gene replacement, once thought to be difficult in *E. coli*, can now be dealt with by using the techniques that were developed by Wanner and colleagues (Datsenko and Wanner, 2000). The systematic attempts to construct these resources for functional genomics is now rapidly raising *E. coli* to one of the leading organisms in the field of functional genomics.

ORF clones

ORF clones will provide a basic genetic tool for studying gene function, since they provide a template for PCR amplification and for preparing purified gene products etc. To clone all of the genes of *E. coli*, a plasmid vector with the following properties was constructed: (1) high copy number plasmid, (2) IPTG inducible expression of cloned ORF and repression of expression by *lac*I^q, (3) a Histidine tag coding sequence attached to the N-terminal of ORF, (4) in-frame fusion with GFP coding sequence at the C-terminal end, (5) generation of *Sfi*I restriction sites at both boundaries of the cloned ORF, (6) possibility of GFP fragment removal by *Not*I (Kitagawa, in preparation). The whole set of PCR amplified ORF fragments were cloned into the *Stu*I site of this vector. As far as we know, these clones represent the only comprehensive collection of *E. coli* ORFs that is currently publicly available. Fig. 1 shows the structure of these clones.

Knock-out mutants

A systematic mutational analysis of genes in their chromosomal location should provide basic information and insight into their function. A large number of these mutants have already been established, primarily by random insertion mutagenesis (Miki *et al.*, personal communication). Although this effort provided the research community with a unique collection of mutants, it was necessary to determine the insertion position of the transposons. In addition, some complications could not be avoided due to the nature of transposon mutagenesis, such as incomplete disruption of the targeted gene or polar effects on the downstream genes. Regarding this last point, only the set of in-frame deletion mutants can avoid this problem. Most bacteria, including *E. coli*, are not readily transformable with linear DNA because of the presence of intracellular exonucleases that degrade transformed linear DNA. In contrast, genes can be directly disrupted in *Saccharomyces cerevisiae* by transformation with PCR fragments encoding a selectable marker having only 35 nt of flanking DNA homologous to the chromosome (Baudin *et al.*, 1993). On the other hand, it has long been known that many bacteriophages encode their own homologous recombination systems (Smith, 1988). It was recently shown that the λ Red (*g, b, exo*) function promotes a greatly enhanced rate of recombination over that exhibited by the *recBC sbcB* or *recD* mutants when using linear DNA. Wanner and his colleagues developed a convenient procedure based on the l Red system that provides an efficient way to isolate replacement mutants using PCR fragments encoding an antibiotic resistance gene and having only 40 to 50 nt of flanking regions (Datsenko and Wanner, 2000). A comprehensive and clean deletion mutant library for all *E. coli* ORFs (the KO collection, Knock-Out and KEIO University) using the λ Red system is now under construction and will be open to the public soon (Baba, in preparation). The design of isolating these deletions is illustrated in Fig. 2.

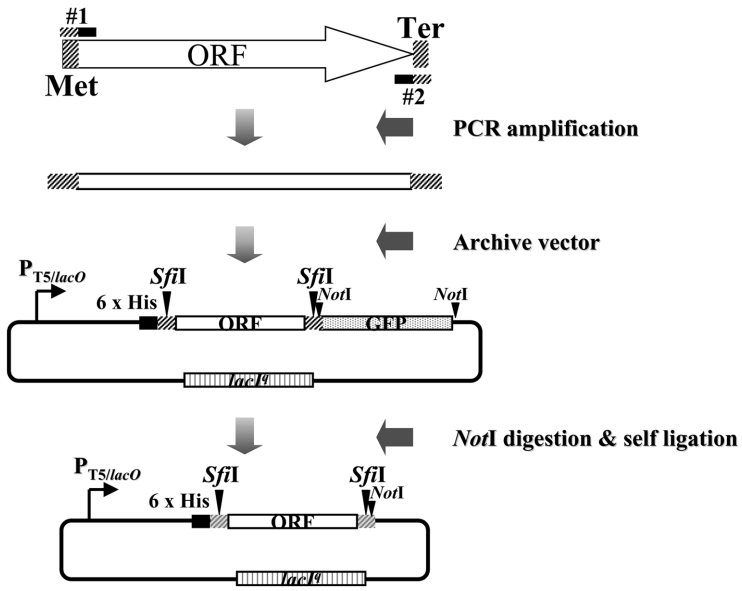

Fig. 1. Construction of Archive clones.

