[SI-1]

Functional Genomics in the Context of Biocatalysis and Biodegradation

Sung-Cheol Koh* and Byung-Hyuk Kim

Division of Civil and Environmental Systems Engineering Korea Maritime University, Busan 606-791, Korea

Abstract

Functional genomics aims at uncovering useful information carried on genome sequences and at using it to understand the mechanisms of biological function. Elucidating the unknown biological functions of new genes based upon the genomics rationales will greatly speed up the extensive understanding of biocatalysis and biodegradation in biological world including microorganisms. DNA microarrays generate a system for the simultaneous measurement of the expression level of thousands of genes in a single hybridization assay. Their data mining (transcriptome) strategy has two categories: differential gene expression and coordinated gene expression. Furthermore, measurement of proteins (proteome) generates information on how the transcribed sequences end up as functional characteristics within the cell, and quantitation of metabolites yields information on how the functional proteins act to produce energy and process substrates (metabolome). Various composite functional genomics databases containing genetic, enzymatic and metabolic information have been developed and will contribute to the understanding of the life blue print and the new discoveries and practices in biocatalysis and biodegradation that could enrich their industrial and environmental applications.

Introduction

Genomics or the complete sequencing and analysis of an organism's DNA, will be transforming the way in which the biologists approach their sciences. While DNA sequence analysis is yielding important patterns, the greatest genomic richness results from assigning metabolic functions to individual genes and deriving a biological usefulness of the gene in the context of the organism and its environment (Wackett and Hershberger, 2000). Functional genomics aims at uncovering useful information carried on genome sequences and at using it to understand the mechanisms of biological function. Elucidating the unknown biological functions of new genes based upon the genomics rationales will greatly speed up the extensive understanding of biocatalysis and biodegradation in biological world including microorganisms.

In this paper, we will present an overview of the application of DNA microarray technologies to the studies of gene expression in microbial biocatalysis and biodegradation systems. The understanding these gene expression systems will greatly facilitate the development of technologies to resolve the issues of various biotechnology-related industries including medicine, agriculture, food, environmental biotechnology-related businesses.

Functional genomics: concept and its technologies

Traditionally, the analysis of the regulation and function of genes has largely been performed by sequential studies of individual genes and proteins. In the last decade, however, a paradigm shift has been observed in which we may be able to handle many different genes in a highly parallel and rapidly serialized way: the development of DNA microarrays. These arrays consist of a highly ordered matrix of thousands of different DNA sequences that can be used to measure DNA and RNA variation in applications that include gene expression profiling, comparative genomics and genotyping (Case-Green *et al*, 1998; Lander, 1999; Brown and Botstein, 1999; Ferea and Brown, 1999). DNA microarrays generate a system for the simultaneous measurement of the expression level of thousands of genes in a single hybridization assay (Harrington *et al.*, 2000). Fluorescently labeled RNA or DNA prepared from messenger RNA is hybridized to complementary DNA on the array and then detected by laser scanning. Hybridization intensities for each DNA sequence on the array are determined using an automated process and converted to a quantitative read-out of relative gene expression levels. The data can then be further analyzed to identify expression patterns and variation that correlate with cellular development, physiology and function. A description of the typical microarray system is shown in the Figure 1.

Following acquisition and processing of the fluorescent array image, there are three essential steps to meet for efficient and effective data analysis: data normalization, data filtering, and pattern identification. The methods used for data mining and interpretation are varied, ranging from simple lists of increased and decreased genes based on user-defined thresholds to the implementation of sophisticated clustering and visualization programs, such as hierarchical clustering (Eisen *et al.*, 1998) and self-organizing maps also called k-means clustering (Tamayo *et al.*, 1999; Törönen *et al.*, 1999; Tavazoie *et al.*, 1999) .

