Human Genome Center

Laboratory of Genome Database

ゲノムデータベース分野

Professor	Toshihisa Takagi, Ph.D.
Associate Professor	Kenta Nakai, Ph.D.
Research Associate	Takako Takai, Ph.D.

教　授	工学博士	高　木　利　久
助教授	理学博士	中　井　謙　太
助　手	理学博士	高　井　貴　子

（平成13年12月迄在籍）

In analyzing human genome data, the importance of maintaining databases of various facts and knowledge is unquestionable. Thus, the main mission of our laboratory is to provide worldwide genome-research communities with useful resources, including supercomputer facilities and Internet services. Not only maintaining established databases but also the development of newer databases and technologies for better mining biological and medical content from accumulated data is our important project.

1. Development of ontology and database for the cell signaling system

Takako TAKAI and Toshihisa TAKAGI

In the post-genome sequencing era, the most significant issue is the reconstruction of living organisms in computers, based on their genome information. Reconstruction and analysis of molecular interactions among gene products, pathways, and networks could be addressed as its first step. We analyzed conceptual structure of the cell signaling system and specified the concepts as SIGNAL-ONTOLOGY. The ontology consists of concepts of molecules, molecular interactions, pathway motifs, and cellular functions. We also develop a database for the cell signaling system, SPARK, based on the ontology. The database system is constructed by XML-database, compound graph representation, and ontology. All the data contained in the database are collected from literatures according to controlled vocabulary in the ontology. SIGNAL-ONTOLOGY and SPARK are opened from http://www.ontology.jp/.

2. Signal transduction pathways and logical inferences

Ken-ichiro FUKUDA[1] and Toshihisa TAKAGI: [1]Computational Biology Research Institute

Our group focuses on providing methods required to develop Signal Transduction Pathway (STP) databases. The problem is broken down into two subproblems, i.e., knowledge representation design of STPs and its utilization to infer relevant biological hypothesis. The knowledge representation model is based on a Compound graph model and can cope with knowledge fragmentation, complex hierarchies and various levels of details on specific bodies of knowledge (heterogeneous knowledge granularity). Equipped with the ontologies for STPs, the model is able to formalize the knowledge described with natural language or drawings of diagrams. Then, the inference procedure that a biologist performs is modeled as a hypothetical reasoning framework on a case-base, and a prototype knowledge base was implemented, which infers cross-talk of pathways.

3. Kinase Pathway Database: An Integrated Resource of Automatically Extracted Information and Sequence Analysis

Asako KOIKE, Yoshiyuki KOBAYASHI, and

Toshihisa TAKAGI

Protein kinases play a crucial role in the regulation of cellular functions. Various kinds of information on these molecules are important for understanding signaling pathways and organism characteristics. We have developed the Kinase Pathway Database, an integrated database involving major completely sequenced eukaryotes. It contains the classification of protein kinases and their functional conservation, ortholog tables among species, protein-protein, protein-gene and protein-compound interaction data, domain information, and structural information. It also provides an automatic pathway graphic image interface. The protein, gene and compound interactions are automatically extracted from abstracts for all genes and proteins by natural language processing using phrase patterns and the GENA protein, gene and compound name dictionary, which was developed by our group. With this database, pathways are easily compared among species using data with more than 35,000 protein interactions and protein kinase ortholog tables. The database is

available for querying and browsing at http://kinasedb.ontology.ims.u-tokyo.ac.jp/. The automatically extracted biological function information and various pathway deductive function are also open to public in near future.

4. Simulation of linkage disequilibrium in subdivided populations

Osamu OGASAWARA and Toshihisa TAKAGI

Simulation studies were undertaken to explore strategies for detecting loci underlying rare and common disorders in subdivided population. We simulated the evolution of a many-site haplotype system (infinite-site model), under three different demographic scenarios, one with constant population size and two with population growth. The simulation was designed to observe the effects of population history, recombination fraction, and mutation rate on allele and haplotype frequencies, haplotype diversity, frequency of ancestral alleles, and linkage disequilibrium. The known ancestral haplotypes were often found at low frequencies and even became extinct after 1, 000 generations, especially with small effective population sizes.

5. Database of Human UV-regulated Genes

Goro TERAI, Makoto YAMAZAKI, Yoshinori SUGIYAMA[1], Hiroko AO, Shintaro INOUE[1], Toshihisa TAKAGI: [1]Kanebo LTD.

Ultraviolet ray (UV) is one of the most serious environmental risk factors of various skin problems such as wrinkles, stains and cancers. To make clear the mechanism of these condition of patient, it is important to know what kind of genes are activated or inactivated in UV-irradiated skin. Microarray and DNA chip techniques can measure large-scale gene expression data. But the reliability of those data is still low. So, at the first step, we have developed a database for known information about UV-regulated genes at mRNA level by surveying literatures. This database is useful for validation of comprehensive results with referring previous experimental data. Last year, we got the information about 279 genes from 113 literatures. This year, we used another strategy of data collection and could get the information about 258 genes from 119 literatures further. Out of 537 genes, 513 genes were assigned at least one Gene Ontology (GO) term for gene function annotation. We renewed Data Input System and developed useful web based analysis tools. For example, you can select genes that have similar expression pattern and get some significant gene function of this gene group. A big challenge is to integrate information on signal transduction pathways into our database. Another direction is to collect information on expression of other skin cells, such as melanocytes and fibloblasts. At present, we are planning to collect information on signal transductions from review articles.