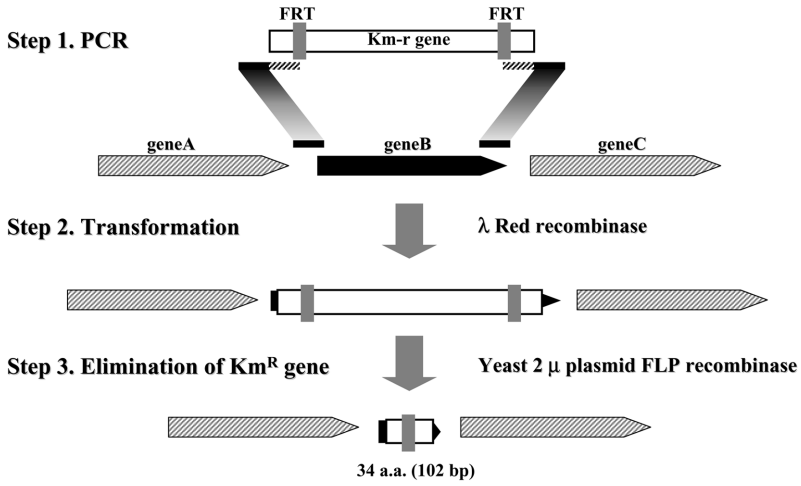

Fig. 2. Construction of deletion mutants.

Large deletion mutants

Two recent publications describe deletions of large chromosomal segments from the *E. coli* genome. Kim and his colleagues made two large libraries of independent transposon mutants using modified Tn5 transposons with two different selection markers, thereby precisely mapping their chromosomal location (Yu *et al.*, 2002). This method allows the integration of the mapped transposon insertion carrying *loxP* site from each of the mutant libraries into the same chromosome followed by excision of the flanking genomic segments by site-specific Cre mediated *loxP* recombination. Alternately, Blattner and his colleagues used a deletion "by design approach" and have already generated lineages in which more than 10% of the genome has been eliminated without loss of viability (Kolisnychenko *et al.*, 2002). Another large-scale deletion construction has been undertaken by Katoh (J. Katoh, personal communication). Basically, the approach consists of markerless gene replacement, and has already generated a genome lacking more than 25% of the original genome DNA. Intermediates have been saved and descendents with larger deletions are being generated. By examining representatives of these collections for growth defects under specific environmental conditions, the effects of losing many genes can be traced simultaneously. Data generated in this type of experiment will be a great asset for cell modeling. These approaches are likely to be especially advantageous when single gene mutations display no discernable phenotypic changes.

Other approaches for functional geneomics

Two different types of technology were recently developed to study protein expression and protein-protein interaction. One

is based on novel tandem affinity purification (TAP) of tag fusion protein. This is a generic procedure to purify target proteins expressed at their natural level under native conditions (Rigaut *et al.*, 1999). To investigate heteromeric protein complexes of unknown composition, standard systems for protein overexpression may lead to the assembly of overexpressed proteins as non-physiological complexes. To overcome this problem, a TAP tag fusion cassette was developed which encodes the calmodulin-binding peptide (CBP), a TEV protease recognition site, and proteinA of *Staphylococcus aureus* (ProtA).

The other method depends on the genome-wide, registered collection of *E. coli* bioluminescent reporter gene fusions. Each of the random fusions of *E. coli* chromosomal DNA fragment to the *Photorhabdus luminescens luxCDBE* reporter gene was precisely mapped by sequencing (Van Dyk *et al.*, 2001). To identify and quantify changes in the expression level, the authors tested this type of fusion and analyzed alterations in the expression levels of heat shock, SOS response and oxidative stress genes.

Transcriptome analysis To exploit the rapid progress in genome research, many novel techniques have been developed including DNA microarray or DNA chip technology that are extremely useful for the analysis of global gene expression (Lockhart *et al.*, 1996). Generally, a DNA microarray is defined as an orderly arrangement of tens to hundreds of thousands of unique DNA molecules of known sequence, usually on a glass slide. Unique DNA molecules are either individually synthesized on a rigid silicon plate (generally referred to as DNA chips and developed by Affymetrix Co.) or prepared from pre-synthesized DNA (synthetic oligo-nucleotides or PCR products) that are spotted and immobilized on a slide glass. The use of DNA microarrays to study *E. coli* gene regulation was first illustrated by Blattner and colleagues. This has rapidly expanded and been applied to study various aspects of transcriptional regulation (Tao *et al.*, 1999). To elucidate the whole transcriptional regulatory network, several systematic approaches using DNA microarray have been performed (Oshima *et al.*, 2002; Masuda and Church, 2003). These analyses were done not only under different growth conditions but also using either the overexpression or deletion of a target regulatory gene. As previously described, large experimental resources including complete sets of clones and deletion mutants of all of the predicted genes of *E. coli* are being established. These resources will contribute in the acceleration of the systematic transcriptome analyses. Accumulation of the results from such these large-scale analyses using DNA microarrays will also assist more traditional biological research, as well as the construction of large databases that will be very beneficial. Public availability of this data is not limited to supplemental data derived from publications on the ftp site of journal publishers, but is also available from integrated databases such as the KEGG database as a systematic collection of microarray data (Kanehisa *et al.*, 2002).

