British Journal of Nutrition (2005), **94**, 493–495
DOI: 10.1079/BJN20051536

https://doi.org/10.1079/BJN20051536 Published online by Cambridge University Press

Short communication

A nutrigenomics database – integrated repository for publications and associated microarray data in nutrigenomics research

Kenji Saito[1], Soichi Arai[2] and Hisanori Kato[1]*

[1]*Department of Applied Biological Chemistry, Graduate School of Agriculture and Life Sciences, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo, Japan*

[2]*Department of Nutritional Science, Tokyo University of Agriculture, 1-1-1 Sakuragaoka, Setagaya-ku, Tokyo, Japan*

(Received 14 November 2004 – Revised 17 May 2005 – Accepted 6 June 2005)

In the current situation where microarray data in the field of nutritional genomics (nutrigenomics) are accumulating rapidly, there is imminent need for an efficient data infrastructure to support research workflow. We have established a web-based, integrated database of the publications and microarray expression data in the field of nutrigenomics. The registered data include links to external databases such as PubMed of the National Center for Biotechnology Information and public microarray databases that contain Minimum Information About a Microarray Experiment-compliant microarray expression data. Using this database, all data sets created will be effectively utilized and shared with other researchers. This database is built on an open-source database system and is freely accessible via the World Wide Web (http://a-yo5.ch.a.u-tokyo.ac.jp/index.phtml).

Microarray: Database: Nutrigenomics

Gene expression analysis using high-throughput technology is becoming more and more popular in a variety of research fields (Goodman & Lory, 2004; Mirnics & Pevsner, 2004; Straume, 2004). Of all the methods for analysing differential gene expression, the DNA microarray technique is the most powerful because of its ability to simultaneously monitor the expression of several thousand genes (Schena *et al.* 1995). In the field of nutritional research, one principal application of DNA microarrays is their use as a comprehensive screening tool for identifying the genes regulated by nutrients. With microarray technologies, many scientists have already demonstrated that the expression of a host of genes is differentially controlled by nutritional status and various nutrient stimuli (Weindruch *et al.* 2001; Endo *et al.* 2002; Sreekumar *et al.* 2002). The new term, 'nutrigenomics', was coined for the area of study targeting the genome-wide influence of dietary signals, and this field has proved highly promising as the next generation of nutrition research. At present, nutrigenomics research remains in its infancy and is saddled with a variety of problems (Kato & Kimura, 2003; Muller & Kersten, 2003). Nevertheless, the massive expansion of microarray studies has caused this research field to develop much faster than originally expected. Hence, the building of an adequate information infrastructure to support research workflow is one of the prime tasks for the community of nutrigenomics researchers.

Each microarray experiment produces a vast amount of data. The rapid accumulation of microarray data in recent years has created problems with regard to data organization, storage and analysis. These problems have primarily been overcome with advances in computational biology, including the advent of a variety of microarray databases and bioinformatic tools (Gardiner-Garden & Littlejohn, 2001; Quackenbush, 2001). However, there are still considerable obstacles with regard to data accessibility and comparability. For example, the array data are distributed in different data formats over several public repositories such as the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo) of the National Center for Biotechnology Information (NCBI) and ArrayExpress (http://www.ebi.ac.uk/arrayexpress) of the European Bioinformatics Institute, as well as hundreds of individual laboratory servers. As a consequence, access to all published data has been crucial for data collection and further analysis. The public availability of gene expression data, especially of raw microarray data, offers another consideration. Currently, the raw array data obtainable through public repositories are quite limited as compared with the literature sources, exemplified by PubMed. This limitation is primarily due to many academic journals still not requiring that a complete or summary subset of array data be deposited at the time of submission. In addition, one of the major public microarray databases does not hold image data (raw data) that would provide reliable information regarding the quality of the analysis; such information enables one to check

Abbreviations: BASE, BioArray Software Environment; MIAME, Minimum Information About a Microarray Experiment; NCBI, National Center for Biotechnology Information.

* **Corresponding author:** Dr Hisanori Kato, fax +81 3 5841 5114, email akatoq@mail.ecc.u-tokyo.ac.jp

https://doi.org/10.1079/BJN20051536 Published online by Cambridge University Press

possible contamination as to each reporter and to re-normalize values with different normalization factors.

Now that efficient utilization of the microarray data of nutrigenomics research is a pressing issue, a centralized nutrigenomics database capable of accumulating and managing information is indispensable. We have therefore established an integrated database, the 'Nutrigenomics Database', which allows us to share the publications and gene expression information relating to nutrigenomic research. As a primary phase of the study, information from over 200 publications, some of which are accompanied by links to associated array data (both provided by NCBI in the public domain at www.ncbi.nlm.nih.gov), has been collected and organized in the database. All information in the database is freely available and helpful for food and nutrition scientists concerned with global expression data.