The data mining strategy used depends on the experimental design and can be generally divided into two categories: differential gene expression and coordinated gene expression (Claverie, 1999). The differential gene expression approach generally consists of paired comparisons between normal/abnormal samples such as from healthy and pathological specimens or wild-type and mutant
genotypes. Coordinated gene expression analysis involves the assessment of the expression levels of a large number of genes over a period of time or through a series of experimental conditions,

Functional genomics in the area of biocatalysis and biodegradation

Enzymes contain the largest and most diverse group of all proteins, catalyzing various chemical reactions in the metabolism of all organisms. The completion of the human genome sequencing and those of many other organisms, including several dozens of bacteria, speeded up post-genome projects aiming at elucidating the blueprint of life from a scientific perspective. They are also targeting at uncovering new drugs and other useful materials, and at finding out biodegradation pathways of xenobiotic chemicals such as pollutants and toxins from medical, industrial and environmental aspects. All of them, however, require chemical information, which is not found in the genome, in addition to information about genes and proteins, which is derived from the

genome, and chembioinformatics has been considered as one of the important research fields in the post-genome age.

Biocatalysis

The study of secondary metabolites that organisms such as microbes and plants have evolved, largely for the purpose of their own survival, has historically turned out to be immense benefit in drug discovery, providing a rich source of structurally novel bioactive molecules, many of which have become life-saving drugs (Nisbet and Moore, 1997). The exploitation of structural chemical databases comprising a wide variety of chemotypes, in conjunction with databases on target genes and proteins, will facilitate the creation of new chemical species through computational molecular modeling for pharmacological evaluation.

Measurement of mRNA provides information on what genetic information is transcribed (the transcriptome), measurement of proteins (the proteome) yields information on how transcribed sequences end up as functional entities within the cell, and measurement of metabolites gives information on how the functional proteins act to produce energy and process materials (the metabolome) (Delneri *et al.*, 2001).

Frishman *et al.* (2001) introduced the PEDANT software system for high-throughput analysis of large biological sequence sets and the genome analysis server associated with it. The principal features of PEDANT are: (i) completely automatic processing of data using a wide range of bioinformatics methods, (ii) manual refinement of annotation, (iii) automatic and manual assignment of gene products to a number of functional and structural categories, (iv) extensive hyperlinked protein reports, and (v) advanced DNA and protein viewers. The main purpose of the PEDANT genome database has been to spread well-organized information on completely sequenced and unfinished genomes. In particular, the availability of structural prediction data for over 300 000 genomic proteins makes PEDANT one of the most extensive structural genomics resources available on the web (Figure 2).