The accumulation of publicly-available DNA microarray analyses data promotes rapid computational analysis of gene expression profiles. In general, bacterial genes form operon in which multiple ORFs are transcribed from the same promoter to form a single mRNA transcript. Prediction of operons using DNA microarray experiments is one approach used to reconstruct gene regulatory networks at the whole genome scale (Sabatti *et al.*, 2002). Reconstruction of global regulatory networks using genome-scale gene expression data sets has also been reported recently (Gutierrez-Rios *et al.*, 2003; Herrgard *et al.*, 2003).

Proteome One of the basic technologies for global analysis of cellular proteins has been developed by O'Farrell as two-dimensional polyacrylamide gel electrophoresis (O'Farrell, 1975). *E. coli* has a long history of protein cataloging by 2D gel (Vanbogelen, 1996). The major limitations of 2D electrophoresis are related to the difficulty in assigning gene identities to observed spots and the lack of spots for proteins that do not separate well on 2D gels. In addition, the method displays a general bias against membrane proteins and proteins of low abundance. However, matrix-assisted laser desorption ionization-mass spectrometry (MALDI-MS) has greatly contributed to the acceleration of protein identification (Figeys *et al.*, 1998). A new gel based separation system (RFHR) has also significantly expanded the separation range of proteins (Wada *et al.*, 1993). The assignment of 2D gel data will be coordinated between two major ongoing initiatives, the CyberCell project (Ellison, personal communication) and the efforts of a Japanese group (Wada, personal communication). iIn addition, several alternative methods, based on an analysis of peptides from whole cell lysates by LC/MS, were developed and successfully used (Gygi *et al.*, 2000; Corbin *et al.*, 2003).

Protein localization

Even though the bacterial cell have no complex intracellular compartments that are hallmarks of eukaryotic cells, there is comprehensive information about the location of proteins within the bacterial cell, which is important for understanding their functions and interactions. Large-scale analyses of protein localization in *S. cerevisiae* have been reported (Ross-Macdonald *et al.*, 1999; Kumar *et al.*, 2002). Recently, Niki *et al.* also reported the comprehensive analysis of protein localization using clones of individual ORFs that were fused with GFP protein under non-induced growth conditions (Niki, personal communication). Localized GFP fluorescence was successfully observed for about 4,000 out of 4,300 genes tested. The patterns of localization were classified roughly into 4 distinct categories. Protein localization, based on subcellular fractionation, is also underway as part of the CyberCell project (Ellison, personal communication) and in parallel at Harvard University (G. Church, personal communication).

Protein-protein interaction
Most cellular processes are carried out by multiprotein complexes. The identification and analysis of these protein complexes provide further insight into the physiological function and molecular mechanisms of the functional units. Following identification and cataloging of all of the proteins that are expressed in a cell, a global analysis of the protein-protein interaction becomes critical for understanding cellular processes. In *S. cerevisiae*, a genome-scale analysis of the protein complexes was performed using the yeast two-hybrid system (Fromont-Racine *et al*., 1997; Uetz *et al.*, 2000; Ito *et al.*, 2001), protein chips (Zhu *et al.*, 2001), or affinity tagged system (Gavin *et al*., 2002). In *E. coli*, there are presently two comprehensive analyses underway using affinity-tagged proteins, chromosomally tagged with TAP (J. Greenblatt, personal communication) and plasmid clones containing an histidine tag (Arifuzzaman, in preparation), to identify the protein-protein interaction, as was done in *S. cerevisiae*. These will allow identification by mass spectrometry of proteins that co-purify with the tagged baits that are thus candidate-interacting proteins. Greenblatt and his colleagues are focusing on studying the interaction of nearly 200 highly conserved proteins that are known to be essential, such as DNA and RNA polymerases. They have developed an interaction network for these proteins. On the other hand, Arifuzzaman and his colleagues performed high-throughput analysis using plasmid clones, although most of the membrane proteins were only poorly purified and failed to function as bait. However, out of 4,300 total ORFs, more than 2,700 ORFs were successfully purified from the plasmid clones and candidates that can interact with His-tagged bait proteins were identified. The total number of observed interactions in this set of 2,700 proteins amounts to an impressive total of about 14,000 potential interactions (Arifuzzaman, in preparation). An analysis of this complex protein interaction network is now underway.

