*Human Genome Center*

# Laboratory of Genome Database
# Laboratory of Sequence Analysis
# ゲノムデータベース分野
# シークエンスデータ情報処理分野


| | | | | |
|---|---|---|---|---|
| Professor | Minoru Kanehisa, Ph.D. | 教　授 | 理学博士 | 金　久　　實 |
| Research Associate | Toshiaki Katayama, M.Sc. | 助　手 | 理学修士 | 片　山　俊　明 |
| Research Associate | Michihiro Araki, Ph.D. | 助　手 | 薬学博士 | 荒　木　通　啓 |



*Owing to continuous developments of high-throughput experimental technologies, ever-increasing amounts of data are being generated in functional genomics and proteomics. We are developing a new generation of databases and computational technologies, beyond the traditional genome databases and sequence analysis tools, for making full use of such large-scale data in biomedical applications, especially for elucidating cellular functions as behaviors of complex interaction systems.*


## 1. Community databases for linking genomes to cellular functions

**Miho FURUMICHI, Yoko SATO, Toshiaki KATAYAMA, and Minoru KANEHISA**

The availability of the genome sequence and a complete set of genes for any organism is a great boon to researchers in the field, accelerating their research and resulting in a number of publications. However, accelerated research poses a database problem. When new gene functions are identified, the published result is stored in PubMed, but not necessarily in the public sequence repositories of GenBank, EMBL, and DDBJ. The lack of an up-to-date, well-annotated database is a major problem for most of the prokaryotic genomes thus far sequenced. Perhaps, the best solution to the current database problem is to get the research community actively involved in the annotation process. We are promoting a community database as a vehicle by which the community collectively annotates available genomes. In return the community can benefit from customization of KEGG - the community database provides an organism-specific view for exploring cellular functions and comparative genomics. Thus far, the most successful community databases have been developed for cyanobacteria and *Bacillus subtilis*.

## 2. Automatic assignments of orthologs and paralogs in complete genomes

**Toshiaki KATAYAMA and Minoru KANEHISA**

In addition to the human intensive efforts in the community databases, we are developing a computational method for automating genome annotations. It is based on a graph analysis of the KEGG SSDB database, containing sequence similarity relations among all the genes in the completely sequenced genomes. The nodes of the SSDB graph are genes (currently about 600,000 genes in over 150 genomes) and the

edges are the Smith-Waterman sequence similarity scores computed by the SSEARCH program (currently over 300 million edges above the threshold score of 100). The edges are not only weighted but also directed, indicating the best (top-scoring) hit when a gene in an organism is compared against all genes in another organism. Thus, a highly connected cluster of nodes containing a number of bidirectional best hits might be considered an ortholog cluster consisting of functionally identical genes. Such a cluster can be found by our heuristic method for finding "quasi-cliques", but the SSDB graph is too large to perform quasi-clique finding at a time. Therefore, we introduce a hierarchy (evolutionary relationship) of organisms and treat the SSDB graph as a nested graph.

### 3. Carbohydrate sequence database and glycoinformatics

**Michihiro ARAKI and Minoru KANEHISA**

Despite the importance as a third major class of biological macromolecules, the database efforts and associated informatics studies for glycans lag far behind those for nucleic acids and proteins. In contrast to the polynucleotide chain of a nucleic acid and the polypeptide of a protein, the polysaccharide chain of a glycan is branched. It is a tree structure requiring different data representation and data analysis methods. KEGG GLYCAN is a new database for glycan structures, most of which are derived from defunct CarbBank and some of which are entered from KEGG pathways. Although new glycomics projects have been initiated around the world, most informatics efforts seem to be targeted to large-scale profiling data and not to high-quality sequence data. This is probably because carbohydrate sequence determination is still a laborious work. Thus, we have started surveying all literature in Medline indexed with "carbohydrate sequence" for possible inclusion in GLYCAN.

## Publications


Kanehisa, M. and Bork, P. Bioinformatics in the post-sequence era. Nat. Genet. 33: 305-310, 2003.

Vert, J.-P. and Kanehisa, M. Graph-driven features extraction from microarray data using diffusion kernels and kernel CCA. Advances in Neural Information Processing Systems 15. Edited by Becker, S., Thrun, S., and Obermayer, K. (MIT Press, MA). pp. 1425-1432, 2003.

Yamanishi, Y., Vert, J.-P., and Kanehisa, M. Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis. Bioinformatics 19: i323-i330, 2003.

Park, K.-J. and Kanehisa, M. Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. Bioinformatics 19: 1656-1663, 2003.

Vert, J.-P. and Kanehisa, M. Extracting active pathways from gene expression data. Bioinformatics 19: ii238-ii234, 2003.

Hattori, M., Okuno, Y., Goto, S., and Kanehisa, M. Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. J. Am. Chem. Soc. 125: 11853-11865, 2003.

Aoki, K.F., Yamaguchi, A., Okuno, Y., Akutsu, T., Ueda, N., Kanehisa, M., and Mamitsuka, H. Efficient tree-matching methods for accurate carbohydrate database queries. Genome Informatics 14: 134-143, 2003.

Hattori, M., Okuno, Y., Goto, S., and Kanehisa, M. Heuristics for chemical compound matching. Genome Informatics 14: 144-153, 2003.

Ota, K., Yamada, T., Yamanishi, Y., Goto, S., and Kanehisa, M. Comprehensive analysis of delay in transcriptional regulation using expression profiles. Genome Informatics 14: 302-303, 2003.

Kawano, S., Okuno, Y., Hashimoto, K., Yamamoto, H., Takematsu, H., Kozutsumi, Y., Goto, S., and Kanehisa, M. Prediction of glycan structures from glycosyltransferase expression profiles. Genome Informatics 14: 304-305, 2003.

Kotera, M., Hattori, M., Goto, S., and Kanehisa, M. Analysis of reactive modules in the metabolic pathways. Genome Informatics 14: 366-367, 2003.

Limviphuvadh, V., Okuno, Y., Katayama, T., Goto, S., Yoshizawa, A.C., and Kanehisa, M. Metabolic pathway reconstruction for malaria parasite Plasmodium falciparum. Genome Informatics 14: 368-369, 2003.

Tanaka, M., Okuno, Y., Yamada, T., Goto, S., Uemura, S., and Kanehisa, M. Extraction of a thermodynamic property for biochemical reactions in the metabolic pathway. Genome Informatics 14: 370-371, 2003.

Honda, W., Kawashima, S., and Kanehisa, M. Self-nonself discrimination based on incompatibility of amino acid sequences of human and viruses. Genome Informatics 14: 432-433, 2003.

Itoh, M., Yoshizawa, A.C., Okuda, S., Goto, S., and Kanehisa, M. Analysis of domain combinations in eukaryotic genomes. Genome Informatics 14: 434-435, 2003.

Yoshizawa, A.C., Itoh, M., Okuda, S., Limviphuvadh, V., Sakiyama, T., Kawashima, S., and Kanehisa, M. Comprehensive survey of intracellular transport system-related proteins in complete genomes and draft genomes. Genome Informatics 14: 436-437, 2003.

Yamanishi, Y., Yoshizawa, A.C., Itoh, M., Katayama, T., and Kanehisa, M. Extraction of organism groups from whole genome comparisons. Genome Informatics 14: 438-439 2003.

Sato, T., Yamanishi, Y., Horimoto, K., Toh, H., and Kanehisa, M. Prediction of protein-protein interactions from phylogenetic trees using partial correlation coefficient. Genome Informatics 14: 496-497, 2003.

Minowa, Y., Katayama, T., Nakaya, A., Goto, S., and Kanehisa, M. Classification of protein sequences into paralog and ortholog clusters using sequence similarity profiles of KEGG/SSDB. Genome Informatics 14: 528-529, 2003.

Park, K.-J., Kanehisa, M., and Akiyama, Y. PLOC: prediction of subcellular location of proteins. Genome Informatics 14: 559-560, 2003.

Aoki, K.F., Yamaguchi, A., Okuno, Y., Akutsu, T., Ueda, N., Kanehisa, M., and Mamitsuka, H. Statistical siginficance of tree similarity scores. Genome Informatics 14: 579-580, 2003.

Igarashi, Y., Okuno, Y., Hattori, M., Goto, S., and Kanehisa, M. Common features in substrates of multidrug resistance transporters. Genome Informatics 14: 601-602 2003.

Goto, N., Nakao, M.C., Kawashima, S., Katayama, T., and Kanehisa, M. BioRuby: open-source bioinformatics library. Genome Informatics 14: 629-630, 2003.

Hashimoto, K., Hamajima, M., Goto, S., Masumoto, S., Kawasima, M., and Kanehisa, M. GLYCAN: the database of carbohydrate structures. Genome Informatics 14: 649-650, 2003.

Sakiyama, T., Kawashima, S., Yoshizawa, A.C., and Kanehisa, M. The construction of a database for ubiquitin signaling cascade. Genome Informatics 14: 653-654, 2003.

Furumichi, M., Sato, Y., Katayama, T., Kawashima, S., and Kanehisa, M. Development of community annotation databases for linking genomes to cellular functions. Genome Informatics 14: 657-658, 2003.

Kawashima, S., Katayama, T., Sato, Y., and Kanehisa, M. KEGG API: a web service using SOAP/WSDL to access the KEGG system. Genome Informatics 14: 673-674, 2003.

Kanehisa, M. Integration of bioinformatics and cheminformatics in KEGG. Genome Informatics 14: 715-716, 2003.

*Human Genome Center*

# Laboratory of Genome Structure Analysis
## ゲノム構造解析分野

Associate Professor Sumio Sugano, M.D., D.M.Sc.
Research Associate Yutaka Suzuki, Ph.D.

助教授 医学博士 菅野純夫
助手 理学博士 鈴木穣

*The main project of our laboratory is to identify and collect human genes en masse in the form of full-length cDNA clones. The sequence informations of full-length cDNA are indispensable for elucidating exon-intron structures as well as promoters of genes. Furthermore, full-length cDNA clones are valuable resource for the functional analysis of proteins coded by the genes. Thus, the direction of our Laboratory is a mass determination of gene structures and functions. Following are topics in the year 2003.*

### Identification and isolation of human full-length cDNA clones by 1 pass sequencing.

**Yutaka Suzuki, Hiroko Kozuka-Hata, Kiyomi Yoshitomo-Nakagawa , Junko Mizushima-Sugano, Tomohiro Hasui and Sumio Sugano:**

We have sequenced 5′ end of randomly picked cDNA clones from full-length enriched cDNA libraries made by "oligo-capping" method. We have sequenced about 100,000 clones this year. Of these clones, about 80% of them contained already known genes. About 80% of the known clones seemed to be full. Now, we have about 30,000 putative full-length cDNA clones with unknown function. Using 5′ end 1 pass sequence data, we identified mRNA start sites of 8000 genes and now making human promoter data using these data.

With FLJ cDNA sequencing consortium, the entire sequence was determined 30,000 clones. The average length of cDNA is about 2200bp which distribute from 1kb to 7kb. About half of them had ORF longer than 120 amino acid residues (AA). The average ORF length is about 390 AA. About 16% of these clones had membrane-spanning sequence and 3.6% signal sequences. Further more, about 25% of the clones with ORF longer than 120 AA had some type of motifs or showed some homology to known proteins. We are also mapping these fully sequenced clones to the draft sequence of the human genomes. The sequence data were deposited on the Genbank database and the clones will be available from several suppliers.

### Identification of putative mitochondria proteins by high-throughput subcellular localization analysis.

**Takushi Togashi, Yutaka Suzuki, Sumio Sugano**

The rapid accumulation of both human genome and cDNA sequences give us bases for understand the human as a complex molecular system. High-throughput analysis of subcellular localization is one of such analysis. We are putting human ORFs in to GATEWAY system and started using them for high-throughput functional analysis. We performed high-throughput analysis of sub-cellular localization of proteins coded by the cDNAs using GFP fusion proteins. About 16.3% of the fusion proteins localized in nucleus, 16.7% in cytoplasm, 41.5% both in cytoplasm and nucleus, 11% in ER and only 5.3% in

small organelles such as peroxisome and mitochondria. The 38 putative mitochondria genes were then analyzed by sequence homology. 12 cDNA clones showed homology to previously known mitochondria genes. In addition, we identified 5 gene that are not known to localize to mitochondria previously and 21 new mitochondria genes.

### Proteomic analysis of human small proteins

**Masaaki Oyama, Hiroko Hata, Yutaka Suzuki, Sumio Sugano**

Proteomic analysis of small proteins expressed in human leukemia K562 cells was performed by high-resolution nanoflow liquid chromatography coupled with electrospray ionization tandem mass spectrometry. Our analysis led to the identification of 59 proteins whose amino acid length was not over 100 in total, including 7 novel proteins. Five out of seven novel coding sequences were located upstream of the longest open reading frame and the other two were the longest in the corresponding mRNA. To our surprise, we found three coding sequences overlapping each downstream longest open reading frame that bears functional constraint on its sequence. Our finding indicates the translation of short open reading frame occurs in vivo whether or not there exists a longer potential coding region at the downstream of mRNA. This evidence supports the generality of the classical translation model in which the first ATG codon is recognized as a translation initiation site.

### Monkey cDNA project

**Munetomo Hida, Yutaka Suzuki, Sumio Sugano**

In collaboration with Prof. Momoki Hirai in Faculty of Science and Dr. Katsuyuki Hashimoto in National Institute of Infectious Diseases, we started monkey cDNA identification similar to that of human described above. The target organ for the isolation of full-length cDNAs is brain. We made "Oligo-capping" cDNA libraries from various parts of Macaca brain and more than 40,000 cDNA clones were sequenced at their 5′ end and the comparison between human data is in progress.

## Reference


Ota T. et al., Nature Genetics, in press

Watanabe, J., Sasaki, M., Suzuki, Y. and Sugano, S. Nucleic Acids Res. in press.

Suzuki, Y., Yamashita, R., Sugano, S. and Nakai, K. Nucleic Acids Res. in press.