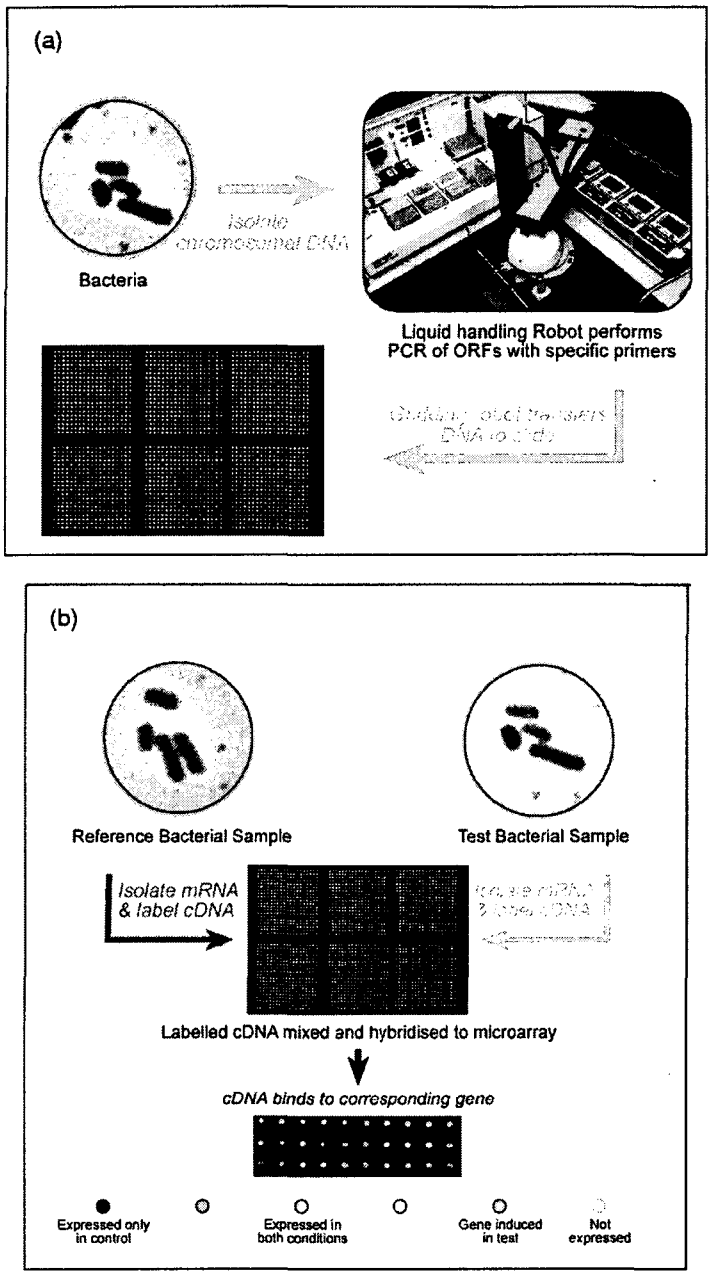

Figure 1. Making (a) and using (b) DNA microarray for gene expression profiling (http://www. microarrays.org)

The LIGAND (Goto *et al.*, 2002) is a composite database composed of three sections: COMPOUND for the information about metabolites and other chemical compounds, REACTION for the collection of substrate-product relations representing metabolic and other reactions, and ENZYME for the information about enzyme molecules (Figure 3). This has been established to fill in the gap between genomic information and chemical information, and applied to actual reconstruction of metabolic pathways in the completely sequenced organisms in the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Goto *et al.*, 1998; Kanehisa, 1997; Kanehisa *et al.*, 2002). The current version of COMPOUND includes xenobiotic chemicals such as environmental pollutants and toxins, because KEGG has an agreement with UM-BBD to include biodegradation pathways of xenobiotic

chemicals in KEGG/PATHWAY (Ellis *et al.*, 2001).

BRENDA (BRaunschweig Enzyme DAtabase), launched in 1987 by Dietmar Schomburg, is a comprehensive protein function database, containing enzymatic and metabolic information derived from the primary literatures (Schomburg *et al.*, 2002). Currently, the database carries data on more than 40 000 enzymes and 4460 different organisms, and includes information about enzyme–ligand relationships with numerous chemical compounds.

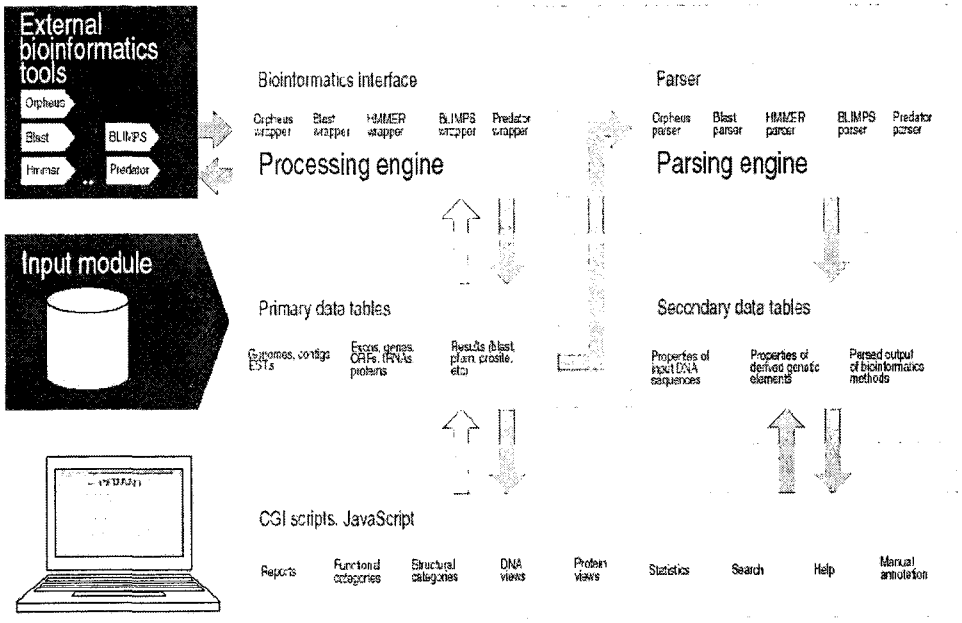

Figure 2. The PEDANT architecture, a genome analysis server, is available at http://pedant.mips.biochem.mpg.de (Frishman *et al.*, 2001)