Metabolome A substantial portion of the *E. coli* genome encodes enzymes that interconvert metabolites, synthesize cofactors, and regulate small molecule metabolism. Metabolites can in turn control the gene expression and are allosteric regulators of enzymes. The metabolome can be described as the total complement of metabolites in a cell (Tweeddale *et al.*, 1998). Metabolome analyses can be performed using several approaches, such as metabolite profiling, flux analysis using isotopic tracer, and pathway reconstruction etc. This allows insight into metabolic and physiological responses within a cell. Global metabolite profiling will provide deeper insight not only into metabolism but also into cellular physiology and functional genomics (Fiehn, 2002). This approach has been used for functional genomics studies in plants (Fiehn *et al.*, 2000) and yeast (Raamsdonk *et al.*, 2001). In *E. coli*, metabolites were labeled with C-14 glucose and identified by 2-dimensional thin-layer chromatography after extraction by cold methanol (Maharjan and Ferenci, 2003). Using this approach, the authors identified about 100 metabolite spots. Recently, Soga and colleagues developed a powerful analytical method using capillary electrophoresis-electrospray ionization mass spectrometry (CE-ESI-MS) that dramatically increases the number of metabolites that can be measured simultaneously (Soga *et al.*, 2002a, 2002b, 2003).

E. coli Genome as a Model for Systems Biology

Biology itself is now at a turning point between the past descriptive science and the emerging modern quantitative systems biology. Understanding the causal relationships between the genotype and phenotype will require a very significant expansion of the traditional toolbox that is used by molecular biologists. It must include concepts and techniques from many other scientific disciplines such as physics, mathematics, numerical analysis, stochastic processes, and control theory. Many novel tools must be developed to understand how dynamic, robust but adapting, and developing systems can emerge from the information buried in the genome (Ehrenberg *et al.*, 2003). The genome sequencing efforts and the subsequent bioinformatics analyses have not only defined the molecular parts for a number of living organisms, but also opened up possibilities to reconstruct the metabolic pathways. The stoichiometric coefficients for each enzyme in the *E. coli* metabolic map were assembled to construct a genome-specific stoichiometric matrix. This matrix was used to define the systems characteristics and the capabilities of this organisms metabolism (Edwards and Palsson, 2000). In that report, the authors showed the result of comparisons between the in silico predictions and experimental observations using deletions of genes in the central metabolic pathways. These approaches have now been expanded to the genome-scale reconstruction not only of metabolic network (Forster *et al.*, 2003; Reed *et al.*, 2003), but also of heterogeneous network types including transcription and translation (Shen-Orr *et al.*, 2002; Allen *et al.*, 2003; Gutierrez-Rios *et al.*, 2003). A shift in biology from a component-based perspective to a systems view of the cell is occurring based on genome sequence accumulation and high-throughput post-genomic data generation. Modeling cellular functions according to a systems biology is not new but this approach is now expanding to reach the genome-scale. The total number of genes and biochemical elements that are integrated into a single model has now reached ~2000 (Reed and Palsson, 2003). In parallel, several software environments for the quantitative simulation of cellular processes, including metabolic pathways, based on the numerical integration of rate equations, have been developed (Goryanin *et al.*, 1999; Tomita *et al.*, 1999; Mendes and Kell, 2001; Hucka *et al.*, 2003).