Methods

The nutrigenomics database was developed using the BioArray Software Environment (BASE) system (Saal *et al.* 2002), a free web-accessible database released under the GNU General Public License. BASE is a customizable database that runs on GNU/Linux (http://www.linux.org/), written in PHP language (http://www.php.net/) with a MySQL (http://www.mysql.com/) backend. The organization and interface of the original BASE system have been designed for the efficient management of microarray productions, but it is not for published articles and manually created annotations. So we modified the source codes of the BASE software (version 1.12.10) in several ways considering the needs of nutrigenomics researchers. Among these modifications, the following are the major ones:

(1) Addition and reassembling of the tables to manage the collected and summarized publication information;
(2) The addition of hyperlinks to public microarray databases and NCBI's PubMed; and
(3) The addition of guest login mode to be a public database and to improve the usability and accessibility of the website.

The database is available at http://a-yo5.ch.a.u-tokyo.ac.jp/index.phtml.

Results and discussion

The database contents include publications derived from nutrigenomics research and a limited number of associated microarray data at the moment. To browse the list of data, all users must log in to the database either with a username and password combination that is authorized or as our 'guest' without a password. Guest login is restricted from using extended features and from referring to the detailed information, but is more user-friendly than registered use as minimal database operations are needed. Clicking 'Experiments' at the 'Analyze data' category on the left frame of the screen will let one retrieve the list of experimental sets (Fig. 1). Additional information can be obtained by clicking on the nutrient name at 'Name'. Also possible is direct access to the raw data under 'Raw data sets' (if it exists) although the number of accessible raw data sets is limited thus far.

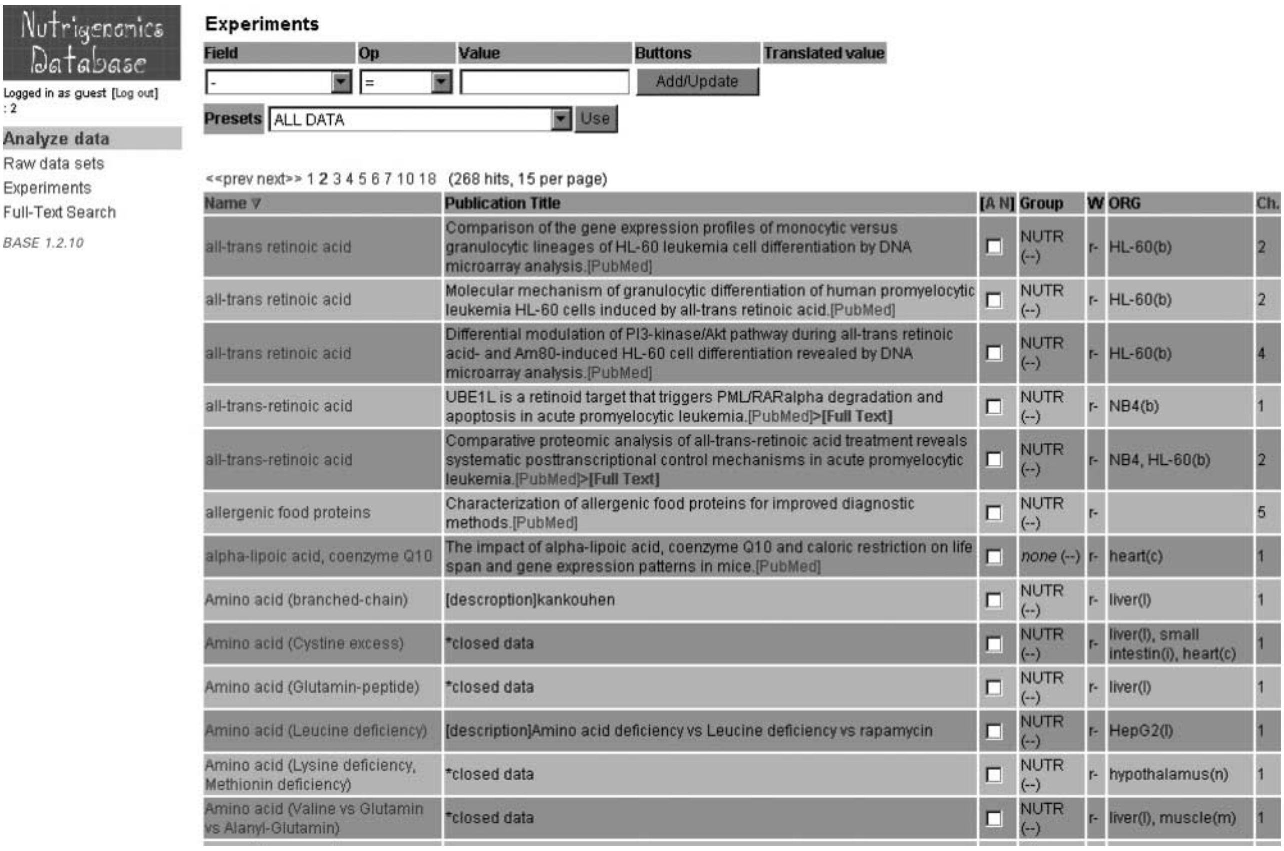
Probably the most common way of using this database is intuitive access to the pre-categorized resource of publication data by selecting a keyword from the pull-down menu, 'Presets' feature, which contains a few dozen keywords derived from nutrition-specific information sources. These auxiliary annotations were added manually based on the information extracted from discrete publications. The presets data are principally categorized by the names and types of dietary components, including some functional food components such as 'flavonoids', and also by the target tissues used in microarray experiments. In addition, the filter function at the top of the screen enables one to extract the data records containing the arbitrary keyword from any fields of a data set. Such accessibility to potentially relevant studies may be the key advantage of the database; it would greatly help to reduce working time and thereby improve research productivity since the literature-based analytical stage is still one of the most time-consuming steps in microarray studies.