Clark, M.S., Edwards, Y.J., Peterson, D., Clifton, S.W., Thompson, A.J., Sasaki, M., Suzuki, Y., Kikuchi, K., Watabe, S., Kawakami, K. Sugano, S. et al. Fugu ESTs: new resources for transcription analysis and genome annotation. Genome Res. 13: 2747-2753, 2003.

Ito Y, Oike Y, Yasunaga K, Hamada K, Miyata K, Matsumoto S, Sugano S, Tanihara H, Masuho Y, Suda T. Inhibition of angiogenesis and vascular leakiness by angiopoietin-related protein 4. Cancer Res. 63: 6651-6657, 2003.

Ohira M, Morohashi A, Inuzuka H, Shishikura T, Kawamoto T, Kageyama H, Nakamura Y, Isogai E, Takayasu H, Sakiyama S, Suzuki Y, Sugano S, Goto T, Sato S, Nakagawara Expression profiling and characterization of 4200 genes cloned from primary neuroblastomas: identification of 305 genes differentially expressed between favorable and unfavorable subsets. Oncogene. 22: 5525-5536. 2003

Tsai CC, Chung YD, Lee HJ, Chang WH, Suzuku Y, Sugano S, Lin JY. Large-scale sequencing analysis of the full-length cDNA library of human hepatocellular carcinoma. J Biomed Sci 10: 636-643. 2003.

Ohira M, Morohashi A, Nakamura Y, Isogai E, Furuya K, Hamano S, Machida T, Aoyama M, Fukumura M, Miyazaki K, Suzuki Y, Sugano S, Hirato J, Nakagawara A. Neuroblastoma oligo-capping cDNA project: toward the understanding of the genesis and biology of neuroblastoma. Cancer Lett. 197: 63-68, 2003.

Kikuchi S, et al., Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice. Science. 301: 376-379, 2003.

Shibui-Nihei A, Ohmori Y, Yoshida K, Imai J, Oosuga I, Iidaka M, Suzuki Y, Mizushima-Sugano J, Yoshitomo-Nakagawa K, Sugano S. The 5′ terminal oligopyrimidine tract of human elongation factor 1A-1 gene functions as a transcriptional initiator and produces a variable number of Us at the transcriptional level. Gene. 311: 137-45, 2003.

Furukawa K, Horie M, Okutomi K, Sugano S, Furukawa K. Isolation and functional analysis of the melanoma specific promoter region of human GD3 synthase gene. Biochim Biophys Acta. 1627: 71-78, 2003.

Arai M, Yokosuka O, Chiba T, Imazeki F, Kato M, Hashida J, Ueda Y, Sugano S, Hashimoto K, Saisho H, Takiguchi M, Seki N. Gene Ex-

pression Profiling Reveals the Mechanism and Pathophysiology of Mouse Liver Regeneration. J Biol Chem. 278: 29813-29818, 2003.

Matsuda A, Suzuki Y, Honda G, Muramatsu S, Matsuzaki O, Nagano Y, Doi T, Shimotohno K, Harada T, Nishida E, Hayashi H, Sugano S. Large-scale identification and characterization of human genes that activate NF-kappaB and MAPK signaling pathways. Oncogene. 22: 3307-3318, 2003.

Sakate R, Osada N, Hida M, Sugano S, Hayasaka I, Shimohira N, Yanagi S, Suto Y, Hashimoto K, Hirai M. Analysis of 5′-end sequences of chimpanzee cDNAs. Genome Res. 13: 1022-1026, 2003.

Suzuki Y, Sugano S. Construction of a full-length enriched and a 5′-end enriched cDNA library using the oligo-capping method. Methods Mol Biol. 221: 73-91, 2003.

Harada T, Matsuzaki O, Hayashi H, Sugano S, Matsuda A, Nishida E. AKRL1 and AKRL2 activate the JNK pathway. Genes Cells. 8: 493-500, 2003.

Jin Y, Suzuki H, Maegawa S, Endo H, Sugano S, Hashimoto K, Yasuda K, Inoue K. A vertebrate RNA-binding protein Fox-1 regulates tissue-specific splicing via the pentanucleotide GCAUG. EMBO J. 22: 905-912, 2003.

Xie GX, Han X, Ito E, Yanagisawa Y, Maruyama K, Sugano S, Suzuki Y, Wang Y, Gabriel A, Stevens SK, Mitchell J, Sharma M, Palmer PP. Gene structure, dual-promoters and mRNA alternative splicing of the human and mouse regulator of G protein signaling GAIP/RGS19. J Mol Biol. 325: 721-732, 2003.

*Human Genome Center*

# Laboratory of DNA Information Analysis
## DNA情報解析分野


| | | | | | |
|---|---|---|---|---|---|
| Professor | Satoru Miyano, Ph.D. | | 教　授 | 医学博士 | 宮　野　　悟 |
| Research Associate | Seiya Imoto, Ph.D. | | 助　手 | 理学博士 | 井　元　清　哉 |
| Research Associate | Hideo Bannai, M.Sc. | | 助　手 | 理学修士 | 坂　内　英　夫 |
| Instructor | Michiel J. L. de Hoon, Ph.D. | | 特任教員 | Ph.D. | Michiel J. L. de Hoon |



*The aim of the research at this laboratory is to establish computational methodologies for discovering and interpreting information of nucleic acid sequences, proteins and some other experimental data arising from researches in Genome Science. Our current concern is to realize a system which can deal with the relationship between sequence information and biological functions by extracting biological knowledge encoded on sequences and by using knowledge bases developed so far. Apart from the research activity, the laboratory has been providing bioinformatics software tools and has been taking a leading part in organizing an international forum for Genome Informatics.*


### 1. Gene Network Inference and Its Applications

#### a. Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks


**Seiya Imoto, Tomoyuki Higuchi[1], Takao Goto, Kousuke Tashiro[2], Satoru Kuhara[2], Satoru Miyano: [1]Institute of Statistical Mathematics and [2]Graduate School of Genetic Resources Technology, Kyushu University**


We propose a statistical method for estimating a gene network based on Bayesian networks from microarray gene expression data together with biological knowledge including protein-protein interactions, protein-DNA interactions, binding site information, existing literature and so on. Unfortunately, microarray data do not contain enough information for constructing gene networks accurately in many cases. Our method adds biological knowledge to the estimation method of gene networks under a Bayesian statistical framework, and also controls the trade-off between microarray information and biological knowledge automatically. We conduct Monte Carlo simulations to show the effectiveness of the proposed method. We analyze *Saccharomyces cerevisiae* gene expression data as an application.

#### b. Use of gene networks for identifying and validating drug targets


**Seiya Imoto, Christopher J. Savoie[3], Sachiyo Aburatani[2], Sunyong Kim, Kousuke Tashiro[2], Satoru Kuhara[2], Satoru Miyano: [3]Gene Networks, Inc.**


We propose a new method for identifying and validating drug targets by using gene networks, which are estimated from cDNA microarray gene expression profile data. We created novel gene disruption and drug response microarray gene expression profile data libraries for the purpose of drug target elucidation. We use two types of microarray gene expression

profile data for estimating gene networks and then identifying drug targets. The estimated gene networks play an essential role in understanding drug response data and this information is unattainable from clustering methods, which are the standard for gene expression analysis. In the construction of gene networks, we use the Bayesian network model. We use an actual example from analysis of the *Saccharomyces cerevisiae* gene expression profile data to express a concrete strategy for the application of gene network information to drug discovery.

### c. Use of gene networks from full genome microarray libraries to identify functionally relevant drug-affected genes and gene regulation cascades

**Christopher J. Savoie[3], Sachiyo Aburatani[2], Shouji Watanabe[4], Yukihiro Eguchi[4], Shigeru Muta[2], Seiya Imoto, Satoru Miyano, Satoru Kuhara[2], Kousuke Tashiro[2]: [4]Mitsui Knowledge Industry, Co. Ltd.**

We developed an extensive yeast gene expression library consisting of full-genome cDNA array data for over 500 yeast strains, each with a single-gene disruption. Using this data, combined with dose and time course expression experiments with the oral antifungal agent griseofulvin, whose exact molecular targets were previously unknown, we used Boolean and Bayesian network discovery techniques to determine the gene expression regulatory cascades affected directly by this drug. Using this method we identified CIK1 as an important affected target gene related to the functional phenotype induced by griseofulvin. Cellular functional analysis of griseofulvin showed similar tubulin-speci.c morphological effects on mitotic spindle formation to those of the drug, in agreement with the known function of CIK1p. Further, using the nonparametric, nonlinear Bayesian gene networks we were able to identify alternative ligand-dependant transcription factors and G protein homologues upstream of CIK1 that regulate CIK1 expression and might therefore serve as alternative molecular targets to induce the same molecular response as griseofulvin.

### d. Inferring gene networks from time series microarray data using dynamic Bayesian networks

**Sunyoung Kim, Seiya Imoto, Satoru Miyano**

Dynamic Bayesian networks (DBNs) are considered as a promising model for inferring gene networks from time series microarray data. DBNs take over the advantages of Bayesian networks and can construct cyclic regulations using time delay information. In this paper, we summarize a general framework of DBN modeling. Both discrete and continuous DBN models are constructed systematically and criteria for learning network structures are introduced from a Bayesian statistical viewpoint. The detailed survey is presented for several applications of DBNs over past years. We also show real data applications for *S. cerevisiae* time series gene expression data.

### e. Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data

**Sunyong Kim, Seiya Imoto, Satoru Miyano**

We propose a dynamic Bayesian network and nonparametric regression model for constructing a gene network from time series microarray gene expression data. The proposed method can overcome a shortcoming of the Bayesian network model in the sense of the construction of cyclic regulations. The proposed method can analyze the microarray data as continuous data and can capture even nonlinear relations among genes. It can be expected that this model will give a deeper insight into the complicated biological systems. We also derive a new criterion for evaluating an estimated network from Bayes approach. We demonstrate the effectiveness of our method by analyzing *Saccharomyces cerevisiae* gene expression data.

### f. Finding optimal gene networks using biological constraints

**Sascha Ott, Satoru Miyano**

Finding gene networks from microarray data has been one focus of research in recent years. Given search spaces of super-exponential size, researchers have been applying heuristic approaches like greedy algorithms or simulated annealing to infer such networks. However, the accuracy of heuristics is uncertain, which-in combination with the high measurement noise of microarrays-makes it very difficult to draw conclusions from networks estimated by heuristics. We present a method that finds optimal Bayesian networks of considerable size and show first results of the application to yeast data. Having removed the uncertainty due to the heuristic methods, it becomes possible to evaluate the power of different statistical models to find biologically accurate networks.

### g. Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection

**Yoshinori Tamada, Sunyong Kim, Hideo Bannai, Seiya Imoto, Kousuke Tashiro[2], Satoru Kuhara,[2] Satoru Miyano**

We present a statistical method for estimating gene networks and detecting promoter elements simultaneously. When estimating a network from gene expression data alone, a common problem is that the number of microarrays is limited compared to the number of variables in the network model, making accurate estimation a difficult task. Our method overcomes this problem by integrating the microarray gene expression data and the DNA sequence information into a Bayesian network model. The basic idea of our method is that, if a parent gene is a transcription factor, its children may share a consensus motif in their promoter regions of the DNA sequences. Our method detects consensus motifs based on the structure of the estimated network, then re-estimates the network using the result of the motif detection. We continue this iteration until the network becomes stable. To show the effectiveness of our method, we conducted Monte Carlo simulations and applied our method to *Saccharomyces cerevisiae* data as a real application.

## 2. Biopathway Simulations and Systems Biology

### a. Inference, modeling and simulation of gene networks

**Satoru Miyano**

Systems biology may be soundly explored by development of computational tools and capabilities which enable us to understand complex biological systems. Strongly anticipated matters for systems biology are scientific contributions that also create practical benefits such as biomedical applications, solutions for environmental problems, etc. Thus, computational challenges in systems biology must be highly motivated with this direction and should not be regarded as yet another applications of techniques from computer science. This paper overviews our computational strategy for systems biology and indicates problems and further challenges.

One is computational inference of gene networks from gene expression profile data obtained from various perturbations such as gene disruptions, shocks, etc. We have considered three gene network models for gene network inference. The other is computational modeling and simulation of biological systems. We have developed a tool Genomic Object Net so that it would be a platform with which biological scientists can comfortably model and simulate dynamic causal interactions and processes in the cell such as gene regulations, metabolic pathways, and signal transduction cascades. An application of computational methods in systems biology is also shown. Recently, we have succeeded in discovering a drug target gene by analyzing gene networks constructed from gene expression profile data based on gene disruptions and drug doses. We have shown how this was achieved with gene network inference methods and laboratory works.

### b. Genomic Object Net: I. a platform for modeling and simulating biopathways

**Masao Nagasaki, Atsushi Doi, Hiroshi Matsuno[5], Satoru Miyano: [5]Faculty of Science, Yamaguchi University**

Genomic Object Net (GON) 1.0 is a software package for creating models and simulations of biopathways. Its core architecture employs the notion of a hybrid functional Petri net with extension (HFPNe). HFPNe can seamlessly handle discrete and continuous objects and events while keeping the model components themselves simple. With the feature and graphical model editor, biopathways can be modeled intuitively and simulated on GON. The subsequent output of the simulation results can be evaluated in customised views on GON Visualizer by writing an XML file. Additionally, GON provides a tool to transform biopathway models in KEGG and BioCyc to the GON XML files for modeling and simulation. The tool avoids a lot of tedious work by users so that they can focus on the biological model.

The trial version of Genomic Object Net can be obtained from GNI Inc (http://gene-networks.com). All XML files of example biopathways are available from the GON Project page (http://genomicobject.net/).

### c. Genomic Object Net: II. Modelling biopathways by hybrid functional Petri net with extension

**Atsushi Doi, Masao Nagasaki, Hiroshi Matsuno[5], Satoru Miyano**

This reseach demonstrates how to create an HFPNe (hybrid functional Petri net with extension) model, using the *lac* operon gene regula-

tory mechanism and glycolytic pathway as an example. Using this example, readers can then model other biopathways of interest. Simulations of the HFPNe model were performed using the software package Genomic Object Net.