Table 2. Useful websites for *Escherichia coli*

http://ecoli.aist-nara.ac.jp	GenoBase database at Nara Institute of Science and Technology, Japan
http://www.genome.ad.jp	KEGG database at Kyoto University, Japan
http://www.shigen.nig.ac.jp/ecoli/pec/index.jsp	PEC database at the Institute of National Genetics, Japan
http://gib.genes.nig.ac.jp	GIB database at the Institute of National Genetics, Japan
http://genome.gen-info.osaka-u.ac.jp/bacteria/o157	Web site of *E. coli* O157 at Osaka University, Japan
http://www.cifn.unam.mx/Computational_Genomics/regulondb	RegulonDB at Universidad Nacional Autonoma de Mexico, Mexico
http://redpoll.pharmacy.ualberta.ca/CCDB	CyberCell Project at University of Alberta, Canada
http://www.uni-giessen.de/~gx1052/ECDC/ecdc.htm	*E. coli* Database Collection at Justus-Liebig-University, Germany
http://genolist.pasteur.fr/Colibri	Colibri database at the Institute Pasteur, France
http://web.bham.ac.uk/bcm4ght6/res.html	The *E. coli* Index at the University of Birmingham, UK
http://colibase.bham.ac.uk	Colibase dataset at the University of Birmingham, UK
http://us.expasy.org/ch2d	SWISS-2DPAGE in SWISS-PROT database at Swiss Institute of Bioinformatics
http://kr.expasy.org/enzyme	Enzyme database at Swiss Institute of Bioinformatics
http://www.EcoliCommunity.org	*E. coli* Community at Purdue University, USA
http://biocyc.org/ecocyc	EcoCyc database at SRI International, USA
http://www.genome.wisc.edu	Genome Project at University of Wisconsin ñ Madison, USA
http://bmb.med.miami.edu/EcoGene/EcoWeb	EcoGene database at University of Miami School of Medicine, USA
http://www.vetsci.psu.edu/ecoli.cfm	*E. coli* Reference Center at Penn State University, USA
http://www.ncbi.nlm.nih.gov	National Center for Biotechnology Information, USA
http://cgsc.biology.yale.edu	*E. coli* Genetic Stock Center at Yale University, USA
http://www.ecoli.princeton.edu/index.php	*E. coli* Bioinformatics/Resources Initiative at Princeton University, USA
http://genomics.lbl.gov/~ecoreg/index.html	EcoReg, The *Escherichia coli* Regulation Consortium, USA

International Consortium for Large Scale *E. coli* Modeling

The International *E. coli* Alliance (IECA, http://www.EcoliCommunity.org) was formed in November 2002 to tackle the fundamental biological problem in developing the first comprehensive computational model of a living cell. IECAs mission is to consolidate global efforts to understand a living bacterial cell. Scientists around the world are working together to create a complex computer model, integrating all of the dynamic molecular interactions that are required for the life of a simple, self-replicating cell. An *E. coli* cell model will have immediate practical benefits in biology and bioengineering and should significantly contribute to advancing the field of computational systems biology. The generation of a computerized *E. coli* cell will also add powerful new tools to our existing arsenal for functional discovery, including virtual experimentation and mathematical simulation. Ultimately, these biological and computational tools could be useful in both drug discovery and in the design of bioenhanced nanomachines. Furthermore, the development of a virtual system for experimentation on the *E. coli* cell will be extremely useful for understanding more complex cells and contribute to the development and validation of in silico models of human cells and whole multicellular organisms. Biology is now evolving to become a “big science”. The tiny *E. coli* is well positioned to become one of the giant players in the new biology era, based on the determining role it played in the field of molecular genetics.

Useful *E. coli* Websites

As internet technology advances, and the scale of experimental approaches grows exponentially, the importance of biological information and data repository websites cannot be overstated. Some useful websites for *E. coli* biology are listed in Table 2.

Epilog

Robert Hooke first used the term “cell” to describe the basic structural unit of cork in 1665. Biology still has a long way to go for a complete understanding of a cell, even though the complete genetic blueprints are available. In the last four centuries, biology has developed and supported intensive activities based on traditional small-scale research. This type of research will always be important and needs to increase in the future in order to build a more precise quantitative model of biological processes or a cell itself. On the other hand, it is also absolutely true that solving large scale and complex biological networks is far beyond the conventional approach. As previously, biology itself is now standing at a turning point. I hope these new approaches will flower with traditional ones, because these are exactly complementary.

References

Aiba, H., Baba, T., Hayashi, K., Inada, T., Isono, K., Itoh, T., Kasai, H., Kashimoto, K., Kimura, S., Kitakawa, M. *et al.* (1996) A 570-kb DNA sequence of the *Escherichia coli* K-12 genome corresponding to the 28.0-40.1 min region on the linkage map. *DNA Res.* **3**, 363-377.

Allen, T. E., Herrgard, M. J., Liu, M., Qiu, Y., Glasner, J. D., Blattner, F. R. and Palsson, B. O. (2003) Genome-scale analysis of the uses of the *Escherichia coli* genome: model-driven analysis of heterogeneous data sets. *J. Bacteriol.* **185**, 6392-6399.

Arifuzzaman, M. (in preparation).

Baba, T. (in preparation).

Baudin, A., Ozier-Kalogeropoulos, O., Denouel, A., Lacroute, F. and Cullin, C. (1993) A simple and efficient method for direct gene deletion in *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **21**, 3329-3330.

Bergler, H., Hogenauer, G. and Turnowsky, F. (1992) Sequences of the envM gene and of two mutated alleles in *Escherichia coli*. *J. Gen. Microbiol.* **138**, 2093-2100.

Blattner, F. R., Plunkett, G., 3rd, Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F. *et al.* (1997) The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453-1474.