Statistical approaches are typically used to analyse a large amount of gene expression data (Datta & Datta, 2003; Gottardo *et al.* 2003). Thus, numerous analytical tools are preinstalled on the original BASE system to avoid the need to download the array data for analysis by third-party software. Analytical tools include functions such as normalization, multidimensional scaling, principal components analysis and clustering, but thus far these do not fully support the single channel array. Detailed instructions for each application are available from BASE website (http://base.thep.lu.se/).

The database has been designed to store comprehensive gene expression information that provides researchers with clues to understanding the biological actions of nutrients at the molecular, cellular and individual animal levels. Some data sets do not comply with the above criteria since the definition of 'nutrigenomics' is not yet standardized. We are currently accumulating and summarizing any kind of data set that is likely to be valuable to nutrigenomics researchers. For instance, some data relating to the effects of pathological states such as diabetes are included. Data are updated manually as needed for the time being, because the number of available data sets that are directly tied to nutrition and food is still limited. The database currently carries over 200 nutrigenomics studies and over 300 pharmaco- and toxicogenomics studies, which will provide suggestive information for nutrition studies. Some data sets are not authorized for access because they include unpublished or closed data.

Currently, our database does not fully support Minimum Information About a Microarray Experiment (MIAME)-compliant information. MIAME aims at standardizing microarray information in order to facilitate the sharing and comparing of data among laboratories and researchers (Brazma *et al.* 2001). One of the common problems encountered in inter-experimental comparison of microarray data is the comparability of the data produced by means of various experimental procedures. We are therefore expanding the functions of the database so that it can handle the MIAME data as well as more nutrition-specific annotations. It is hoped that such features will facilitate the standardization of experimental conditions to develop an optimized workflow, a most challenging part of nutrigenomics study. Another feature to be added in the near future is the ability to search for experimental sets according to the gene expression changes of interest. In addition, we have already begun to develop and implement an algorithm that allows us to predict the functions of diets and nutrients based on a similarity of gene expression patterns. Detailed information about new features will be displayed on the top page of the database.

At this time, only a limited amount of gene expression data and raw array data is stored in the database because of its poor public availability. With the wealth of data generated from microarray experiments, however, there is clearly a need for data storage and



Experiments

Field	Op	Value	Buttons	Translated value
-	=		Add/Update	

Presets: ALL DATA

<<prev next>> 1 2 3 4 5 6 7 10 18 (268 hits, 15 per page)