MetaCyc is a metabolic-pathway database that annotates 445 pathways and 1115 enzymes occurring in 158 organisms (Karp *et al.*, 2002). Applications of MetaCyc are pathway analysis of genomes, metabolic engineering and biochemistry education. The modification of a metabolic map through genetic engineering involves (i) inserting a new enzyme or pathway into an organism, (ii) replacing an existing enzyme or pathway with a substitute or (iii) removing an enzyme or pathway.

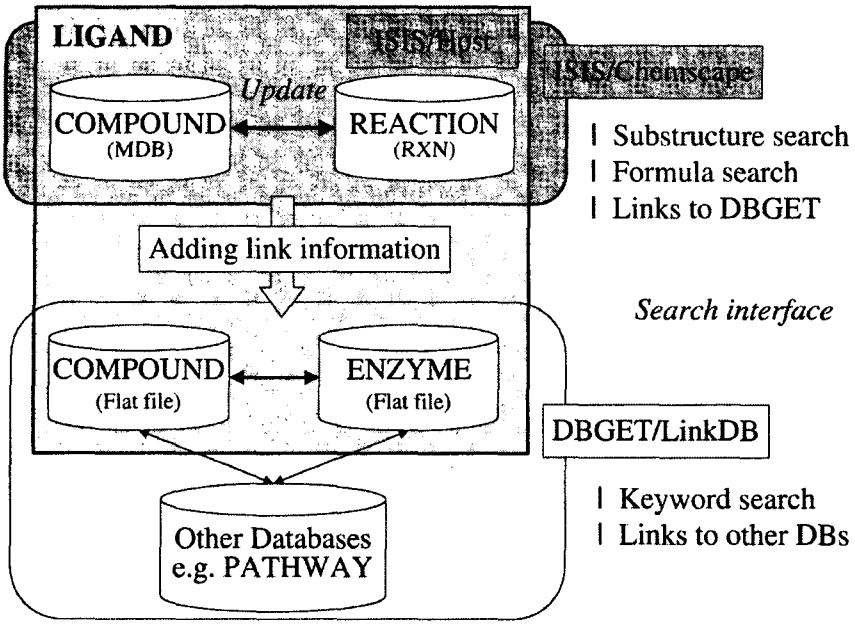

Figure 3. Relationship between the ISIS version and the DBGET version of LIGAND (Goto *et al.*, 2002).

The University of Minnesota Biocatalysis/Biodegradation Database (UM-BBD, http://www.labmed.umn.edu/umbbd/index.html) first became available on the web in 1995 to provide information on microbial biocatalytic reactions of, and biodegradation pathways for, organic chemical compounds, especially those produced by humans (Figure 4) (Ellis *et al.*, 2002). The database includes the diversity of known microbial metabolic routes, organic functional groups, and environmental conditions under which biodegradation occurs. The database could contribute to enhance understanding of basic biochemistry, biocatalysis leading to specialty chemical manufacture, and biodegradation of environmental pollutants. It will be also a resource for functional genomics, since it contains information on enzymes and genes involved in specialized metabolism not found in intermediary metabolism databases, and thus can assist in assigning functions to genes homologous to such less common genes (Figure 5) (Ellis *et al.*, 2002).