Corbin, R. W., Paliy, O., Yang, F., Shabanowitz, J., Platt, M., Lyons, C. E., Jr, Root, K., McAuliffe, J., Jordan, M. I., Kustu, S., Soupene, E. and Hunt, D. F. (2003) Toward a protein profile of *Escherichia coli*: comparison to its transcription profile. *Proc. Natl. Acad. Sci. USA* **100**, 9232-9237.

Datsenko, K. A. and Wanner, B. L. (2000) One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc. Natl. Acad. Sci. USA* **97**, 6640-6645.

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. USA* **97**, 5528-5533.

Ehrenberg, M., Elf, J., Aurell, E., Sandberg, R. and Tegner, J. (2003) Systems biology is taking off. *Genome Res.* **13**, 2377-2380.

Fiehn, O. (2002) Metabolomics--the link between genotypes and phenotypes. *Plant Mol. Biol.* **48**, 155-171.

Fiehn, O., Kopka, J., Dormann, P., Altmann, T., Trethewey, R. N. and Willmitzer, L. (2000) Metabolite profiling for plant functional genomics. *Nat. Biotechnol.* **18**, 1157-1161.

Figeys, D., Gygi, S. P., Zhang, Y., Watts, J., Gu, M. and Aebersold, R. (1998) Electrophoresis combined with novel mass spectrometry techniques: powerful tools for the analysis of proteins and proteomes. *Electrophoresis* **19**, 1811-1818.

Forster, J., Famili, I., Fu, P., Palsson, B. O. and Nielsen, J. (2003) Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. *Genome Res.* **13**, 244-253.

Fromont-Racine, M., Rain, J. C. and Legrain, P. (1997) Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. *Nat. Genet.* **16**, 277-282.

Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., Cruciat, C. M. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141-147.

Goryanin, I., Hodgman, T. C. and Selkov, E. (1999) Mathematical simulation and analysis of cellular metabolism and regulation. *Bioinformatics* **15**, 749-758.

Gutierrez-Rios, R. M., Rosenblueth, D. A., Loza, J. A., Huerta, A. M., Glasner, J. D., Blattner, F. R. and Collado-Vides, J. (2003) Regulatory network of *Escherichia coli*: consistency between literature knowledge and microarray profiles. *Genome Res.* **13**, 2435-2443.

Gygi, S. P., Rist, B. and Aebersold, R. (2000) Measuring gene expression by quantitative proteome analysis. *Curr. Opin. Biotechnol.* **11**, 396-401.

Hayashi, T., Makino, K., Ohnishi, M., Kurokawa, K., Ishii, K., Yokoyama, K., Han, C. G., Ohtsubo, E., Nakayama, K., Murata, T. *et al.* (2001) Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res.* **8**, 11-22.

Herrgard, M. J., Covert, M. W. and Palsson, B. O. (2003) Reconciling gene expression data with known genome-scale regulatory network structures. *Genome Res.* **13**, 2423-2434.

Hill, C. W. and Harnish, B. W. (1981) Inversions between ribosomal RNA genes of *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **78**, 7069-7072.

Hill, T. M. (1996) in *Escherichia coli* and Salmonella: Cellular and Molecular Biology. Neidhart. F. C. (ed.), ASM Press, Washington, USA.

Horiuchi, T. (in preparation).

Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., Arkin, A. P., Bornstein, B. J., Bray, D., Cornish-Bowden, A. *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19**, 524-531.

Ishino, Y., Shinagawa, H., Makino, K., Amemura, M. and Nakata, A. (1987) Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in *Escherichia coli*, and identification of the gene product. *J. Bacteriol.* **169**, 5429-5433.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. and Sakaki, Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA* **98**, 4569-4574.

Itoh, T., Aiba, H., Baba, T., Hayashi, K., Inada, T., Isono, K., Kasai, H., Kimura, S., Kitakawa, M., Kitagawa, M. *et al.* (1996) A 460-kb DNA sequence of the *Escherichia coli* K-12 genome corresponding to the 40.1-50.0 min region on the linkage map. *DNA Res.* **3**, 379-392.

Itoh, T., Okayama, T., Hashimoto, H., Takeda, J., Davis, R. W., Mori, H. and Gojobori, T. (1999) A low rate of nucleotide changes in *Escherichia coli* K-12 estimated from a comparison of the genome sequences between two different substrains. *FEBS Lett.* **450**, 72-76.

Kanehisa, M., Goto, S., Kawashima, S. and Nakaya, A. (2002) The KEGG databases at GenomeNet. *Nucleic Acids Res.* **30**, 42-46.

Kitagawa, M. M. H. (in preparation).