### d. Biopathways representation and simulation on hybrid functional Petri net

**Hiroshi Matsuno[5], Yukiko Tanaka[5], Hitoshi Aoshima[5], Atsushi Doi, Mika Matsui[5], Satoru Miyano**

The following two matters should be resolved for biosimulation tools in order to be accepted by users in biology/medicine; (1) Remove issues which are irrelevant to biological importance, and (2) Allow users to represent biopathways intuitively and understand/manage easily the details of representation and simulation mechanism. From these criteria, we firstly define a novel notion of Petri net called hybrid functional Petri net (HFPN). Then, we introduce a software tool, Genomic Object Net, for representing and simulating biopathways, which we have developed by employing the architecture of HFPN. In order to show the effectiveness of Genomic Object Net for representing and simulating biopathways, we show some typical biopathway modelings related to gene regulation (switching mechanism of λ-phage, circadian rhythm of *Drosophila*, *lac* operon regulatory mechanism of *E. coli*), metabolic pathway (glycolytic pathway), and signal transduction (Fas ligand induced apoptosis), which cover the basic aspects in biopathways.

### e. Towards biopathway modeling and simulation

**Hiroshi Matsuno[5], Sachie Fujita[5], Atsuhi Doi, Masao Nagasaki, Satoru Miyano**

Petri net has been employed for modeling metabolic pathways as well as signal transduction pathways and gene regulatory networks. The purpose of this paper is to introduce an extension of Petri net called hybrid functional Petri net (HFPN) which allows us to model biopathways naturally and effectively. The method for creating biopathways with HFPN is demonstrated through a well-known biopathway example lac operon gene regulatory mechanism and glycolytic pathway." In order to evaluate this biopathway model, simulations of five mutants of the *lac* operon are carried out by Genomic Object Net which is a biopathway simulator based on the HFPN architecture. The software Genomic Object Net and the HFPN files presented in this paper can be downloaded from http://www.GenomicObject.Net/.

## 3. Software and Algorithms

### a. Open source clustering software

**Michiel J.L. de Hoon, Seiya Imoto, J. Nolan[6], Satoru Miyano: [6]University of California, Santa Cruz Extension in Silicon Valley**

We have implemented *k*-means clustering, hierarchical clustering, and self-organizing maps in a single multipurpose open-source library of C routines, callable from other C and C++ programs. Using this library, we have created an improved version of Michael Eisen's well-known Cluster program for Windows, Mac OS X, and Linux/Unix. In addition, we generated a Python and a Perl interface to the C Clustering Library, thereby combining the flexibility of a scripting language with the speed of C.

The C Clustering Library and the corresponding Python C extension module Pycluster were released under the Python License, while the Perl module Algorithm:: Cluster was released under the Artistic License. The GUI code Cluster 3.0 for Windows, Macintosh, and Linux/Unix, as well as the corresponding command-line program, were released under the same license as the original Cluster code.

The complete source code is available at http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster. Alternatively, Algorithm:: Cluster can be downloaded from CPAN, while Pycluster is also available as part of the Biopython distribution.

### b. Efficiently finding regulatory elements using correlation with gene expression

**Hideo Bannai, Shunsuke Inenaga[7], Ayumi Shinohara[7], Masayuki Takeda[7], Satoru Miyano: [7]Department of Informatics, Kyushu University**

We present an efficient algorithm for detecting putative regulatory elements in the upstream DNA sequences of genes, using gene expression information obtained from microarray experiments. Based on a generalized suffix tree, our algorithm looks for motif patterns whose appearance in the upstream region is most correlated with the expression levels of the genes. We are able to find the optimal pattern, in time linear in the total length of the upstream sequences. We implement and apply our algorithm to publicly available microarray gene expression data, and show that our method is able to discover biologically significant motifs, in-

cluding various motifs which have been reported previously using the same dataset. We further discuss applications for which the efficiency of the method is essential, as well as possible extensions to our algorithm.

## Publications


Akutsu, T., Kuhara, S., Maruyama, O., Miyano, S. Identification of genetic networks by strategic gene disruptions and gene overexpressions under a boolean model. Theoretical Computer Science. 298(1): 235-251, 2003.

Akutsu, T., Miyano, S., Kuhara, S. A simple greedy algorithm for finding functional relations: efficient implementation and average case analysis. Theoretical Computer Science. 292(2): 481-495, 2003.

Bannai, H., Inenaga, S., Shinohara, A., Takeda, M., Miyano, S. Efficiently finding regulatory elements using correlation with gene expression. J. Bioinformatics and Computational Biology, in press.

Bannai, H., Inenaga, S., Shinohara, A., Takeda, M., Inferring strings from graphs and arrays. Proceedings of the 28th International Symposium on Mathematical Foundations of Computer Science (MFCS2003), Lecture Notes in Computer Science. 208-217, 2003.

De Hoon, M.J.L., Imoto, S., Nolan, J., Miyano, S. Open source clustering software. Bioinformatics, in press.

De Hoon, M., Imoto, S., Kobayashi, K., Ogasawara, N., Miyano, S. Inferring gene regulatory networks from time-ordered gene expression data of Bacillus subtilis using differential equations. Pacific Symposium on Biocomputing. 8: 17-28, 2003.

De Hoon, M.J.L, Chapman, B., Friedberg, I. Bioinformatics and computational biology with Biopython. Genome Informatics. 14: 298-299, 2003.

Doi, A., Nagasaki, M., Matsuno, H., Miyano, S. Genomic Object Net: II. Modelling biopathways by hybrid functional Petri net with extension. Applied Bioinformatics. 2(3): 185-188, 2003.

Gribskov, M., Kanehisa, M., Miyano, S., Takagi, T. (Eds.) Genome Informatics. 14, 2003.

Imoto, S., Kim, S., Goto, T., Aburatani, S., Tashiro, K., Kuhara, S., Miyano, S. Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network. Journal of Bioinformatics and Computational Biology. 1(2): 231-252, 2003.

Imoto, S., Higuchi, T., Goto, T., Tashiro, K., Kuhara, S., Miyano, S. Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. Journal of Bioinformatics and Computational Biology. 2 (2): in press.

Imoto, S., Konishi, S. Selection of smoothing parameters in *B*-spline nonparametric regression models using information criteria. Annals of the Institute of Statistical Mathematics, 55: 671 -687, 2003.

Imoto, S., Savoie, C.J., Aburatani, S., Kim, S., Tashiro, K., Kuhara, S., Miyano, S. Use of gene networks for identifying and validating drug targets. Journal of Bioinformatics and Computational Biology. 1(3): 459-474, 2003.

Jeong, E., Chung, I.F., Miyano, S. Prediction of residues in protein-RNA interaction sites by neural networks. Genome Informatics. 14: 506-507, 2003.

Kamimura, T., Shimodaira, H., Imoto, S., Kim, S., Tashiro, K., Kuhara, S., Miyano, S. Multiscale bootstrap analysis of gene networks based on Bayesian network and nonparametric regression. Genome Informatics, 14: 350-351, 2003.

Kim, S., Imoto, S., Miyano, S. Inferring gene networks from time series microarray data using dynamic Bayesian networks. Briefings in Bioinformatics. 4(3): 228-235, 2003.

Kim, S., Imoto, S., Miyano, S. Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data. Biosystems, in press.

Konishi, S., Ando, T., Imoto, S. Bayesian information criteria and smoothing parameter selection in radial basis function networks. Biometrika, 91, in press.

Kono, T., Noda, R., Kitakaze, H., Nagasaki, M., Doi, A., Matsuno, H., Miyano, S. Distributed client-server system architechture high performance simulations on Genomic Object Net. Genome Informatics, 14: 619-620, 2003.

Matsui, M., Fujita, S., Suzuki, S., Matsuno, H., Miyano, S. Simulated cell division processes of the *Xenopus* cell cycle pathway by Genomic Object Net, Journal of Integrative Bioinformatics, 0003, 2004 (Online Journal http://journal.imbio.de/index.php ? paper_id=3, 2004).

Matsuno, H., Murakami, R., Yamane, R., Yamasaki, N., Fujita, S., Yoshimori, H., Miyano, S. Boundary formation by notch signaling in Drosophila multicellular systems: experimental observations and gene network modeling by Genomic Object Net. Pacific Symposium on Biocomputing. 8: 152-63, 2003.

Matsuno, H., Tanaka, Y., Aoshima, H., Doi, A., Matsui, M., Miyano, S. Biopathways representation and simulation on hybrid functional Petri net. In Silico Biology. 3(3):. 389-404, 2003.

Matsuno, H., Fujita, S., Doi, A., Nagasaki, M., Miyano, S. Towards biopathway modeling and simulation. Proceedings of 24th International Conference on Applications and Theory of Petri Nets (ICATPN 2003). Lecture Notes in Computer Science. 2679: 3-22, 2003.

Miyano, S. Inference, modeling and simulation of gene networks. Proceedings of International Workshop on Computational Methods in Systems Biology. Lecture Notes in Computer Science. 2602: 207-211, 2003.

Miyano, S. Computational strategy for systems biology. Genome Informatics, 14: 723-725, 2003.

Nagasaki, M., Doi, A., Matsuno, H., Miyano, S. Integrating biopathway databases for large-scale modeling and simulation. Proceedings of Second Asia-Pacific Bioinformatics Conference (APBC2004). Vol. 29, in press.

Nagasaki, M., Doi, A., Matsuno, H., Miyano, S. Recreating biopathway databases towards simulation. Proceedings of International Workshop on Computational Methods in Systems Biology. Lecture Notes in Computer Science. 2602: 191-192, 2003.

Nagasaki, M. Doi, A., Matsuno, H., Miyano, S. Genomic Object Net: I. a platform for modeling and simulating biopathways. Applied Bioinformatics. 2(3): 181-184, 2003.

Nariai, N., Kim, S., Imoto, S., Miyano, S. Using protein-protein interactions for refining gene networks estimated from microarray data by Bayesian networks. Pacific Symposium on Biocomputing, 9, in press.

Nagasaki, M., Doi, A., Ueno, K., Torikai, E., Matsuno, H., Miyano, S. BPE: Biopathway executer for large-scale biopathway modeling and simulation. Genome Informatics, 14: 296-297, 2003.

Obara, W., Iida, A., Suzuki, Y., Tanaka, T., Akiyama, F., Maeda, S., Ohnishi, Y., Yamada, R., Tsunoda, T., Takei, T., Ito, K., Honda, K., Uchida, K., Tsuchiya, K., Yumura, W., Ujiie, T., Nagane, Y., Nitta, K., Miyano, S., Narita, I., Gejyo, F., Nihei, H., Fujioka, T., Nakamura, Y. Association of single-nucleotide polymorphisms in the polymeric immunoglobulin receptor gene with immunoglobulin A nephropathy (IgAN) in Japanese patients. J Hum Genet. 48(6): 293-299, 2003.

Ott, S., Tamada, Y., Bannai, H., Nakai, K., Miyano, S. Intrasplicing - analysis of long intron sequences. Pacific Symposium on Biocomputing. 8: 339-350, 2003.

Ott, S., Miyano, S. Finding optimal gene networks using biological constraints. Genome Informatics. 14: 124-133, 2003.

Ott, S., Imoto, S., Miyano, S. Finding optimal models for small gene networks. Pacific Symposium on Biocomputing, 9. in press.

Ott, S., Miyano, S. Enumeration of likely gene networks network motif extraction for large gene networks. Genome Informatics, 14: 354-355, 2003.

Savoie, C.J., Aburatani, S., Watanabe, S., Eguchi, Y., Muta, S., Imoto, S., Miyano, S., Kuhara, S., Tashiro, K. Use of gene networks from full genome microarray libraries to identify functionally relevant drug-affected genes and gene regulation cascades. DNA Research. 10(1): 19-25, 2003.

Sumii, E., Bannai, H., The extension of ML with hypothetical views for discovery science: formalization and implementation. J. Functional and Logic Programming, 2003.

Takeda, M., Inenaga, S., Bannai, H., Shinohara, A., Arikawa, S., Discovering most classificatory patterns for very expressive pattern classes. Proceedings of the 6th International Conference on Discovery Science (DS 2003), Lecture Notes in Computer Science. 2843: 486-493, 2003.

Takei, Y., Inoue, K., Ogoshi, M., Kawahara, T., Bannai, H., Miyano, S. Identification of novel adrenomedullin in mammals: a potent cardiovascular and renal regulator. FEBS Letters, 556(1-3): 53-58, 2003.

Tamada, Y., Kim, S., Bannai, H., Imoto, S., Tashiro, K., Kuhara, S., Miyano, S. Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. Bioinformatics. 19 (Suppl.2): ii227-ii236, 2003.

Tamada, Y., Kim, S., Bannai, H., Imoto, S., Tashiro, K., Kuhara, S., Miyano, S. Combining gene expression data with DNA sequence information for estimating gene networks using Bayesian network model. Genome Informatics, 14: 352-353, 2003.

Yamane, R., Umezaki, J., Matsuno, H., Murakami, R., Yamasaki, N., Miyano, S. Simulation of Drosophila boundary cell formation in forced-expression of Notch$^{\Delta E}$, Genome Informatics, 14: 617-618, 2003.

土井淳，長崎正朗，松野浩嗣，宮野悟．細胞反応のシミュレーション．わかる実験医学シリーズ「バイオインフォマティクスがわかる」．菅原秀明編集．羊土社．pp. 81–84，2003.

長崎正朗，土井淳，松野浩嗣，宮野悟．バイオパスウェイモデリングとシミュレーションを実現するためのシステム—Genomic Object Net—．人工知能学会誌, 18：12–20，2003.