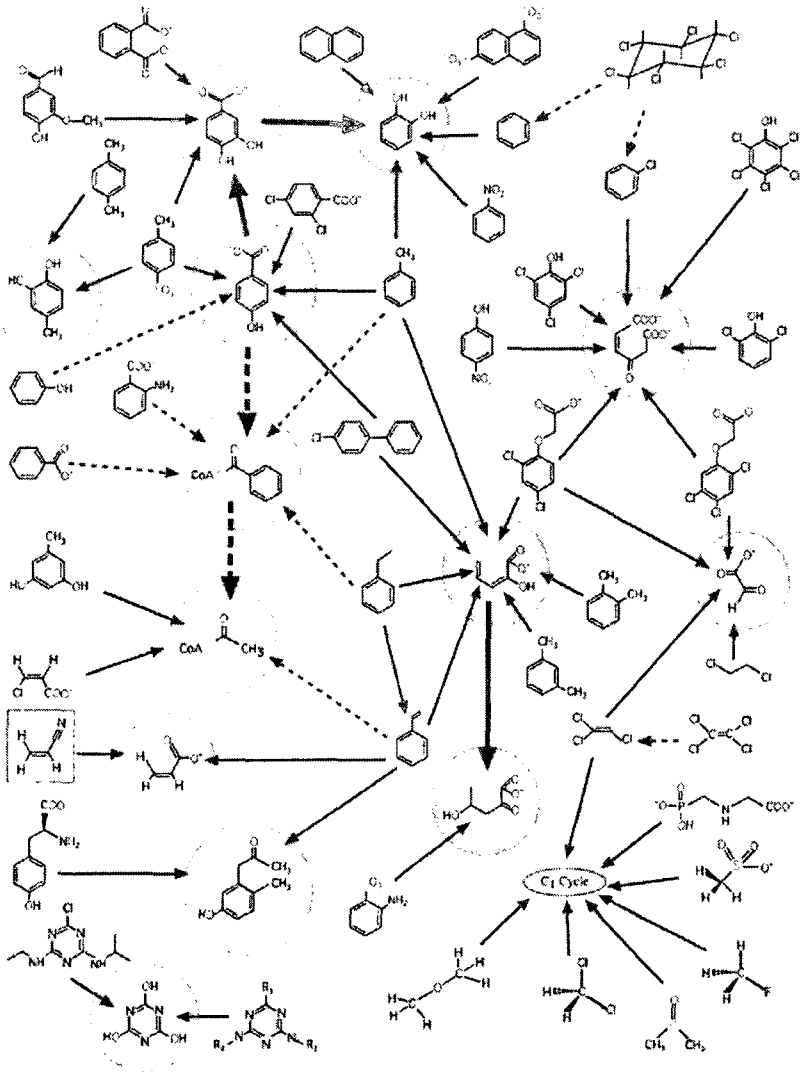

Figure 4. Graphical overview of UM-BBD content. The circled compounds lead to intermediary metabolism (Ellis *et al.*, 2002).