Kolisnychenko, V., Plunkett, G., 3rd, Herring, C. D., Feher, T., Posfai, J., Blattner, F. R. and Posfai, G. (2002) Engineering a reduced *Escherichia coli* genome. *Genome Res.* **12**, 640-647.

Koonin, E. V., Tatusov, R. L. and Rudd, K. E. (1995) Sequence similarity analysis of *Escherichia coli* proteins: functional and evolutionary implications. *Proc. Natl. Acad. Sci. USA* **92**, 11921-11925.

Kumar, A., Agarwal, S., Heyman, J. A., Matson, S., Heidtman, M., Piccirillo, S., Umansky, L., Drawid, A., Jansen, R., Liu, Y., Cheung, K. H., Miller, P., Gerstein, M., Roeder, G. S. and Snyder, M. (2002) Subcellular localization of the yeast proteome. *Genes Dev.* **16**, 707-719.

Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H. and Brown, E. L. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* **14**, 1675-1680.

Maharjan, R. P. and Ferenci, T. (2003) Global metabolite analysis: the influence of extraction methodology on metabolome profiles of *Escherichia coli*. *Anal. Biochem.* **313**, 145-154.

Mahillon, J. and Chandler, M. (1998) Insertion sequences. *Microbiol. Mol. Biol. Rev.* **62**, 725-774.

Mahillon, J., Leonard, C. and Chandler, M. (1999) IS elements as constituents of bacterial genomes. *Res. Microbiol.* **150**, 675-687.

Masuda, N. and Church, G. M. (2003) Regulatory network of acid resistance genes in *Escherichia coli*. *Mol. Microbiol.* **48**, 699-712.

Mendes, P. and Kell, D. B. (2001) MEG (Model Extender for Gepasi): a program for the modellingmodeling of complex, heterogeneous, cellular systems. *Bioinformatics* **17**, 288-289.

Mori, H., Isono, K., Horiuchi, T. and Miki, T. (2000) Functional genomics of *Escherichia coli* in Japan. *Res Microbiol.* **151**, 121-128.

Nakata, A., Amemura, M. and Makino, K. (1989) Unusual nucleotide arrangement with repeated sequences in the *Escherichia coli* K-12 chromosome. *J. Bacteriol.* **171**, 3553-3556.

O'Farrell, P. H. (1975) High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.* **250**, 4007-4021.

Oshima, T., Aiba, H., Baba, T., Fujita, K., Hayashi, K., Honjo, A., Ikemoto, K., Inada, T., Itoh, T., Kajihara, M. *et al.* (1996) A 718-kb DNA sequence of the *Escherichia coli* K-12 genome corresponding to the 12.7-28.0 min region on the linkage map. *DNA Res.* **3**, 137-155.

Oshima, T., Aiba, H., Masuda, Y., Kanaya, S., Sugiura, M., Wanner, B. L., Mori, H. and Mizuno, T. (2002) Transcriptome analysis of all two-component regulatory system mutants of *Escherichia coli* K-12. *Mol. Microbiol.* **46**, 281-291.

Perna, N. T., Plunkett, G., 3rd, Burland, V., Mau, B., Glasner, J. D., Rose, D. J., Mayhew, G. F., Evans, P. S., Gregor, J., Kirkpatrick, H. A. *et al.* (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* **409**, 529-533.

Raamsdonk, L. M., Teusink, B., Broadhurst, D., Zhang, N., Hayes, A., Walsh, M. C., Berden, J. A., Brindle, K. M., Kell, D. B., Rowland, J. J., Westerhoff, H. V., van Dam, K. and Oliver, S. G. (2001) A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nat. Biotechnol.* **19**, 45-50.

Reed, J. L. and Palsson, B. O. (2003) Thirteen years of building constraint-based in silico models of *Escherichia coli*. *J. Bacteriol*. **185**, 2692-2699.

Reed, J. L., Vo, T. D., Schilling, C. H. and Palsson, B. O. (2003) An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biol*. **4**, R54.

Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M. and Seraphin, B. (1999) A generic protein purification method for protein complex characterization and proteome exploration. *Nat. Biotechnol*. **17**, 1030-1032.

Riley, M. and Labedan, B. (1997) Protein evolution viewed through *Escherichia coli* protein sequences: introducing the notion of a structural segment of homology, the module. *J. Mol. Biol*. **268**, 857-868.

Ross-Macdonald, P., Coelho, P. S., Roemer, T., Agarwal, S., Kumar, A., Jansen, R., Cheung, K. H., Sheehan, A., Symoniatis, D., Umansky, L. *et al.* (1999) Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature* **402**, 413-418.

Rudd, K. E. (1999) Novel intergenic repeats of *Escherichia coli* K-12. *Res. Microbiol.* **150**, 653-664.