6. Analyzing the effect of motif combinations on gene expression profiles

Goro Terai and Toshihisa Takagi

It is important to reveal combinations of cis-regulatory elements (motifs) which confer specific gene expression profiles. Computational methods, however, analyzing combinatorial effects of motifs are still in its infancy. Especially, the order and distance of motifs, which can influence interaction of transcription factors that binds their respective motifs, are rarely addressed so far. In this study, we explore the possibility of finding novel motif combinations considering their context on promoter regions. We extracted motif patterns, in which order of motifs and distance between motifs as well as kinds of motifs were specified, from upstream regions of Saccharomyces cerevisiae. To reduce computational time, we used the Sequential Pattern Mining algorithm to extract motif patterns. Then, we have revealed motif combinations based on the statistical significance of the similarity of expression profiles of genes containing each pattern. Using large-scale gene expression profiles during cell cycle and after treatment of DNA damaging agents, we found several combinations, some of which were significant only when the order and distance were considered. This result suggests that the order and distance is really important for gene regulation in eukaryotes. Because gene regulatory mechanism of higher eukaryotic organisms is supposed to be much more complicated, motif combinations play a greater role. Thus, it should be valuable to use our method for the analysis of higher eukaryotes.

7. Database for transcriptional start sites and its analysis

Riu YAMASHITA, Yutaka SUZUKI[4], Sumio SUGANO[4], and Kenta NAKAI: [4]Laboratory of Genome Structure Analysis

Transcriptional start sites (TSS) are indispensable for promoter analysis; however, there have not been sufficient data in current databases. We have constructed DataBase of Transcription Start Site (DBTSS) to specify exact start points of transcription in the human genome since last year. This year, we added mouse data to it and performed comprehensive analysis of promoter regions based on the sequences contained in DBTSS. We found that genes are classified into 3 groups based on the standard deviation of TSSs; namely, genes that have strictly regulated TSSs, genes with ambiguous TSSs, or genes with tissue-specific ones. We could obtain 251 genes that have strictly regulated TSSs from human DBTSS. Sixty-three percent of them had the TATA-box within the -34 to -28 region of their upstream sequence. These strictly-regulated TSSs included one remarkable cluster, which have a characteristic oligopyrimidine tract around the TSS, was rich in genes of translation-related protein such as ribosomal proteins. Interestingly, we found that the highest score of both the GC content and the CpG likelihood score were exactly corresponded with the TSS region. Moreover, there was a strong correlation between the existence of CpG islands and whether they are housekeeping genes or not. Our result shows basic and comprehensive features of TSSs. DBTSS should be also useful for further promoter analysis.

8. “Melina” - Novel tool for elucidation of consensus motif in the promoter regions of functionally related DNA sequences

Natalia POLULIAKH, Michiko KONNO[5], Toshihisa TAKAGI, and Kenta NAKAI: [5]Ochanomizu University

Extraction of potential cis-elements (motifs) from the upstream regions of co-regulated genes is one of the most complicated and important problems in bioinformatics. Many programs for this task are available nowadays, but one difficulty is that it is almost impossible to conclude which one is better but it is time-consuming to try all of them. Another difficulty is that the User can not even guess about the motif's characteristics, i.e., is it short or long, single or multiple, subtle or well-conserved. In trial to help many biological researches to overcome the above difficulties we constructed an User-friendly professional Analyzer, called “Melina” (Motif ELucidator In Nucleotide sequence Assembly), comprises several famous motif extraction algorithms such as Consensus, MEME, GIBBS sampler and Coresearch. Resulting outputs of the used algorithms can be compared at a glance from a very convenient graphical view, thus facilitating the validation of the extracted motifs (‘algorithm comparison’ mode). In the mode of ‘parameter optimization’ users can calculate the results for several sets of parameters, applied concomitantly to one dataset. We tested Melina on various elucidation tasks and defined optimum parameters sets for several typical situations. We insist that the proper usage of parameters can significantly improve the programs’ sensitivity to subtle motifs. Melina is opened to the public usage (http://www.hgc.ims.u-tokyo.ac.jp/Melina/) and we hope that it can be a helpful and convenient tool for many biological researchers.