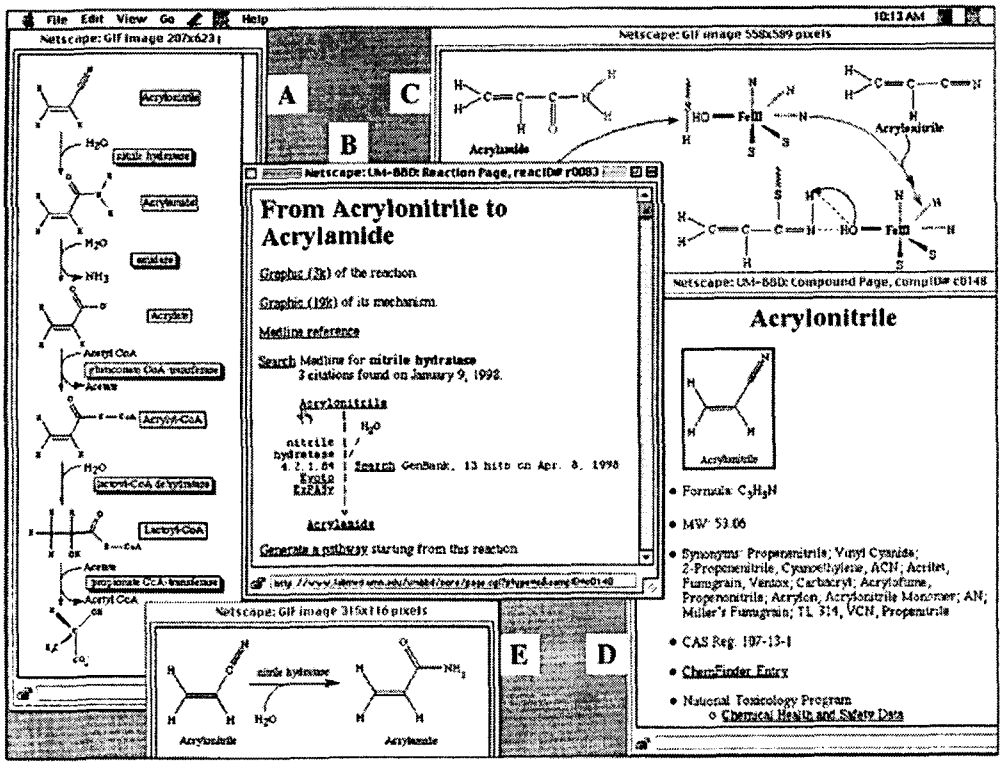

Figure 5. Example UM-BBD pathway information. (A) Graphical pathway map for the acrylonitrile pathway (B) reaction page for the nitrile hydrolase reaction; (C) reaction mechanism graphic for nitrile hydrolase (D) compound page for acrylonitrile; (E) reaction graphic for nitrile hydrolase. URL http://www.labmed.umn.edu/umbbd/acr/acr_map.html (Ellis *et al.*, 2002).

Biodegradation

Although DNA microarrays have been initially developed for medical and diagnostic applications, they are also an ideal technology to assess the sequence diversity of 16S rRNA samples from the environments. Soils are an immense source of microbial diversity, which mostly remains unexplored. Some novel methods utilizing rRNA and rDNA sequence analyses have uncovered a portion of the soil microbial diversity. An advanced step for the microbial ecological studies will be to elucidate genomic, evolutionary and functional information from microbial artificial chromosome libraries of the soil community genomes (*i.e.*, the metagenome). Sophisticated analyses that are based upon molecular phylogenetics, DNA microarrays, functional genomics and *in situ* activity measurements will provide a large amount of new data, potentially enhancing our understanding of the structure and function of soil microbial ecosystems, and the interactions that occur within them (Phelps *et al.*, 2002; Torsvik and Øvreås, 2002).

Studies of the natural diversity and abundance of sulfate-reducing bacteria (SRB) in relationship to hydrocarbon degradation could be greatly facilitated by application of microarray technology. In order to assess the utility of the microarray format for analysis of environmental samples, oil-contaminated sediments from the coast of Kuwait were analyzed (Koizumi *et al.*, 2002). Here the DNA microarray successfully detected bacterial nucleic acids from these samples, but probes targeting specific groups of SRB did not give positive signals. The results of this study demonstrate the limitations and the potential utility of DNA microarrays for microbial community analysis.

To determine the potential of DNA array technology for assessing functional gene diversity and distribution, a prototype microarray was constructed with genes involved in nitrogen cycling: nitrite reductase (*nirS* and *nirK*) genes, ammonia mono-oxygenase (*amoA*) genes, and methane mono-oxygenase (*pmoA*) genes from pure cultures and those cloned from marine sediments (Wu *et al.*, 2001). Here hybridization signal intensity within a certain range of sequence identity and size was affected by sequence divergence and probe length, respectively. The prototype functional gene array did show differences in the apparent distribution of *nir* and *amo*A and *pmo*A gene families in sediment and soil samples. These results indicate that the glass-based microarray hybridization technique has potential as a tool to reveal functional gene composition in natural microbial communities. However, more efforts are needed to improve sensitivity and quantitation and to understand the issues of specificity.

For cultivation-independent detection of sulfate-reducing prokaryotes (SRPs) an oligonucleotide microarray consisting of 132 16S rRNA gene-targeted oligonucleotide probes (18-mers)) was designed in the way that it had hierarchical and parallel (identical) specificity for the detection of all known lineages of sulfate-reducing prokaryotes (SRP-PhyloChip and subsequently evaluated with 41 suitable pure cultures of SRPs (Loy *et al.*, 2002). Consistent with previous studies, the SRP-PhyloChip showed the occurrence of *Desulfomicrobium* spp. in the tooth pockets and the presence of *Desulfonema*- and *Desulfomonile*-like SRPs (together with other SRPs) in the chemocline of the mat.