*Human Genome Center*

# Laboratory of Genome Technology
# Laboratory of Molecular Medicine
## ゲノムシークエンス解析分野，シークエンス技術開発分野


| | | | | |
|---|---|---|---|---|
| Professor | Yusuke Nakamura, M.D., Ph.D. | 教　授 | 医学博士 | 中　村　祐　輔 |
| Associate Professor | Yoichi Furukawa M.D., Ph.D. | 助教授 | 医学博士 | 古　川　洋　一 |
| Assistant Professor | Toyomasa Katagiri, Ph.D. | 助　手 | 医学博士 | 片　桐　豊　雅 |
| Assistant Professor | Ryuji Hamamoto, Ph.D. | 助　手 | 医学博士 | 浜　本　隆　二 |
| Assistant Professor | Yataro Daigo, M.D., Ph.D. | 助　手 | 医学博士 | 醍　醐　弥太郎 |
| Assistant Professor | Hidewaki Nakagawa, M.D., Ph.D. | 助　手 | 医学博士 | 中　川　英　刀 |



*The major goal of the Human Genome Project is to identify genes predisposing to diseases, and to develop new diagnostic and therapeutic tools. We have been attempting to isolate genes involving in carcinogenesis and also those causing or predisposing to other diseases such as cardiovascular disease, deafness and some allergic diseases. By means of technologies developed through the genome project including a high-resolution SNP map, a large-scale DNA sequencing, and the cDNA microarray method, we have isolated a number of biologically and/or medically important genes.*


### 1. Genes playing significant roles in human cancer

### a. Genes that are inducible by p53

**Hirofumi Arakawa, Takashi Kimura, Ching C. Ng, Chizu Tanikawa, Yoshio Anazawa, Koji Ueda, Park Woong Ryeon and Yusuke Nakamura**

We found a novel p53-target gene, designated *p53RDL1* (p53-regulated Receptor for Death and Life), whose product containing a death-domain in the cytoplasmic C-terminal region is highly homologous to rat Unc5H2, a dependence receptor involved in apoptosis-regulation as well as axon guidance and migration of neural cells. p 53RDL1 mediated p53-dependent apoptosis. Conversely, when its ligand, Netrin-1, was present, the p53RDL1 signaling blocked p53-dependent apoptosis by its interaction with Netrin-1. p53RDL1 appears to be a previously unrecognized p53-target that may define a new pathway for p53-dependent apoptosis. We suggest that p53 might regulate both cell death and survival of damaged cells, by balancing regulation of the p53RDL1-Netrin-1 signaling for survival and cleavage of p53RDL1 for apoptosis, thereby helping to maintain the integrity of the genome.

The DNA-damage checkpoint plays a critical role in preventing genomic instability by regulating the cell cycle and DNA repair. Inactivation of the checkpoint may impair the DNA-repair mechanism and increase susceptibility of cells to genotoxic agents. p53, one of the critical checkpoint genes, is frequently mutated in cancers of various types. Genomic instability is

often observed in cancers carrying p53 mutations, but its mechanism is not fully understood; however, the discovery of p53R2 provided an important clue for clarifying it. A recently identified ribonucleotide reductase (RR), p53R2, is directly regulated by p53 for supplying nucleotides to repair damaged DNA. To investigate the notion that this enzyme might play a role in DNA repair by supplying deoxyribonucleotides for resting cells *in vivo*, we generated a strain of mice lacking *Rrm2b* (encoding p53R2). These mice developed normally until weaned, but thereafter displayed growth retardation and early mortality. Pathological examination indicated dysfunction of multiple organs, and all *Rrm2b*-null mice died from severe renal failure by the age of 14 weeks. TUNEL staining showed an increase of apoptotic cells in kidneys of 8-week-old *Rrm2b*$^{-/-}$ mice. p53 protein was activated in kidney tissues of these mice, leading to transcriptional induction of p53-target genes. *Rrm2b*$^{-/-}$ embryonic fibroblasts (MEFs) became immortal much earlier than *Rrm2b*$^{+/+}$ MEFs. dNTP pools were severely attenuated in *Rrm2b*$^{-/-}$ MEFs under oxidative stress. *Rrm2b*-deficiency caused increase of spontaneous mutation rates in the kidney of *Rrm2b*$^{-/-}$ mice. Our results suggest that p53R2 plays a pivotal role in maintaining dNTP levels for repair of DNA in arresting cells, and that impairment of this pathway may enhance spontaneous mutation frequency and activate p53-dependent apoptotic pathway(s) *in vivo*, causing severe renal failure, growth retardation and early mortality.

A mutant version of p53 (p53-121F), in which phenylalanine replaces the 121$^{st}$ serine residue, can induce apoptosis more effectively than wild-type p53 (wt-p53). In view of that observation we considered that one or more apoptosis-related p53-target genes might be preferentially induced by p53-121F. We carried out cDNA microarray analysis to identify such genes, using mRNAs isolated from LS174T colon-cancer cells infected by adenovirus vectors containing either p53-121F (Ad-*p53-121F*) or wild type p53 (Ad-wt*p53*). The *STAG1* gene was one of the transcripts showing higher expression levels in cells infected with Ad-p53-121F as opposed to Ad-wtp53. The encoded product appears to contain a transmembrane domain, and binding motifs for SH3 and WW. In two other cancer-cell lines, expression of *STAG1* mRNA was induced in response to various genotoxic stresses in a p53-dependent manner; moreover, enforced expression of *STAG1* led to apoptosis in several additional cancer-cell lines. Suppression of endogenous *STAG1* using the RNA-interference method reduced the apoptotic response, whether induced by Ad-*p53-121F* or Ad-*p53*. These results suggest that *STAG1*, a novel transcriptional target for p53, mediates p53-dependent apoptosis, and might be a good candidate for next-generation gene

Dual-specificity phosphatase 5 (*DUSP5*), a VH 1-like enzyme that hydrolyzes nuclear substrates phosphorylated on both tyrosine and serine/threonine residues, has a potential role in deactivation of mitogen- or stress-activated protein kinases. Using cDNA microarray technology, we found that expression of *DUSP5* mRNA was dramatically increased by exogenous p53 in U 373MG, a p53-mutant glioblastoma cell line. Transcription of *DUSP5* was also remarkably activated by endogenous p53 in response to DNA damage in colon-cancer cells (p53$^{+/+}$) that contained wild-type p53, but not in p53$^{-/-}$ cells. Chromatin-immunoprecipitation (ChIP) and reporter assays demonstrated that endogenous p 53 protein would bind directly to the promoter region of the DUSP5 gene, implying p53-dependent transcriptional activity. Over-expression of *DUSP5* suppressed growth of several types of human cancer cells, in which Erk1/2 was significantly dephosphorylated. If, as the results suggest, *DUSP5* is a direct target of p53, it represents a novel mechanism by which p53 might negatively regulate cell-cycle progression by down-regulating mitogen-or stress-activated protein kinases.

We also found that Introduction of exogenous p53 into a glioblastoma cell line lacking wild-type p53 (U373MG) dramatically induced expression of *Semaphorin3B* mRNA. An electrophoretic mobility-shift assay and a reporter assay confirmed that a potential p53-binding site present in the promoter region had p53-dependent transcriptional activity. Expression of endogenous semaphorin3B was induced in response to genotoxic stresses caused by adriamycin treatment or UV irradiation in a p53-dependent manner. Ectopic-expression of semaphorin3B in p53-defective cells reduced the number of colonies in colony-formation assays. These results suggest that Semaphorin3B might play some role in regulating cell growth, as a mediator of p53 tumor-suppressor activity.

The *p53AIP1* gene, which we recently identified as a novel p53-target, mediates p53-dependent apoptosis. We evaluated the effects of adenovirus-mediated introduction of *p53AIP1* (Ad-*p53AIP1*) on 30 human cancer-cell lines *in vitro*, and two cell lines *in vivo*, in comparison with the effects of p53 (Ad-*p53*). In 20 of the 30 cell lines, p53AIP1-induced apoptosis was observed, and in 12 of these p53AIP1-sensitive cancer cell lines, the apoptotic effects of p53AIP1 were greater than those of p53 itself. Cancers with wild-type p53, which were thought to be p

53-resisitant, were likely to be sensitive to p53 AIP1-induced apoptosis. p53-resistant cancers such as LS174T ($p53^{+/+}$) and A549 ($p53^{+/+}$), in which no increase of *p53AIP1* mRNA expression was observed when Ad-*p53* was introduced, were killed effectively by Ad-*p53AIP1*. Furthermore, co-introduction of *p53* and *p53AIP1* had synergistic effects on the induction of apoptosis regardless of *p53* status. Finally, adenovirus-mediated introduction of *p53AIP1* suppressed tumor growth *in vivo*. These results suggested that *p53AIP1* gene transfer might become a new strategy for the treatment of p53-resistant cancers.

mRNA-expression of *hCDC4b*, a gene whose product was one of four subunits of SCF complex responsible for Cyclin E degradation as the ubiquitin protein ligases, was dramatically up-regulated by infection of Ad-p53. An electrophoretic mobility shift assay and a chromatin immunoprecipitation assay indicated a potential p 53-binding site (p53BS) could bind to p53, which located in exon1b of the ef *hCDC4* gene. Moreover, a reporter assay confirmed that the p53BS had p53-dependent transcriptional activity. Expression of endogenous *hCDC4b*, but not the other transcript of this gene, *hCDC4a*, was induced in response to genotoxic stresses caused by UV irradiation and adriamycin treatment in a p53 dependent manner, suggesting a different role of each transcript. These results suggest that *hCDC4b* is a novel p53-target gene, and that, in addition to the machinery of $p21^{WAF1}$ for p53-dependent cell-cycle control, p53 might stop cell -cycle progression at G0-G1 phase through negative-regulation of Cyclin E by transcriptional induction of *hCDC4b*.

We also reported isolation of a novel transcriptional target of p53, designated *p53RFP* (p 53-inducible RING-finger protein), whose product has E3 ubiquitin ligase activity. Its expression was negatively correlated to that of $p21^{WAF1}$ protein; p53RFP appeared to play a significant role in degradation of this protein through direct interaction with, and ubiquitination of, p $21^{WAF1}$. p53RFP appears to represent the second known example, the first being MDM2, of an E3 ubiquitin ligase as a p53-target. Our results further suggest that p53 might destabilize $p21^{WAF1}$ through transcriptional regulation of *p53RFP*, and this feature may represent a novel mechanism for a p53-dependent cell-cycle checkpoint.

## b. Colon, Liver, and Gastric cancers

**Yoichi Furukawa, Ryuji Hamamoto, Li Meihua, Meiko Takahashi, Takashi Shimokawa, Takeshi Watanabe, Daisuke Yuki, Kazutaka Obama, Michihiro Sakai, Kazuya Ohyasu, Katsuaki Ura, Takaaki Kobayashi, Pittella Fabio, Natini Jinawath and Yusuke Nakamura**

Through a genome-wide cDNA microarray, we identified that the paternally expressed gene 10 (*PEG10*) was highly expressed in a great majority of hepatocellular carcinomas (HCCs) although its expression was absent in normal liver cells. Exogeneous expression of *PEG10* conferred oncogenic activity and transfection of hepatoma cells with anti-sense S-oligonucleotides suppressing *PEG10* resulted in their growth inhibition. Further experiments revealed that PEG10 protein associated with SIAH1, a mediator of apoptosis, and that overexpression of PEG10 decreased the cell death mediated by SIAH1. These findings suggested that development of drug(s) inhibiting PEG10 activity could be a novel approach for the treatment of HCCs.

We also reported isolation and characterization of a novel human gene at chromosomal band 1p36.13, termed *DDEFL1* (development and differentiation enhancing factor-like 1), encoding a product that shared structural features with centaurin-family proteins. The deduced 903 -amino-acid sequence showed 46% homology to DDEF/ASAP1 (development and differentiation enhancing factor), and contained an Arf GTPase-activating protein (ArfGAP) domain and two ankyrin repeats. Gene transfer of *DDEFL1* promoted proliferation of cells that lacked endogenous expression of this gene. Furthermore, reduction of *DDEFL1* expression by transfection of anti-sense S-oligonucleotides inhibited the growth of SNU475 cancer cells, in which *DDEFL 1* expression was highly up-regulated. Our results provide novel insight into hepatocarcinogenesis and may contribute to development of new strategies for diagnosis and treatment of HCC.

To search for potential molecular targets for development of novel anti-cancer drugs, we have been analyzing expression profiles of clinical samples from cancer patients using a genome-wide cDNA microarray. In experiments with colon-cancer cells, the gene encoding fibroblast growth factor 18 (*FGF18*) was among those that showed elevated expression. The promoter region of this gene was found to contain putative Tcf4-binding motifs; moreover a reporter-gene assay using the luciferase activity as a marker, as well as an electromobility-shift assay, indicated that *FGF18* is a downstream transcription target in the β-catenin/Tcf4 pathway. We showed that exogenous FGF18 promoted growth of NIH3T3 cells in an autocrine manner, and that transfection of *FGF18* siRNAs suppressed growth of colon-cancer cells in culture. Our results indicate that FGF18 is activated in

colon cancers as a direct downstream target of the Wnt signaling pathway, and that it might represent a marker for early diagnosis and a molecular target for treatment of this life-threatening tumor.

To uncover mechanisms underlying progression of colorectal carcinogenesis and to identify genes associated with liver metastasis, we analyzed expression profiles of 14 primary colorectal cancers (CRCs) with liver metastases, and compared them with profiles of 11 non-metastatic carcinomas and those of 9 adenomas of the colon. A hierarchical cluster analysis using data from a cDNA microarray containing 23,040 genes indicated that the cancers with metastasis had different expression profiles from those without metastasis, although a number of genes were commonly up-regulated in primary cancers of both categories. We documented 54 genes that were frequently up-regulated and 375 that were frequently down-regulated in primary tumors with metastases to liver, but not in tumors without metastasis. Subsequent quantitative PCR experiments confirmed that *PRDX4, CKS2, MAGED2*, and an EST (GenBank accession number W38659) were expressed at significantly higher levels in tumors with metastasis. These data should contribute to a better understanding of the progression of colorectal tumors, and facilitate prediction of their metastatic potential.