Name ▽	Publication Title	[A N]	Group	W	ORG	Ch.
all-trans retinoic acid	Comparison of the gene expression profiles of monocytic versus granulocytic lineages of HL-60 leukemia cell differentiation by DNA microarray analysis.[PubMed]	□	NUTR (--)	r-	HL-60(b)	2
all-trans retinoic acid	Molecular mechanism of granulocytic differentiation of human promyelocytic leukemia HL-60 cells induced by all-trans retinoic acid.[PubMed]	□	NUTR (--)	r-	HL-60(b)	2
all-trans retinoic acid	Differential modulation of PI3-kinase/Akt pathway during all-trans retinoic acid- and Am80-induced HL-60 cell differentiation revealed by DNA microarray analysis.[PubMed]	□	NUTR (--)	r-	HL-60(b)	4
all-trans-retinoic acid	UBE1L is a retinoid target that triggers PML/RARalpha degradation and apoptosis in acute promyelocytic leukemia.[PubMed]>**[Full Text]**	□	NUTR (--)	r-	NB4(b)	1
all-trans-retinoic acid	Comparative proteomic analysis of all-trans-retinoic acid treatment reveals systematic posttranscriptional control mechanisms in acute promyelocytic leukemia.[PubMed]>**[Full Text]**	□	NUTR (--)	r-	NB4, HL-60(b)	2
allergenic food proteins	Characterization of allergenic food proteins for improved diagnostic methods.[PubMed]	□	NUTR (--)	r-		5
alpha-lipoic acid, coenzyme Q10	The impact of alpha-lipoic acid, coenzyme Q10 and caloric restriction on life span and gene expression patterns in mice.[PubMed]	□	*none* (--)	r-	heart(c)	1
Amino acid (branched-chain)	[descroption]kankouhen	□	NUTR (--)	r-	liver(l)	1
Amino acid (Cystine excess)	*closed data	□	NUTR (--)	r-	liver(l), small intestin(i), heart(c)	1
Amino acid (Glutamin-peptide)	*closed data	□	NUTR (--)	r-	liver(l)	1
Amino acid (Leucine deficiency)	[description]Amino acid deficiency vs Leucine deficiency vs rapamycin	□	NUTR (--)	r-	HepG2(l)	1
Amino acid (Lysine deficiency, Methionin deficiency)	*closed data	□	NUTR (--)	r-	hypothalamus(n)	1
Amino acid (Valine vs Glutamin vs Alanyl-Glutamin)	*closed data	□	NUTR (--)	r-	liver(l), muscle(m)	1

Fig. 1. A snapshot of the data-browsing page. The list of experimental sets includes links to external databases such as PubMed of the National Center for Biotechnology Information and public databases that contain Minimum Information About a Microarray Experiment-compliant microarray expression data. In addition, users can intuitively retrieve the pre-categorized resource of publication data by selecting a keyword from the pull-down menu, 'Presets'.

management optimized for nutrigenomics study. It is surely essential to consistently accumulate the array data in collaboration with other research groups and organizations.

Acknowledgements

This work is in part supported by Special Coordination Funds of the Japan Ministry of Education, Culture, Sports, Science and Technology (to S. A.) and a Grant-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (to H. K.). The authors thank Mr Shunsuke Aoki (TeamLab Co., Ltd) for his technical help. We are grateful to Dr Manabu Sami and Mt Yutaka Ohta (Asahi Breweries, Ltd) for pertinent discussions and sharing microarray data.

References

Brazma A, Hingamp P, Quackenbush J, *et al.* (2001) Minimum Information About a Microarray Experiment (MIAME) – towards standards for microarray data. *Nat Genet* **29**, 365–371.

Datta S & Datta S (2003) Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics* **19**, 459–466.

Endo Y, Fu Z, Abe K, Arai S & Kato H (2002) Dietary protein quantity and quality affect rat hepatic gene expression. *J Nutr* **132**, 3632–3637.

Gardiner-Garden M & Littlejohn TG (2001) A comparison of microarray databases. *Brief Bioinform* **2**, 143–158.

Goodman AL & Lory S (2004) Analysis of regulatory networks in *Pseudomonas aeruginosa* by genomewide transcriptional profiling. *Curr Opin Microbiol* **7**, 39–44.

Gottardo R, Pannucci JA, Kuske CR & Brettin T (2003) Statistical analysis of microarray data: a Bayesian approach. *Biostatistics* **4**, 597–620.

Kato H & Kimura T (2003) Evaluation of the effects of the dietary intake of proteins and amino acids by DNA microarray technology. *J Nutr* **133**, Suppl. 1, 2073S–2077S.

Mirnics K & Pevsner J (2004) Progress in the use of microarray technology to study the neurobiology of disease. *Nat Neurosci* **7**, 434–439.

Muller M & Kersten S (2003) Nutrigenomics: goals and strategies. *Nat Rev Genet* **4**, 315–322.

Quackenbush J (2001) Computational analysis of microarray data. *Nat Rev Genet* **2**, 418–427.

Saal LH, Troein C, Vallon-Christersson J, Gruvberger S, Borg A & Peterson C (2002) BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data. *Genome Biol* **3**, SOFTWARE0003.

Schena M, Shalon D, Davis RW & Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470.

Sreekumar R, Unnikrishnan J, Fu A, Nygren J, Short KR, Schimke J, Barazzoni R & Nair KS (2002) Impact of high-fat diet and antioxidant supplement on mitochondrial functions and gene transcripts in rat muscle. *Am J Physiol Endocrinol Metab* **282**, E1055–E1061.

Straume M (2004) DNA microarray time series analysis: automated statistical assessment of circadian rhythms in gene expression patterning. *Methods Enzymol* **383**, 149–166.

Weindruch R, Kayo T, Lee CK & Prolla TA (2001) Microarray profiling of gene expression in aging and its alteration by caloric restriction in mice. *J Nutr* **131**, Suppl., 918S–923S.