Even if whole genomic DNA-DNA hybridization technique has an advantage in microbial species determination, it is not widely used because of its laborious implementation procedure. Cho and Tiedje (2001) have developed a method based on random genome fragments and DNA microarray technology that avoid the

disadvantages of whole-genome DNA-DNA hybridization. The cluster analysis of the hybridization profiles using this technique elucidated taxonomic relationships between bacterial strains tested at resolution of species to strain level, indicating that this approach is useful for the bacterial identification as well as determining the genetic distance among bacteria. To quantify target genes in biological samples using DNA microarrays, Cho and Tiedje (2002) recently have employed reference DNA to normalize variations in spot size and hybridization This approach for designing quantitative microarrays and the equation derived from the study could provide a simple and convenient way to estimate the target gene concentration from the hybridization signal ratio.

Data Analysis

Artificial neural networks (ANNs) may provide a useful tool for recognizing patterns in complex, nonlinear data sets such as those associated with predicting gene expression patterns from the DNA microarray data (Urakawa *et al.*, 2002). ANNs will be particularly advantageous over the conventional statistical methods because they can deal with the inherent variability associated with biological data. The self-organizing map (SOM) is an unsupervised neural network protocol that is particularly useful for data visualization (Kohonen, 1997). The SOM algorithm simultaneously allows a representative set of reference vectors of the training data and positions them on a regular two-dimensional grid of neurons so that it can easily visualized (Wang, et al., 2001).

Conclusions and Perspectives

From the standpoints of medicine, agriculture, and food industries, metabolic engineering has been directed primarily to useful metabolite overproductions. The parallel and recent advances in the resolution and acquisition time of biological data, especially structural and functional genomics, will greatly contribute to the systemic view of biology of which the metabolic engineering may take advantage. Moreover, microorganisms play a significant role in ecosystem function and sustainability. Dissecting the structure and composition of microbial communities and their responses and adaptations to environmental perturbations such as toxic pollutants, agricultural and industrial practices will be critically important in finding ways of working out these complicated ecological and environmental issues. The functional genomics connected by large-scale enzyme databases will greatly facilitate new discoveries and practices in biocatalysis and biodegradation.

References

1. Brown, P.O., and D. Botstein. 1999. Exploring the new world of the genome with DNA microarrays. Suppl. Nat. Genet. 21, 33-37.
2. Case-Green, S.C., K.U. Mir, C.E. Pritchard., and E.M. Southern. 1998. Analyzing genetic information with DNA arrays. Curr. Opin. Chem. Biol. 404-410.
3. Cho, J.C. and J.M. Tiedje. 2001. Bacterial Species Determination from DNA-DNA Hybridization by Using Genome Fragments and DNA Microarrays. Appl. Envir. Microbiol. 67, 3677-3682.
4. Cho, J.C. and J.M. Tiedje. 2002. Quantitative Detection of Microbial Genes by Using DNA Microarrays. Appl. Envir. Microbiol. 68, 1425-1430.

5. Claverie, J.M. 1999. Computational methods for the identification of differential and coordinated gene expression. Hum. Mol. Genet. 8, 1821-1832.

6. Delneri, D., F.L. Brancia., and S.G. Oliver. 2001. Towards a truly integrative biology through the functional genomics of yeast. Curr. Opin. Biotechnol.12, 87-91

7. Eisen, M.B., P.T. Spellman., P.O. Brown., D. Botstein. 1998. Cluster analysis and display of genome-wide expression patterns. Pro.c Natl. Acad. Sci. USA. 95, 14863-14868.

8. Ellis, L.B.M., C.D. Hershberger., E.M. Bryan., and L.P. Wackett. 2001. The University of Minnesota Biocatalysis/Biodegradation Database: emphasizing enzymes. Nucleic Acids Res. 29, 340–343.

9. Ellis, L.B.M., C.D. Hershberger., and L.P. Wackett. 2002. The University of Minnesota Biocatalysis/Biodegradation Database: specialized metabolism for functional genomics. Nucleic Acids Res. 27, 373-376

10. Ferea, T.L., and P.O. Brown. 1999. Observing the living genome. Curr. Opin. Genet. Dev. 9, 715-722.

11. Frishman, D., K. Albermann., J. Hani., K. Heumann., A. Metanomski., A. Zollner., and H.W. Mewes. 2001. Functional and structural genomics using PEDANT. *Bioinformatics*. 17, 44-57

12. Goto, S., T. Nishioka., and M. Kanehisa. 1998. LIGAND: chemical database for enzyme reactions. *Bioinformatics*. 14, 591–599.