Gastric cancer is the fourth leading cause of cancer-related death in the world. Two histologically distinct types of gastric cancers, namely intestinal type and diffuse type, have different epidemiological and pathophysiological features, suggesting different mechanisms of carcinogenesis. A number of studies have been carried out to determine the molecular mechanisms of intestinal-type gastric cancer, whereas little is known about those of diffuse-type gastric cancer that has a more invasive phenotype and poorer prognosis. To clarify the mechanisms that underlie development and/or progression of diffuse-type gastric cancer, we compared the expression profiles of 20 diffuse-type gastric cancer tissues with their corresponding non-cancerous mucosae by means of cDNA microarray containing 23,040 genes in combination with laser-microbeam microdissection. We identified 153 genes commonly up-regulated and more than 500 genes commonly down-regulated. Furthermore, comparison of the expression profiles of diffuse-type with those of intestinal-type gastric cancers identified 46 genes that may represent the distinct molecular signature of each type of gastric cancer. The signature of diffuse-type cancer exhibits altered expression of genes related to cell-matrix interaction and extracellular matrix components. Although further investigation of their functions is essential, these data should help to better understand the different mechanisms underlying gastric carcinogenesis and may also provide clues to the identification of novel diagnostic markers and/or therapeutic targets of diffuse-type gastric cancer.

### c. *PTEN*-signaling pathway

**Motoko Unoki and Yusuke Nakamura**

EGR2 plays a key role in the PTEN-induced apoptotic pathway. Using adenovirus-mediated gene transfer to 39 cancer-cell lines, we found that EGR2 could induce apoptosis in a large proportion of these lines by altering the permeability of mitochondrial membranes, releasing cytochrome c and activating caspases-3, -8, and -9. Analysis by cDNA microarray and subsequent functional studies revealed that EGR2 directly transactivates expression of *BNIP3L* and *BAK*. Our results helped to clarify the molecular mechanism of the apoptotic pathway induced by PTEN-EGR2, and suggested that EGR2 may be an excellent target molecule for gene therapy to treat a variety of cancers.

We demonstrated several lines of evidence indicating that the early growth-response 2 (EGR 2) functions as a tumor suppressor, partly on the basis that its expression was often decreased in human tumors and cancer-cell lines. Here we report a possible molecular mechanism to account for down-regulation of EGR2 in tumor cells. Although no genetic mutations in the gene nor alterations in methylation status of its promoter were detected, we found a high degree of methylation at CpG islands in intron1 of *EGR2* in cell lines that were expressing this gene at a high level. Moreover, reporter-gene experiments revealed that methylated intron1 had somehow conferred enhancer-like activity. The data implied the existence of a previously unsuspected mechanism of gene-expression regulation.

### d. cDNA microarray analysis of cancer

**Toyomasa Katagiri, Yataro Daigo, Hidewaki Nakagawa, Yoichi Furukawa, Hitoshi Zembutsu, Takehumi Kikuchi, Soji Kakiuchi, Toru Nakamura, Koichi Okada, Yasuyuki Kaneta, Satoshi Nagayama, Takahide Arimoto, Shingo Ashida, Toshihiro Nishidate, Chie Suzuki, Nobuhisa Ishikawa, Tatsuya Kato, Akira Togashi, Satoshi Hayama, Megumi Iiizumi, Keisuke Taniuchi, and Yusuke Nakamura**

(1) Lung cancer

To investigate genes involved in pulmonary

carcinogenesis and those related to sensitivity of non-small cell lung cancers (NSCLCs) to therapeutic drugs, we performed cDNA microarray analysis of 37 NSCLCs after laser-capture microdissection of cancer cells from primary tumors. A clustering algorithm applied to the expression data easily distinguished two major histological types of non-small cell lung cancer, adenocarcinoma and squamous cell carcinoma. Subsequent analysis of the 18 adenocarcinomas identified 40 genes whose expression levels could separate cases with lymph node metastasis from those without metastasis. In addition, we compared the expression data with measurements of the sensitivity of surgically dissected NSCLC specimens to six anticancer drugs (docetaxel, paclitaxel, irinotecan, cisplatin, gemcitabine, and vinorelbine), as measured by the CD-DST (collagen gel droplet embedded culture-drug sensitivity test) method. We found significant associations between expression levels of dozens of genes and chemosensitivity of NSCLCs. Our results provide valuable information for eventually identifying predictive markers and novel therapeutic target molecules for this type of cancer.

Using the above information, we obtained evidence that the cytochrome c oxidase assembly protein COX17 is a potential molecular target for treatment of lung cancers. By semi-quantitative RT-PCR, we documented increased expression of *COX17* in all of eight primary NSCLCs and in 11 of 15 NSCLC cell lines examined, by comparison with normal lung tissue. Treatment of NSCLC cells with antisense S-oligonucleotides or vector-based small interfering RNAs (siRNAs) of *COX17* suppressed expression of COX 17 and also the activity of cytochrome c oxidase (CCO), and suppressed growth of the cancer cells. As our data imply that up-regulation of COX17 function and increased CCO activity are frequent features of lung carcinogenesis, we suggest that selective suppression of components of the CCO complex might hold promise for development of a new strategy for treating lung cancers.

Although a number of molecules have been implicated in the process of cancer metastasis, the organ-selective nature of cancer cells is still poorly understood. To investigate this issue, we established a metastasis model in mice with multiple organ dissemination by intravenous injection of human SCLC (SBC-5) cells. We analyzed gene-expression profiles of 25 metastatic lesions from four organs (lung, liver, kidney and bone) using a cDNA microarray representing 23,040 genes, and extracted 435 genes that seemed to reflect the organ specificity of the metastatic cells and the cross-talk between cancer cells and microenvironment. Furthermore we discovered 105 genes that might be involved in the incipient stage of secondary-tumor formation by comparing the gene-expression profiles of metastatic lesions classified according to size (<1mm or >2mm) as either "micrometastases" or "macrometastases". This genome-wide analysis should contribute to a greater understanding of molecular aspects of the metastatic process in different microenvironments, and provide indicators for new strategies to predict and prevent metastasis.

(2) Chemosensitivity

One of the most critical issues to be solved in regard to cancer chemotherapy is the need to establish a method for predicting efficacy or toxicity of anti-cancer drugs for individual patients. To identify genes that might be associated with chemosensitivity, we used a cDNA microarray representing 23,040 genes to analyze expression profiles in a panel of 85 cancer xenografts derived from nine human organs. The xenografts, implanted into nude mice, were examined for sensitivity to nine anti-cancer drugs (vinblastine (VLB), vincristine (VCR), cisplatin (DDP), cyclophosphamide (CPM), 5-fluorouracil (5FU), nitrosourea hydrochloride (ACNU), mitomycin C (MMC), methotrexate (MTX), adriamycin (ADR). Comparison of the gene-expression profiles of the tumors with sensitivities to each drug identified 1578 genes whose expression levels correlated significantly with chemosensitivity; 333 of those genes showed significant correlation with two or more drugs and 32 correlated with six or seven drugs. These data should contribute useful information for identifying predictive markers for drug sensitivity that may eventually provide "personalized chemotherapy" for individual patients as well as for development of novel drugs to overcome acquired resistance of tumor cells to chemical agents.

One of the most critical issues to be solved in regard to cancer chemotherapy is establishment of ways to predict efficacy of anti-cancer drugs for individual patients. To develop a prediction system based on expression of specific genes, we analyzed expression profiles of mononuclear cells from 18 patients with chronic myeloid leukemia (CML) who were treated with the tyrosine kinase inhibitor STI571. cDNA microarrays representing 23,040 genes identified 79 genes that were expressed differentially between responders and non-responders to STI571. On the basis of expression patterns of 15 or 30 of these genes among the patients we used, a "Prediction Score" system that could clearly separate the responder group from the non-responder group. Verification of this system using five additional

("test") cases succeeded in predicting the response of each of those five patients to the drug, with 100% accuracy. These results provide the first evidence that gene-expression profiles can predict sensitivity of CML cells to STI571, and may eventually lead to achievement of "personalized therapy" for this disease

To establish a method for predicting the response to chemotherapy for osteosarcoma (OS), we performed expression profile analysis using cDNA microarray consisting of 23,040 genes. Hierarchical clustering based on the expression profiles of 19 biopsy samples of OS demonstrated two major clusters; one consisted exclusively of typical OS, i.e. conventional central OS in long bone of patients in the second decade and the other in the other types of bones in rather middle age. A set of genes was identified to characterize this subgroup, some of which were previously indicated possible relation to carcinogenesis of osteosarcoma. Thirteen of the 19 patients were treated with an identical protocol of chemotherapy with doxorubicin, cisplatin and ifosfamide, and histological examination of resected specimens after operation classified six cases as responders and seven as non-responders. A comparison of expression profiles of these two groups identified 60 genes whose expression levels were likely to be correlated with the response to this particular chemotherapy ($P$ value of $<0.008$). We developed a drug response scoring (DRS) system on the basis of the expression levels of these genes, and proved this system may be applicable to predict the response to this protocol irrespective to the subclassification of OS. The reliability of the DRS system was further confirmed by testing additional five OS cases. These results indicated that scoring system based on gene-expression profiles might be useful to predict the response to chemotherapy for OS.

(3) Endometriosis

Using a cDNA microarray consisting of 23,040 genes, we analyzed gene-expression profiles of ovarian endometrial cysts from 23 patients in order to identify genes involved in endometriosis. By comparing expression patterns between endometriotic tissues and corresponding eutopic endometria, we identified 15 genes that were commonly up-regulated in the endometrial cysts during both proliferative and secretory phases of the menstrual cycle, 42 that were up-regulated only in the proliferative phase, and 40 that were up-regulated only in the secretory phase. The up-regulated elements included genes encoding some HLA antigens, complement factors, ribosomal proteins, and TGFBI. On the other hand, 337 genes were commonly down-regulated throughout the menstrual cycle, 144 only in the proliferative phase, and 835 only in the secretory phase. The down-regulated elements included the tumor suppressor TP53, genes related to apoptosis such as GADD34, GADD45A, GADD45B and PIG11, and the gene encoding OVGP1, a protein involved in maintenance of early pregnancy. Semi-quantitative RT-PCR experiments supported the results of our microarray analysis. These data should provide useful information for finding candidate genes whose products might serve as molecular targets for diagnosis or treatment of endometriosis.

(4) Testicular cancer

To identify new diagnostic markers for testicular germ cell tumors (TGCTs), including seminomas, as well as potential targets of new drugs for treating the disease, we compared gene-expression profiles of cancer cells from 13 seminomas with normal human testis using laser-capture microdissection and a cDNA microarray representing 23,040 genes. We identified 349 genes that were commonly up-regulated in seminoma cells. The functions of 227 were known to some extent; the remaining 122 included 57 ESTs. On the list were cyclin D2 (CCND2), prostate cancer over-expressed gene 1 (POV1), and junction plakoglobin (JUP), all of which were already known to be over-expressed in seminomas. On the other hand, our protocol selected 593 genes as being commonly down-regulated in seminoma cells. That list included 340 functionally characterized genes; the other 253 included 131 ESTs. To confirm the expression data, we performed semi-quantitative RT-PCR experiments with nine highly up-regulated genes, and the results supported those from of our microarray analysis. The information provided here should prove useful for identifying genes whose products might serve as molecular targets for treatment of TGCTs.

## 2. Genes responsible for other diseases

### a. Deafness

**Satoko Abe, Toyomasa Katagiri, Akihiko Saito -Hisaminato, and Yusuke Nakamura**

Hearing loss that disturbs normal communication is a common sensory disorder worldwide. The incidence of congenital deafness is approximately one in 1,000 newborns, and half of those cases are thought to result from genetic factors. Most congenital or childhood-onset hearing impairments are non-syndromic. So far, more than 70 genetic loci linked to non-syndromic deafness have been described, and 26 genes whose muta-

tions can cause deafness have been cloned (Hereditary Hearing Loss Homepage). Those data indicate that deafness is a highly heterogeneous disorder, and that genes responsible for deafness encode a large diversity of molecules. However, little is known of the molecular basis of inner-ear function because the tissues in question are too small to be investigated in detail. Therefore, we applied a genome-wide cDNA microarray analysis to investigate gene-expression profiles in human cochlea and vestibule, and focused on one of the genes that was expressed at high levels in both of those tissues. Through this approach, we detected strong expression of $\mu$-crystallin (*CRYM*; also known as NADP-regulated thyroid hormone-binding protein) only in these inner-ear tissues. In a subsequent search for mutations of *CRYM* among 192 patients with non-syndromic deafness, we identified two mutations at the C-terminus; one was a *de novo* change (X315Y) in a patient with unaffected parents and the other was a missense mutation (K 314T) that segregated dominantly in the proband's family. When the mutated proteins were expressed in COS-7 cells, their sub-cellular localizations were different from that of the normal protein: the X315Y mutant showed vacuolated distribution in the cytoplasm and the K314 T mutant localized in perinuclear areas; normal protein was distributed homogeneously in the cytoplasm. Aberrant intracellular localization of the mutated proteins might cause dysfunction of the *CRYM* product and result in hearing impairment. *In situ* hybridization analysis using mouse tissues indicated its expression in the lateral region of the spiral ligament and the fibrocytes of the spiral limbus, implying its possible involvement in the potassium-ion recycling system. Our results strongly implicate *CRYM* in normal auditory function and identify it as one of the genes that can be responsible for non-syndromic deafness.

We report three possibly disease-causing point mutations in one of the inner-ear-specific genes, *KIAA1199*. We identified a R187C mutation in one family, an R187H mutation in two unrelated families, and a H783Y mutation in one sporadic case of non-syndromic hearing loss. *In situ* hybridization indicated that the murine homolog of *KIAA1199* mRNA is expressed specifically in Deiters' cells in the organ of Corti at postnatal day zero (P*n*) P0, before the onset of hearing, but expression in those cells disappears by day P7. The signal of *KIAA1199* was also observed in fibrocytes of the spiral ligament and the spiral limbus, through to P21 when the murine cochlea matures. Thus, the gene product may be involved in uptake of potassium ions or trophic factors with a particular role in auditory development. Although the R187C and R187H mutations did not appear to affect the subcellular localization of the gene product *in vitro*, the H783Y mutation did present an unusual cytoplasmic distribution pattern that could underlie the molecular mechanism of hearing impairment. Our data bring attention to a novel candidate for hearing loss and indicate that screening of mutations in inner-ear-specific genes is likely to be an efficient approach to finding genetic elements responsible for deafness.