Sabatti, C., Rohlin, L., Oh, M. K. and Liao, J. C. (2002) Co-expression pattern from DNA microarray experiments as a tool for operon prediction. *Nucleic Acids Res.* **30**, 2886-2893.

Sharples, G. J. and Lloyd, R. G. (1990) A novel repeated DNA sequence located in the intergenic regions of bacterial chromosomes. *Nucleic Acids Res.* **18**, 6503-6508.

Shen-Orr, S. S., Milo, R., Mangan, S. and Alon, U. (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.* **31**, 64-68.

Smith, G. R. (1988) Homologous recombination in procaryotes. *Microbiol. Rev.* **52**, 1-28.

Soga, T., Ohashi, Y., Ueno, Y., Naraoka, H., Tomita, M. and Nishioka, T. (2003) Quantitative metabolome analysis using capillary electrophoresis mass spectrometry. *J. Proteome Res.* **2**, 488-494.

Soga, T., Ueno, Y., Naraoka, H., Matsuda, K., Tomita, M. and Nishioka, T. (2002a) Pressure-assisted capillary electrophoresis electrospray ionization mass spectrometry for analysis of multivalent anions. *Anal. Chem.* **74**, 6224-6229.

Soga, T., Ueno, Y., Naraoka, H., Ohashi, Y., Tomita, M. and Nishioka, T. (2002b) Simultaneous determination of anionic intermediates for Bacillus subtilis metabolic pathways by capillary electrophoresis electrospray ionization mass spectrometry. *Anal. Chem*. **74**, 2233-2239.

Storz, G. (2002) An expanding universe of noncoding RNAs.

Science **296**, 1260-1263.

Tao, H., Bausch, C., Richmond, C., Blattner, F. R. and Conway, T. (1999) Functional genomics: expression analysis of *Escherichia coli* growing on minimal and rich media. *J. Bacteriol.* **181**, 6425-6440.

Tatusov, R. L., Koonin, E. V. and Lipman, D. J. (1997) A genomic perspective on protein families. *Science* **278**, 631-637.

Tomita, M., Hashimoto, K., Takahashi, K., Shimizu, T. S., Matsuzaki, Y., Miyoshi, F., Saito, K., Tanida, S., Yugi, K., Venter, J. C. and Hutchison, C. A., 3rd. (1999) E-CELL: software environment for whole-cell simulation. *Bioinformatics* **15**, 72-84.

Tweeddale, H., Notley-McRobb, L. and Ferenci, T. (1998) Effect of slow growth on metabolism of *Escherichia coli*, as revealed by global metabolite pool ("metabolome") analysis. *J. Bacteriol*. **180**, 5109-5116.

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P. *et al.* (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae. Nature* **403**, 623-627.

Van Dyk, T. K., Wei, Y., Hanafey, M. K., Dolan, M., Reeve, M. J., Rafalski, J. A., Rothman-Denes, L. B. and LaRossa, R. A. (2001) A genomic approach to gene fusion technology. *Proc. Natl. Acad. Sci. USA* **98**, 2555-2560.

Vanbogelen, R. A. (1996) in *Escherichia coli* and Salmonella: Cellular and Molecular Biology. ed. Neidhart. F. C.

Wada, A., Koyama, K., Maki, Y., Shimoi, Y., Tanaka, A. and Tsuji, H. (1993) A 5 kDa protein (SCS23) from the 30S subunit of the spinach chloroplast ribosome. *FEBS Lett.* **319**, 115-118.

Yamamoto, Y., Aiba, H., Baba, T., Hayashi, K., Inada, T., Isono, K., Itoh, T., Kimura, S., Kitagawa, M., Makino, K. *et al.* (1997) Construction of a contiguous 874-kb sequence of the *Escherichia coli*-K12 genome corresponding to 50.0-68.8 min on the linkage map and analysis of its sequence features. *DNA Res.* **4**, 91-113.

Yu, B. J., Sung, B. H., Koob, M. D., Lee, C. H., Lee, J. H., Lee, W. S., Kim, M. S. and Kim, S. C. (2002) Minimization of the *Escherichia coli* genome using a Tn5-targeted Cre/loxP excision system. *Nat. Biotechnol*. **20**, 1018-1023.

Zhu, H., Bilgin, M., Bangham, R., Hall, D., Casamayor, A., Bertone, P., Lan, N., Jansen, R., Bidlingmaier, S., Houfek, T., Mitchell, T., Miller, P., Dean, R. A., Gerstein, M. and Snyder, M. (2001) Global analysis of protein activities using proteome chips. *Science* **293**, 2101-2105.