13. Goto, S., Y. Okuno., M. Hattori., T. Nishioka., and M. Kanehisa. 2002. LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Research*. 30, 402-404

14. Harrington, C.A., A. Rosenow, and J. Retief. 2000. Monitoring gene expression using DNA microarrays. *Curr. Opin. Microbiol*. 3, 285-291

15. Kanehisa, M. 1997. A database for post-genome analysis. *Trends Genet*. 13, 375–376.

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

17. Karp, P.D., M. Riley., S.M. Paley., and A.P. Toole. 2002. The MetaCyc Database. *Nucleic Acids Res*. 30, 59-61

18. Kohonen, T. 1997. Self-organizing maps. 2nd extended edition. Springer, Berlin.

19. Koizumi, Y., J.J. Kelly., T. Nakagawa., H. Urakawa., E.F. Saïd., A.M. Saleh., M. Fukui., Y. Urushigawa., and D.A. Stahl. 2002. Parallel Characterization of Anaerobic Toluene- and Ethylbenzene-Degrading Microbial Consortia by PCR-Denaturing Gradient Gel Electrophoresis, RNA-DNA Membrane Hybridization, and DNA Microarray Technology. *Appl. Envir. Microbiol*. 68, 3215-3225.

20. Lander, E. 1999. Array of hope. *Suppl. Nat. Genet*. 21, 3-4.

21. Loy, A., A. Lehner., N. Lee., J. Adamczyk., H. Meier., J. Ernst., K.H. Schleifer., and M. Wagner. 2002. Oligonucleotide Microarray for 16S rRNA Gene-Based Detection of All Recognized Lineages of Sulfate-Reducing Prokaryotes in the Environment *Appl. Envir. Microbiol*. 68, 5064-5081.

22. Nisbet, L. and M. Moore. 1997. Will natural products remain an important source of drug research for the future? *Curr. Opin. Biotechnol*. 8, 708-712

23. Øvreås, L. 2000. Population and community level approaches for analysing microbial diversity in natural environments. *Ecol. Letts*. 3, 236-251.

24. Phelps, T.J., A.V. Palumbo., and A.S. Beliaev. 2002. Metabolomics and microarrays for improved understanding of phenotypic characteristics controlled by both genomics and environmental constraints. *Curr. Opin. Biotechnol.* 13, 20-24

25. Schomburg, I., A. Chang., O. Hofmann., C. Ebeling., F. Ehrentreich., and D. Schomburg. 2002. BRENDA: a resource for enzyme data and metabolic informationl. *Trend. Biochem.Sci.* 27, 54-56

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

27. Tavazoie, S., J.D. Hughes., M.J. Campbell., R.J. Cho., and G.M. Church. 1999. Systematic determination of genetic network architecture. *Nat Genet.* 22, 281-285.

28. Torsvik, V., and L. Øvreås. 2002. Microbial diversity and function in soil: from genes to ecosystems. *Curr. Opin. Microbiol.* 5, 240-245

29. Törönen, P., M. Kolehmainen., G. Wong., and E. Castrën. 1999. Analysis of gene expression data using self-organizing maps. *FEBS Lett.* 451, 142-146.

30. Urakawa, H., P.A. Noble, S.E. Fantroussi, J.J. Kelly, and D.A. Stahl. 2002. Single-base-pair discrimination of terminal mismatches by sing oligonucleotide microarrays and neural network analyses. *Appl. Envir. Microbiol.* 68, 235-244.

31. Wackett, L.P. and C.D. Hershberger. 2001. Biocatalysis and Biodegradation: Microbial transformation of organic compounds. American Soc. Microbiol., Washington, D.C.

32. Wang, H-C., J. Badger, P. Kearney, and Ming Li. 2001. Analysis of codon usage patterns of bacterial genomes using the self-organizing map. *Mol. Biol. Evol.* 18, 792-800

33. Wu, L., D.K. Thompson., G.H. Li., R.A. Hurt., J.M. Tiedje., and J. Zhou. 2001. Development and Evaluation of Functional Gene Arrays for Detection of Selected Genes in the Environment *Appl. Envir. Microbiol.* 67, 5780-5790.