## b. Bone development

**Mitsuhito Doi and Yusuke Nakamura**

Through expression profile analyses of the human mesenchymal stem cells incubated in the osteogenic supplements, we identified and characterized a novel human cDNA, *EMILIN-5* (Elastin Microfibril Interface Located proteIN-5), that is likely to play a significant role in osteogenic process. The deduced amino acid sequence of *EMILIN-5* consists of 766 amino acids with a cysteine-rich EMI domain at the $NH_2$ terminus. Western blot analysis suggested that *EMILIN-5* expression was detected in various osteoblastic cells. Immunohistochemistry of mouse embryos at 13.5 days post coitus interestingly revealed relatively high levels of *EMILIN-5* protein in perichondrium cells of developing limbs. The present findings suggest that the *EMILIN-5* gene plays an important role in mesenchymal development.

## c. IgA nephropathy

**Fumihiro Akiyama, Toshihiro Tanaka[1], Ryo Yamada[2], Yozo Ohnishi[1], Shiro Maeda[3], Tatsuhiko Tsunoda[4], Takashi Takei[5], Wataru Obara[1], Kyoko Ito[5], Kazuho Honda[5], Keiko Uchida[5], Ken Tsuchiya[5], Kosaku Nitta[5], Kazuko Yumura[5], Hiroshi Nihei[5], Takashi Ujiie[6], Yutaka Nagane[8], Satoru Miyano, Yasushi Suzuki[7], Tomoaki Fujioka[7], Ichiei Narita[9], Fumitake Gejyo[9], and Yusuke Nakamura[1)]**

1) Laboratory for Cardiovascular Diseases,
2) Laboratory for Rheumatic Diseases,
3) Laboratory for Diabetic Nephropathy, and
4) Laboratory for Medical Informatics, SNP Research Center, The Institute of Physical and Chemical Research (RIKEN), Tokyo, Japan;
5) Department of Medicine, Kidney Center, Tokyo Women's Medical University, Tokyo, Japan;
6) Department of Urology, Iwate Prefectural Ofunato Hospital, Iwate, Japan;
7) Department of Urology, Iwate Medical University, Iwate, Japan;

8) Department of Urology, Sanai Hospital, Iwate, Japan.
9) Division of Clinical Nephrology and Rheumatology, Niigata University Graduate School of Medical and Dental Sciences, Niigata, Japan;

Immunoglobulin A nephropathy (IgAN) is a primary glomerulonephritis of common incidence worldwide whose etiology and pathogenesis remain unresolved, although genetic factors are assumed to be involved in the development and progression of this disease. To identify genetic variations that might confer susceptibility to IgAN, we performed a case-control association study involving 389 Japanese IgAN patients and 465 controls. Genome-wide analysis of about 90,000 single-nucleotide polymorphisms (SNPs) identified a significant association between IgAN and six SNPs located in the PIGR (polymeric immuoglobulin receptor) gene at chromosome 1q31-q41. One of them, PIGR-17, caused an amino-acid substitution from alanine to valine at codon 580 in PIGR ($chi^2$=13.05, $p$= 0.00030, odds ratio [OR]=1.59 [95% confidence interval {CI} 1.24-2.05]); the OR of minor homozygotes to others was 2.71 (95%CI 1.31-5.61). Another SNP, PIGR-2, could affect promoter activity ($chi^2$=11.95, $p$=0.00055, OR=1.60 [95% CI 1.22-2.08]); the OR of minor homozygotes to others was 2.08 (95%CI 0.94-4.60). Pairwise analyses demonstrated that all six SNPs were in almost complete linkage disequilibrium (LD). Biopsy specimens from IgAN patients were positively stained by antibody against the secretory component of PIGR, but corresponding tissues from non-IgAN patients were not. Our results suggest that a gene associated with susceptibility to IgAN lies within or close to the PIGR locus on chromosome 1q in the Japanese population.

## Publications

S. Abe, T. Katagiri, A. Saito-Hisaminato, S. Usami, Y. Inoue, T. Tsunoda, and Y. Nakamura: Identification of CRYM as a candidate responsible for non-syndromic deafness, through cDNA microarray analysis of human cochlear and vestibular tissues. Am. J. Human Genetics, 72: 73-82, 2003

S. Tsukada, M. Iwai, J. Nishiu, M. Itoh, H. Tomoike, M. Horiuchi, Y. Nakamura, and T. Tanaka: Inhibition of experimental intimal thickening in mice lacking a novel G-protein-coupled receptor. Circulation, 313-319, 2003

S. Abe, K. Koyama, S. Usami, Y. Nakamura: Construction and characterization of a vestibular -specific cDNA library using T7-based RNA amplification. Journal of Human Genetics, 48: 142-149, 2003

A. Iida, T. Tanaka, and Y. Nakamura: High-density SNP map of human ITR, a gene associated with vascular remodeling. Journal of Human Genetics, 48: 170-172, 2003

A. Iida, and Y. Nakamura: High-resolution SNP map in the 55-kb region containing the selectin gene family on chromosome 1q24-q25. Journal of Human Genetics, 48: 150-154, 2003

T. Kikuchi, Y. Daigo, T. Katagiri, T. Tsunoda, K. Okada, S. Kakiuchi, H. Zembutsu, Y. Furukawa, M. Kawamura, K. Kobayashi, K. Imai, and Y. Nakamura: Expression profiles of non-small cell lung cancers on cDNA microarrays: Identification of genes for prediction of lymph node metastasis and sensitivity to anticancer drugs. Oncogene, 22: 2192-2205, 2003

T. Arimoto, T. Katagiri, K. Oda, T. Tsunoda, T. Yasugi, Y. Osuga, H. Yoshikawa, O. Nishii, T. Yano, Y. Taketani and Y. Nakamura: Genome-wide cDNA microarray analysis of gene-expression profiles involved in ovarian endometriosis. International Journal of Oncology, 22: 551-560, 2003

A. Iida, S. Saito, A. Sekine, C. Mishima, Y. Kitamura, K. Kondo, S. Harigae, S. Osawa, and Y. Nakamura: Catalog of 668 SNPs detected among 31 genes encoding potential drug targets on the cell surface. Journal of Human Genetics, 48: 23-46, 2003

M. Unoki, and Y. Nakamura: EGR2 induces apoptosis in various cancer-cell lines by direct transactivation of BNIP3L and BAK. Oncogene, 22: 2172-2185, 2003

K. Ochi, A. Saito-Hisaminato, Y. Daigo, T. Katagiri, Y. Toyama, H. Matsumoto and Y. Nakamura: Expression profiles of two types of human knee-joint cartilage. Journal of Human Genetics, 48: 177-182, 2003

C. Tanikawa, K. Matsuda, S. Fukuda, Y. Nakamura, and H. Arakawa: p53RDL regulates p53 -dependent apoptosis. Nature Cell Biology, 5: 216-223, 2003

G. Watanabe, H. Nishimori, H. Irifune, Y. Sasaki, S. Ishida, H. Zenbutsu, T. tanaka, S. Kawaguchi, T. Wada, J. hata, M. Kusakabe, K. Yoshida, Y, Nakamura, and T. Tokino: Induction of Tenascin-C by tumor-specific EWS-ETS fusion genes. Genes Chromosomes and Cancer, 36: 224-232, 2003

M. Unoki, J. Okutsu, and Y. Nakamura: Identification of a novel human gene, ZFP91, involved in acute myelogenous leukemia. International Journal of Oncology, 22: 1217-1223,

2003
R. Yamada, T. Miyazaki, L-M. Lu, M. Ono, M. T. Ito, M.Terada, S. Mori, K. Hata, Y. Nozaki, S. Nakatsuru, Y. Nakamura, M. Onji, and M. Nose: Genetic basis of tissue specificity of vasculitis in MRL/lpr mics. Arthritis & Rheumatism, 48: 1445-1451, 2003
S. Saito, A. Iida, A. Sekine, S. Kawagichi, S. Higuchi, C. Ogawa, and Y. Nakamura: Catalog of 680 variations among 8 cytochrome P 450 (CYP) genes, nine esterase genes, and two other genes in the Japanese population. Journal of Human Genetics, 48: 249-270, 2003
R. Ohno, and Y. Nakamura: Prediction of response to Imatinib by cDNA microarray analysis. Seminors in Hematology, 40: 42-49, 2003
S. Kakiuchi, Y. Daigo, T. Tsunoda, S. Yano, S. Sone, and Y. Nakamura: Genome-wide analysis of organ-preferential metastasis of human small cell lung cancer in mice. Molecular Cancer Research, 1: 485-499, 2003
H. Zembutsu, Y. Ohnishi, Y. Daigo, T. Katagiri, T. Kikuchi, S. Kakiuchi, K. Hirata and Y. Nakamura: Gene-Expression profiles of human tumor xenografts in nude mice treated orally with the EGFR tyrosine kinase inhibitor ZD1839. International Journal of Oncology, 23: 29-39, 2003
W. Obara, A. Iida, Y. Suzuki, T. Tanaka, F. Akiyama, S Maeda, Y Ohnishi, R. Yamada, A. Sekine, T. Tsunoda, T. Takei, K. Ito, K. Honda, K. Uchida, K. Tsuchiya, W. Yumura, T. Ujiie, Y. Nagane, K. Nitta, S. Miyano, I. Narita, F. Gejyo, H. Nihei, T. Fujioka, and Y. Nakamura: Association of single-nucleotide polymorphisms in the polymeric immunoglobulin receptor gene with Immunoglobulin A nephropathy (IgAN) in Japanese patients. Journal of Human Genetics, 48: 293-299, 2003
H. Okabe, S. Satoh, Y. Furukawa, T. Kato, S. Hasegawa, Y. Nakajima, Y. Yamaoka, and Y. Nakamura: Involvement of PEG10 in human hepatocellular carcinogenesis through interaction with SIAH1. Cancer Research, 63: 3043-3048, 2003
T. Takeda, M. Kondo, J. Sasaki, H. Kurahashi, H. Kano, K. Arai, K. Misaki, T. Fukui, K. Kobayashi, M. Tachikawa, M. Imamura, Y. Nakamura, T. Shimizu, T. Murakami, Y. Sunada. T. Fujikado, K. Matsumura, T. Terashima and T. Toda: Fukutin is required for maintenance of muscle integrity, cortical histiogenesis and normal eye development. Human Molecular Genetics, 12: 1449-1459, 2003
C.-C. Ng, H. Arakawa, S. Fukuda, H. Kondoh, and Y. Nakamura: p53RFP, a p53-inducible RING-finger protein, regulates the stability of p21WAF1. Oncogene, 22: 4449-4458, 2003
T. Kimura, M. Gotoh, Y. Nakamura and H. Arakawa: hCDC4b, a regulator of Cyclin E, as a direct transcriptional target of p53. Cancer Science, 94: 431-436, 2003
Y. Kaneta, Y. Kagami, T. Tsunoda, R. Ohno, Y. Nakamura, and T. Katagiri: Genome-wide analysis of gene-expression profiles in chronic myeloid leukemia cells using a cDNA microarray. International Journal of Oncology, 23: 681-691, 2003
A. Iida, and Y. Nakamura: Japanese Efforts in Pharmacogenomics. Current Pharmacogenomics, 1: 203-215, 2003
K. Okada, T. Katagiri, T. Tsunoda, Y. Mizutani, Y. Suzuki, M. Kamada, T. Fujioka, T. Shuin, T. Miki, and Y. Nakamura: Analysis of gene-expression profiles in testicular seminomas using a genome-wide cDNA microarray. International Journal of Oncology, 23: 1615-1635, 2003
K. Ueda, H. Arakawa and Y. Nakamura: Dual-Specificity Phosphatase 5 (DUSP5) as a direct transcriptional target of tumor suppressor p 53. Oncogene, 22: 5586-5591, 2003
T. Nakatsura, Y. Yoshitake, S. Senju, M. Monji, H. Komori, Y. Motomura, S. Hosaka, T. Beppu, T. Ishiko, H. Kamohara, H. Ashihara, T. Katagiri, Y. Furukawa, S. Fujiyama, M. Ogawa, Y. Nakamura, and Y. Nishimura: Glypican-3, over-expressed specifically in human hepatocellular carcinoma, is a novel tumor marker. Biochem. Biophy. Research Comm., 306: 16-25, 2003
A. Iida, K. Ozaki, Y. Ohnishi, T. Tanaka and Y. Nakamura: Identification of 46 novel SNPs in the 130-kb region containing a myocardial infarction susceptibility gene on chromosomal band 6p21. Journal of Human Genetics, in press, 2003
T. Shimokawa, Y. Furukawa, M. Sakai, M. Li, N. Miwa, Y. -M. Lin and Y. Nakamura: Involvement of the FGF18 gene in colorectal carcinogenesis, as a novel downstream target of the β -catenin/T-cell factor complex. Cancer Research, 63: 6116-6120, 2003
A. Suzuki, R. Yamada, X. Chang, S. Tokuhiro, T. Sawada, M. Suzuki, M. Nagasaki, M. Nakayama-Hamada, R. Kawaida, M. Ono, M. Ohtsuki, H. Furukawa, S. Yoshino, M. Yukioka, S. Touma, T. Matsubara, S. Wakitani, R. Teshima, A. Sekine, A. Iida, A. Takahashi, T. Tsunoda, Y. Nakamura, and K. Yamamoto: Functional haplotypes of PADI4, encoding citrullinating enzyme peptidylarginine deiminase 4, are associated with rheumatoid arthritis. Nature Genetics, 34: 395-402, 2003
T. Kimura, S. Takeda, Y. Sagiya, M. Gotoh, Y. Nakamura and H. Arakawa: Impaired func-

tion of p53R2 in Rrm2b-null mice causes severe renal failure through attenuation of dNTP pools. Nature Genetics, 34: 440-445, 2003

S. Saito, A. Iida, A. Sekine, S. Kawauchi, S. Higuchi, C. Ogawa, and Y. Nakamura: Catalog of 178 variations in the Japanese population among eight human genes encoding G protein-coupled receptors (GPCRs). Journal of Human Genetics, 48: 461-468, 2003

N. Tanaka, T. Babazono, S. Saito, A. Sekine, T. Tsunoda, M. Haneda, Y. Tanaka, T. Fujioka, K. Kaku, R. Kawamori, R. Kikkawa, Y. Iwamoto, Y. Nakamura and S. Maeda: Associationof solute carrier family 12 (sodium/chloride) member 3 with diabetic nephropathy, identified by genome-wide analyses of Single Nucleotide Polymorphisms. Diabetes, 52: 2848-2853, 2003

C. Suzuki, Y. Daigo, T. Kikuchi, T. Katagiri and Y. Nakamura: Identification of COX17 as a therapeutic target for non-small cell lung cancer. Cancer Research, in press, 2003

S. Abe, S. Usami, and Y. Nakamura: Mutations in the gene encoding KIAA1199 protein, an inner-ear protein expressed in Deiters' cells and the fibrocytes, as the cause of nonsyndromic hearing loss. Journal of Human Genetics, in press, 2003

M. Unoki, and Y. Nakamura: Methylation at CpG-islands in intron1 of EGR2 confers enhancer-like activity. FEBS letters, in press, 2003

H. Okabe, Y. Furukawa, T. Kato, S. Hasegawa, Y. Yamaoka, and Y. Nakamura: Isolation of DDEFL1 (Development and Differentiation Enhancing Factor-Like 1) as a drug target for hepatocellular carcinomas. International Journal of Oncology, in press, 2003

Y. Anazawa, H. Arakawa, H. Nakagawa, and Y. Nakamura: Identification of STAG1 as a key mediator of a p53-dependent apoptotic pathway. Oncogene, in press, 2003

S. Tokuhiro, R. Yamada, X. Chang, A. Suzuki, Y. Kochi, T. Sawada, M. Suzuki, M. Nagasaki, M. Ohtsuki, M. Ono, H. Furukawa, M. Nagashima, S. Yoshino, A. Mabuchi, A. Sekine, S. Saito, A. Takahashi, T. Tsunoda, Y. Nakamura and K. Yamamoto: An intronic SNP in a RUNX1 binding site of SLC22A4, encoding an organic cation transporter, is associated with rheumatoid arthritis. Nature Genetics, in press, 2003

K. Yoshida, M. Monden, Y. Nakamura and H. Arakawa: Adenovirus-mediated p53AIP1 gene transfer as a new strategy for treatment of p53 -resistant tumors. Cancer Science, in press, 2003

M. Li, Y.-M. Lin, S. Hasegawa, T. Shimokawa, K. Murata, M. Kameyama, O. Ishikawa, T. Katagiri, T. Tsunoda, Y. Nakamura and Y. Furukawa: Genes associated with liver metastasis of colon cancer, identified by genome-wide cDNA microarray. International Journal of Oncology, in press, 2003

S. Nagayama, M. Iiizumi, T. Katagiri, J. Toguchida and Y. Nakamura: Identification of PDZK4, a novel human gene with PDZ domains, that is up-regulated in synovial sarcomas. Oncogene, in press, 2003

K. Ochi, Y. Daigo, T. Katagiri, S. Nagayama, T. Tsunoda, A. Myoui, N. Naka, N. Araki, I. Kudawara, M. Ieguchi, Y. Toyama, J. Toguchida, H. Yoshikawa and Y. Nakamura: Prediction of response to neoadjuvant chemotherapy for osteosarcoma by gene-expression profiles. International Journal of Oncology, in press, 2003

Y. Nakamura: Isolation of p53-target genes and their functional analysis (Review). Cancer Science, in press, 2004

T. Sekiya, S. Adachi, K. Kohu, T. Yamada, O. Higuchi, Y. Furukawa, Y. Nakamura, T. Nakamura, K. Tashiro, S. Kuhara, S. Ohwada, and T. Akiyama: Identification of BAMBI, an inhibitor of TGF-β signaling, as a target of the β-catenin pathway in colorectal tumor cells. Journal of Biological Chemistry, in press, 2003

M. Doi, A. Nagano, and Y. Nakamura: Molecular Cloning and Characterization of a Novel Gene, EMILIN-5, and Its Possible Involvement in Skeletal Development. Biochem. Biophy. Research Comm., in press, 2003

K. Yoon, Y. Nakamura and H. Arakawa: p53 directly transactivates ALDH4 by various cellular stresses. Journal of Human Genetics, in press, 2003

The International HapMap Consortium: The International HapMap Project. Nature, 426: 789-796, 2003

T. Nakamura, Y. Furukawa, T. Tsunoda, H. Ohigashi, K. Murata, O. Ishikawa, K. Ohgaki, N. Kashimura, M. Miyamoto, S. Hirano, S. Kondo, H. Katoh, and Y. Nakamura, and T. Katagiri: Genome-wide cDNA microarray analysis of gene-expression profiles in pancreatic cancers Using populations of tumor cells and normal ductal epithelial cells selected for purity by laser microdissection. Oncogene, in press

*Human Genome Center*

# Laboratory of Functional Genomics
## ゲノム機能解析分野


| | | | |
|---|---|---|---|
| Professor | Yoshiyuki Sakaki, Ph.D. | 教　授 | 理学博士　榊　　佳　之 |
| Associate Professor | Hajime Tei, Ph.D. | 助教授 | 農学博士　程　　肇 |
| Research Associate | Yuriko Hagiwara-Takeuti, Ph.D. | 助　手 | 理学博士　萩原(竹内)百合子 |


*We are focusing to sequence-based comparative analysis of human genome, the construction of gene (protein)-gene (protein) interaction map, and molecular mechanism regulating mammalian circadian rhythms, and*

### 1. Comparative genomics of the human genome

**Kunihiko Takamatsu, Kouhei Maekawa, Tomoyo Shirakawa, Kohji Okamura, Tadayuki Takeda[1], Masahira Hattori[1], Todd Taylor[1] and Yoshiyuki Sakaki ([1]RIKEN, Genomic Sciences Center, Yokohama)**

Our group has made considerable contribution to the International Human Genome Sequencing Project, and the Project reached the final goal on April 2003. However, the knowledge obtained from human genome sequence alone (even if completely determined) is limited. One powerful approaches to zoom up important regions of the genome is comparative genomics approach. For this reason, we have done two types of comparative analysis, the one, mouse vs human and the other, chimpanzee vs human. Human-mouse comparison is expected to reveal "conserved" regions of the genome that have potentially important functions. We have done mouse chromosome 16 (MMU16) vs human chr 21q comparison. Our first target is a commonly described "DS critical region," and we found all known genes, plus 144 conserved sequences (CSs)≥100bp long that show≥80% identity between mouse and human but do not match known exons. EST and cDNA evidence indicated that twenty of these 144 CSs are transcribed sequences from chr 21. Eight putative CpG islands are found in conserved positions. Using conditions for comparative sequence analysis that identified portions of every previously identified gene in the region, two HSA21 genes, *DSCR4* and *DSCR8*, did not have counterparts at the corresponding positions on MMU 16 nor elsewhere in the mouse genome. Following zoo blot analysis suggested there genes are primate-specific. We also started human vs chimpanzee comparison, which will zoom up the difference of the two genoms. Such differences must be related to the phenotype difference of the two species. Our initial analysis showed the sequence difference is 1.23% and we found a number of genes are inactivated during the course of evolution.

### 2. A comprehensive analysis of allelic methylation status of CpG islands on human chromsome 11.

**Tomoyo Shirakawa, Todd Taylor[1], Aya**

**Nakayama, Yuriko Hagiwara-Takeuchi, Takashi Ito[2] and Yoshiyuki Sakaki ([1]RIKEN, Genomic Sciences Center, Yokohama, [2]Graduate school of Frontier Sciences, Univ of Tokyo)**

Imprinted genes are often associated with DNA regions subject to allele-specific methylation, termed differentially methylated region (DMR) or methylation imprints, which often share structural features such as tandemly repeats and CpG island-like base composition. Because CpG islands generally lie near promoter regions and escape methylation, monoal-lelically methylated DMRs and the islands on inactivated X chromosome in female stand for exceptions: cells thus bear both methylated and unmethylated alleles for these islands and hence display composite pattern upon methylation analysis. It is thus conceivable that screening for CpG islands with composite methylation pattern serves as a novel method to identify DMRs and associated imprinted genes. To screen for such islands, we used HpaII-McrBC PCR method by exploiting the complementary nature of HpaII, which cuts the un-methylated DNA, and McrBC, which digests the methylated one. We applied the method to a comprehensive methylation analysis of 657 CpG islands on human chromosome 11. While most islands escape methylation as expected (83%), 6 CpG islands display composite methylation pattern. Intriguingly, two of the composite CGIs were methylated in an allele-specific but parental-origin-independent manner. Comparing the distribution of methylated and unmethylated CpG islands on chromosome 11 and 21, we also demonstrated that the distribution pattern of CpG islands may differ chromosome to chromosome. This approach would provide a novel way to identify DMRs or methylation imprints and hence novel imprinted genes in the human genome and may contribute to uncover novel modes of allelic methylation.

## 3. Analysis of a solitary imprinted gene Impact

**Kohji Okamura, Takashi Ito[1] and Yoshiyuki Sakaki ([1]Graduate school of Frontier Sciences, Univ of Tokyo)**

We also analyzed the mouse imprinted gene Impact isolated in our laboratory by allelic message display. Mouse Impact, the sole imprinted gene mapped to chromosome 18, lies between Hrh4 and Osbpl1. While most imprinted genes show close physical clustering, both Hrh4 and Osbpl1 failed to show any evidence for imprinting, suggesting that Impact is a solitary imprinted gene. Solitary imprinted genes identified so far share common features such as fewer introns and occurrence in singular introns of other genes, indicative of retrotransposition. In contrast, Impact lies in an intergenic region and consists of 11 exons, thereby strongly arguing against its retrotranspositional origin. Based on hazardous effects of overexpressed Impact, a genomic segment containing paralogues of Hrh4 and Osbpl1, and strong promoter activity in mouse, we assume that the species-specific solitary imprinting of Impact evolved as a series of response to segmental duplication followed by lineage-specific enhancement of promoter activity.

## 4. Molecular mechanisms regulating mammalian circadian clock

**Hajime Tei, Rika Numano, Nobuya Koike, Shihoko Kojima, Soshi, Kawaguchi, Atsuki Shinozaki, Matsumi Hirose, Miyuki Shimada, Aya Nakayama and Yoshiyuki Sakaki**

Many biochemical, physiological and behavioral processes in many organisms exhibit circadian rhythms. Circadian rhythms are driven by autonomous oscillators and entrained by daily light-dark cycles. The transcription of *Per1* and *Per2*, two mammalian clock genes, oscillates in a circadian manner in the mouse suprachiasmatic nucleus (SCN; the central pacemaker of the mammalian circadian clock) with a peak in the daytime and a trough at night. In addition, the expression of m*Per1* and m*Per2* in the SCN is induced immediately by a light pulse even at night. The function of the circadian expression of the mammalian *Per1* and *Per2* genes is a key question for the regulation of circadian rhythms. For elucidation of the molecular mechanisms controlling the mammalian circadian clock, the genomic sequences of the human and mouse *Per 1* and *Per2* genes in addition to their transcriptional start sites have been determined. Both of the h*Per1* and m*Per1* genomic sequences consisted of 23 exons spanning approximately 16 kb. Comparisons of both genes revealed five and one conserved segments in the 5′ flanking regions and the first introns, respectively. These conserved segments contained several potential regulatory elements such as five E-boxes (the binding site for the Clock-Bmal1 complex). Transfection analyses using a series of deletion and point mutants of the m*Per1::luc* reporter showed that each of the five E-boxes was functional for the *Per1* induction mediated by Clock and Bmal1. Unlike *Per1*, five conserved segments in the 5′ flanking regions of h*Per2*, m*Per 2*, and r*Per2* contained no typical E-box se-

quence. However, transfection analyses using a series of deletion and point mutants of the m*Per 2*::luc reporter showed that the sequences similar to E-box were functional for the *Per2* induction mediated by Clock and Bmal1. Subsequently, we generated a *Per1::luc* transgenic rat line in which luciferase is rhythmically expressed under the control of the mouse *Per1* promoter, and used it to study mammalian circadian organization. Light emission from cultured suprachiasmatic nuclei (SCN) of these rats was invariably and robustly rhythmic. Circadian rhythm of light emission from the SCN shifted more rapidly than did the rhythm of locomotor behavior. Liver, lung, and skeletal muscle expressed damped circadian rhythms *in vitro*. We hypothesize that self-sustained circadian oscillators in the SCN entrain damped circadian oscillators in the periphery to maintain adaptive phase control.

## Publications


Kojima, S., Hirose M., Tokunaga, K., Sakaki, Y., and Tei, H. Structural and functional analysis of 3′ untranslated region of mouse *Period1* mRNA. Biochem. Biophys. Res. Commun. 301: 1-7, 2003.

Yada, T., Totoki, Y., Takaeda, Y., Sakaki, Y. and Takagi, T. DIGIT: a novel gene finding program by combining gene-finders, Proc. of Pacific Sympo. on Biocomputing ’03, 2003.

Ishida, C., Ura, K., Hirao, A., Sasaki, H., Toyoda, A., Sakaki, Y., Niwa, H., Li, E. and Kaneda, Y. Genomic organization and promoter analysis of the Dnmt3b gene. Gene. 310: 151-159, 2003.

Xinh, P-T., Tri, N-K., Nagao, H., Nakazato, H., Taketazu, F., Fujisawa, S., Yagasaki, F., Chen, Y-Z., Hayashi, Y., Toyoda, A., Hattori, M., Sakaki, Y., Tokunaga, K. and Sato, Y. Breakpoint at 1p 36.3 in Three MDS/AML (M4) Patients With t (1; 3) (p 36; q 21) Occur in the First Intron and in the 5’ Region of MELI. Genes Chromosomes Cancer. 36: 313-316, 2003.

Choi, D-K., Yoo, K-W., Hong, S-K., Rhee, M., Sakaki, Y. and Kim, C-H. Isolation and expression of Napor/CUG-BP2 in embryo development. Biochem. Biophys. Res. Commun. 305: 448-454, 2003.

Ikeda, H., Ishikawa, J., Hanamoto, A., Shinose, M., Kikuchi, H., Shiba, T., Sakaki, Y., Hattoril, M. and Omura, S. Complete genome sequence and comparative analysis of the industrial microorganism Streptomyces avermitilis. Nat. Biotechnol. 21: 526-530, 2003.

Ohshima, K., Hattori, M., Yada, T., Gojobori, T., Sakaki, Y. and Okada, N. Whole-genome screening indicates a possible burst of formation of processed pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates. Genome Biol. 4: R74, 2003.

Kamei, H., Adati, N., Arai, Y., Yamamura, K., Takayama, M., Nakazawa, S., Ebihara, Y., Gondo, Y., Akechi, M., Noguchi, T., Hirose, N., Sakaki, Y. and Kojima, T. Association analysis of the *SHC1* gene locus with longevity in the Japanese population. J. Mol. Med. 81: 724-728, 2003.

Takaoka, Y., Ohta, M., Miyakawa, K., Nakamura, O., Suzuki, M., Takahashi, K., Yamamura, K. and Sakaki, Y. Cysteine 10 ls a Key Residue in Amyloidogenesis of Human Transthyretin Val30Met. Am. J. Pathol. 164: 337-345, 2004.

Herzog, ED., Aton, SJ., Numano, R., Sakaki, Y. and Tei, H. Related Articles, Links Abstract Temporal precision in the mammalian circadian system: a reliable clock from less reliable neurons. J. Biol. Rhythms. 19: 35-46, 2004.

*Human Genome Center*

# Laboratory of Functional Analysis *In Silico*
# 機能解析イン・シリコ分野

Professor Kenta Nakai, Ph.D.

教 授 理学博士 中 井 謙 太

*The mission of our laboratory is to conduct leading studies on the functional aspects of genome information computationally* ("in silico"). *Roughly speaking, genome information represents what kind of proteins are synthesized on what conditions. Thus, our study includes not only the analysis of molecular function of each gene product but also the analysis of regulatory information, which will lead us to the understanding of its cellular role represented by the networks of inter-gene interaction.*

## 1. DBTBS: Database of transcriptional regulation in *Bacillus subtilis* and its contribution to comparative genomics

**Yuko Makita, Mitsuteru Nakao, Naotake Ogasawara[1], and Kenta Nakai: [1]Department of Bioinformatics and Genomics, Nara Institute of Science and Technology**

DBTBS (http://dbtbs.hgc.jp) was originally released in 1999 as a reference database of published transcriptional regulation events in *Bacillus subtilis*, one of the best studied bacteria. It is essentially a compilation of transcription factors with their regulated genes as well as their recognition sequences, which were experimentally characterized and reported in literature. In this year, we tried to update the data, which contains the information of 114 transcription factors, including sigma factors, and 633 promoters of 525 genes. The number of references cited in the database increases from 291 to 378. It also newly supports a function to find putative transcription factor binding sites within input sequences by using our collection of weight matrices and consensus patterns. Furthermore, DBTBS aims to contribute to comparative genomics by showing the presence or absence of potentially orthologous transcription factors and their corresponding *cis*-elements on the potentially orthologous promoters of their regulated genes in 50 eubacterial genomes.

## 2. DBTSS (DataBase of Transcriptional Start Sites): Progress Report

**Riu Yamashita, Yutaka Suzuki[2], Sumio Sugano[2] and Kenta Nakai: Laboratory of Genome Structure Analysis**

DBTSS (http://dbtss.hgc.jp) was originally constructed based on a collection of experimentally-determined TSSs of human genes. Since its first release in 2002, it has been updated several times. First, the amount of the stored data has increased significantly: for example, the number of clones that match both to the RefSeq mRNA set and the genome sequence increases from 111,382 to 190,964, now covering 11,234 genes. Second, the positions of SNPs in dbSNP were displayed on the upstream regions of contained human genes. Third, DBTSS now covers other species such as mouse and the human malaria parasite. It will become a central database containing the data of many more species with the oligo-capping and related methods. Lastly, the database now serves for comparative promoter analyses: in the current version, comparative

views of human and mouse potentially orthologous promoters are presented with an additional function of searching potential transcription-factor binding sites, which are either conserved or diverged between species.

### 3. Genome wide analysis reveals strong correlation between CpG islands around promoters and their tissue-specificity

**Riu Yamashita, Yutaka Suzuki[2], Sumio Sugano[2], and Kenta Nakai**

There are several CpG clusters called 'CpG islands' in vertebrate genomes, and they are thought to be around promoter region. There are conflict ideas about correlation between gene expression and CpG islands in promoter region. One of the reasons of the difficulty is uncertain transcription start sites (TSSs) of the cDNA in available databases. Here we obtained reliable information of TSSs from DataBase of Transcriptional start sites (DBTSS). We could classify into 6,600 CpG positive genes and 2,619 CpG negative genes in human while 2,948 CpG positives and 1,830 CpG negatives in mouse. Combined with UniGene expressed information, we found a clear difference between the CpG positives and the negatives. The genes without CpG islands in promoter region usually are expressed with tissue-specificity. We found no significant correlation between spliced mRNA nor transcribed DNA region and tissue-specificity in both human and mouse. Our data suggest that the gene expression pattern can be classified into two major groups with CpG islands in not transcribed DNA region but promoter region.

### 4. Parameter Landscape Analysis for Common Motif Discovery Programs

**Natalia Polulyakh, Paul Horton[3], Michiko Konno[4] and Kenta Nakai: [3]Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology; [4]Graduate School of Humanity and Science, Ochanomizu University**

The identification of regulatory elements as over-represented motifs in the promoters of potentially co-regulated genes is an important and challenging problem in computational biology. Although many motif detection programs have been developed so far, they still seem to be immature practically. In particular the choice of tunable parameters is often critical to success. Thus knowledge regarding which parameter settings are most appropriate for various types of target motifs is invaluable, but unfortunately has been scarce. In this paper, we report our parameter landscape analysis of two widely-used programs (the Gibbs Sampler and MEME). Our results show that the Gibbs Sampler is relatively sensitive to the changes of some parameter values while MEME is more stable. We present recommended parameter settings for the Gibbs Sampler optimized for four different motif lengths. Thus, running the Gibbs Sampler four times with these settings should significantly decrease the risk of overlooking subtle motifs.

### 5. Large-scale analysis of alternatively-spliced protein isoforms

**Mitsuteru C. NAKAO and Kenta NAKAI**

In higher eukaryotic cells, one general mechanism to produce a variety of amino acids from a single gene is the alternative spicing. To characterize this phenomenon, we have developed an objective classification method of protein isoforms produced by alternative splicing. We then classified a number of sequences in the SWISS-PROT database into 37 patterns, among which the pattern having mutually exclusive exons in the C-terminus as observed most frequently. Generally speaking, the C-terminal side was more variable than the N-terminal side. We also found some correlation between some patterns and the presence of specific sequence motifs that are characteristic for some protein function, which means that proteins having a specific function tend to extensively use an isoformal pattern. In addition, there is a strong correlation between the terminal variations of proteins and their differential subcellular localization.

### 6. Analysis of the upstream regions of genes expressed tissue-specifically

**Katsuki Tsuritani, Riu Yamashita, Hiroyuki Aburatani[5] and Kenta Nakai: [5]Center for Collaborative Research, University of Tokyo**

Our aim is to clarify the relationship between the upstream region of human genes and their tissue specificity in the transcriptional level observed using microarray experiments. More specifically, we are developing a query system that will answer to questions like "which of the transcription factors are specifically expressed in a given tissue?" or "In which tissue will the gene with a given upstream sequence be expressed?". For example, we found that the binding site of HNF (hepatocyte nuclear factor) is observed significantly frequently on the upstream regions of genes which are specifically expressed in fetal liver and liver. More general studies are under-

going.

### 7. Analysis of free extracellular DNA sequences found in peripheral blood

**Katsuki Tsuritani, Yoshimi Homma[6] and Kenta Nakai: [6]Institute of Biomedical Sciences, Fukushima Medical University**

We analyzed 562 DNA sequences which were taken from the serum of normal peripheral blood (PBF-DNA: peripheral blood free DNA) to find clues of their production mechanism. We found that the average length of PBF-DNA is 176 bp and that their terminal two nucleotides are G/C-rich. The ratio of their origin is 69 : 26 : 3 for intergenic region, intron, and exon, respectively. In addition, they seem to have been originated from chromosome 19 in a significantly higher ratio. Our tentative conclusion is that PBF-DNAs were generated from random positions in the human genome and that only physically stable ones have survived in peripheral blood.

## Publications

Ott, S., Tamada, Y., Bannai, H., Nakai, K., and Miyano, S., Intrasplicing: analysis of long intron sequences, Pacific Symposium on Biocomputing, 8, 339-350 (2003).

Poluliakh, N., Takagi, T., and Nakai, K., MELINA: motif extraction from promoter regions of potentially co-regulated genes, Bioinformatics, 19 (3), 423-424 (2003).

Gardy, J.L., Spencer, C., Wang, K., Ester, M., Tusnady, G.E., Simon, I., Hua, S., deFays, K., Lambert, C., Nakai, K., and Brinkman, F.S.L., PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria, Nucl. Acids Res., 31 (13), 3613-3617 (2003).

Yamashita, R., Suzuki, Y., Nakai, K., and Sugano, S., Small open reading frames in 5′ untranslated regions of mRNAs, C.R. Biol., 326 (10-11), 987-991 (2003).

Makita, Y., Nakao, M., Ogasawara, N., and Nakai, K., DBTBS: Database of transcriptional regulation in Bacillus subtilis and its contribution to comparative genomics, Nucl. Acids Res., 32, D75-D77 (2004)

Suzuki, Y., Yamashita, R., Sugano, S., and Nakai, K., DBTSS (DataBase of Transcriptional Start Sites): Progress Report 2004, Nucl. Acids Res., 32, D78-D81 (2004).

Ota, T., Suzuki, Y., Nishikawa, T., Otsuki, T., Sugiyama, T., Irie, R., Wakamatsu, A., Hayashi, K., Sato, H., Nagai, K., Kimura, K., Makita, H., Sekine, M., Obayashi, M., Nishi, T., Shibahara, T., Tanaka, T., Ishii, S., Yamamoto, J., Saito, K., Kawai, Y., Isono, Y., Nakamura, Y., Nagahari, K., Murakami, K., Yasuda, T., Iwayanagi, T., Wagatsuma, M., Shiratori, A., Sudo, H., Hosoiri, T., Kaku, Y., Kodaira, H., Kondo, H., Sugawara, M., Takahashi, M., Kanda, K., Yokoi, T., Furuya, T., Kikkawa, E., Omura, Y., Abe, K., Kamihara, K., Katsuta, N., Sato, K., Tanikawa, M., Yamazaki, M., Ninomiya, K., Ishibashi, T., Yamashita, H., Murakaw,a K., Fujimori, K., Tanai, H., Kimata, M., Watanabe, M., Hiraoka, S., Chiba, Y., Ishida, S., Ono, Y., Takiguchi, S., Watanabe, S., Yosida, M., Hotuta, T., Kusano, J., Kanehori, K., Takahashi-Fujii, A., Hara, H., Tanase, T.O., Nomura, Y., Togiya, S., Komai, F., Hara, R., Takeuchi, K., Arita, M., Imose, N., Musashino, K., Yuuki, H., Oshima, A., Sasaki, N., Aotsuka, S., Yoshikawa, Y., Matsunawa, H., Ichihara, T., Shiohata, N., Sano, S., Moriya, S., Momiyama, H., Satoh, N., Takami, S., Terashima, Y., Suzuki, O., Nakagawa, S., Senoh, A., Mizoguchi, H., Goto, Y., Shimizu, F., Wakebe, H., Hishigaki, H., Watanabe, T., Sugiyama, A., Takemoto, M., Kawakami, B., Yamazaki, M., Watanabe, K., Kumagai, A., Itakura, S., Fukuzumi, Y., Fujimori, Y., Komiyama, M., Tashiro. H., Tanigami, A., Fujiwara, T., Ono, T., Yamada, K., Fujii, Y., Ozaki, K., Hirao, M., Ohmori, Y., Kawabata, A., Hikiji, T., Kobatake, N., Inagaki, H., Ikema, Y., Okamoto, S,. Okitani, R., Kawakami, T., Noguchi, S., Itoh, T., Shigeta, K., Senba, T., Matsumura, K., Nakajima, Y., Mizuno, T., Morinaga, M., Sasaki, M., Togashi, T., Oyama, M., Hata, H., Watanabe, M., Komatsu, T., Mizushima-Sugano, J., Satoh, T., Shirai, Y., Takahashi, Y., Nakagawa, K., Okumura, K., Nagase, T., Nomura, N., Kikuchi, H., Masuho, Y., Yamashita, R., Nakai, K., Yada, T., Nakamura, Y., Ohara, O., Isogai, T., Sugano, S., Complete sequencing and characterization of 21,243 full-length human cDNAs Nat. Genet., 36 (1), 40-45 (2004).

Horton, P., Mukai, Y., and Nakai, K., Chapter 14: Protein subcelular localization prediction, in L. Wong (ed.), Practical Bioinformatician, World Scientific Publishing Co., in press

中井謙太・矢田哲士，3章　遺伝子同定・シグナル同定技術，美宅成樹・榊佳之編　応用生命科学シリーズ　9：バイオインフォマティクス，東京化学同人　50-84(2003).