9. Prediction of co-regulated genes in prokaryotic genomes by comparative genomics

Yuko MAKITA, Goro TERAI, Shigeki MITAKU[6], Toshihisa TAKAGI, and Kenta NAKAI: [6]Tokyo University of Agriculture and Technology

Short conserved sequence elements located at promoter regions are often the binding sites for transcription factors that regulate a group of genes involved in a similar cellular function. Thus, the presence of such upstream motifs can provide powerful hypotheses about links in the genetic regulatory network. These motifs can be discovered computationally by local alignment of upstream regions between orthologous genes. We have analyzed the co-regulated genes of Bacillus genus with such an approach and have found some plausible co-regulated genes. This time, we applied the same approach to additional three sets of closely related species belonging to Mycoplasma, Chlamydia and Mycobacterium genera, respectively. We could obtain some biologically plausible clusters. For instance, a cluster containing an inverted repeat sequence is tightly coupled with heat-shock related proteins in Mycoplasma. And the comparison between our results and known binding sites in Mycobacterium genus indicates that we can detect most cis-elements if there are three orthologous genes and unless elements exist on coding regions. We expect that these results can also be useful for delineating the evolution of bacterial transcriptional network.

10.Improvement of PSORT II Protein Sorting Prediction for Mammalian Proteins

Mitsuteru C. NAKAO and Kenta NAKAI

Our group focuses on improving the PSORT II system, a unique tool for the prediction of protein subcellular localization, for analyzing mammalian (human and murine) proteins. We improved PSORT II from three aspects: the employment of mammalian (murine) training data, the optimization of the learning method, and the optimization of the sequence features used. Namely, we first collected amino acid sequence data of mouse proteins the subcellular localization site of which were experimentally confirmed. This screening was performed carefully using the evidence code of the Gene Ontology project. Then, we searched the optimal combination of various sequence features semi-exhaustively. Finally, the parameters of learning methods (k-nearest neighbors rule and Support Vector Machines) were also optimized. As a result, we successfully improved PSORT II for the prediction of mammalian sequences: its total accuracy was estimated to be nearly 60% with the leave-one-out cross-validation test.

11.Sequence analysis of RNA splicing determinants

Hideo BANNAI[7], Yoshinori TAMADA[7], Sascha Ott[7], Satoru MIYANO[7] and Kenta NAKAI: [7]Laboratory of DNA Information Analysis

Elucidating the mechanism of pre-mRNA splicing is an important problem in molecular biology: for example, it has been observed that a significant fraction of mutations in human genes that cause diseases affect the pre-mRNA splicing. Although canonical splicing signals are known, accurately predicting the splice sites of 'long' introns have been difficult. We address this problem from several viewpoints: first, we found that there is some correlation between the strength of acceptor splice sites and the length of the introns containing these sites; second, we propose a novel hypothesis on the mechanism of long intron splicing, which we call intrasplicing. Third, we examine our collection of above-mentioned aberrant splicing events statistically. From these studies, we wish to establish a new hypothesis that can explain splicing events better than the exon-recognition hypothesis.

12.Database and network services for sequence interpretation and information retrieval

a. Wide Area Network

Wide-area computer network is an essential component of the infrastructure for genome research. Thus, we are collaborating with the "GenomeNet" activity at Kyoto University, in cooperation with the IMNet and WIDE computer network groups. Currently, a 6Mbps line from Tokyo to Kyoto and a 6Mbps line to the US are maintained.

b. Computer system

For database and computational services, a supercomputer system is maintained. The system includes:

* SGI-CRAY T94/4128 (vector computer)

* Hitachi SR2201 (massively parallel computer with distributed memory architecture)

* SGI-CRAY Origin2000 (parallel computer with distributed shared memory architecture)

* Sun Ultra Enterprise 10000 (parallel computer with shared memory architecture)

* Sony Petasite (mass storage tape device)

* Sun Ultra1 and SGI Octane (workstations)

c. Database services

We support various database services through the Internet (http://www.hgc.ims.u-tokyo.ac.jp/database.html). Not only standard databases of biological sequences, structures, and literature, but also locally-developed smaller databases are made publicly available by either e-mail or WWW.

Publications

Hirakawa, M., Tanaka, T., Hashimoto, Y., Kuroda, M., Takagi, T., and Nakamura, Y. :JSNP: a database of common gene variations in the Japanese population, Nucleic Acids Res., Vol.30, pp.158-162, 2002.

Yada, T., Totoki, Y., Takaeda, Y., Sakaki, Y. and Takagi, T.: DIGIT: a novel gene finding program by combining gene-finders, Proc. Pacific Sympo. on Biocomputing '03, (in press).

Suzuki, Y., Yamashita, R., Nakai, K., and Sugano, S., DBTSS: database of human transcriptional start sites and full-length cDNAs, Nucl. Acids Res., 30(1), 328-331, 2002.

Bannai, H., Tamada, Y., Maruyama, O., Nakai, K., and Miyano, S., Extensive feature detection of N-terminal protein sorting signals, Bioinformatics, 18(2), 298-305, 2002.

Nakai, K. and Vert, J.-P., Genome informatics for data-driven biology: A report on the twelfth international conference on genome informatics, Tokyo, Japan, December 17-19, 2001, Genome Biology, 3(4), reports4010.1-4010.3 , 2002.

Nakai, K. Chapter 14: Signal peptides, in Ulo Langel (ed.), Cell-Penetrating Peptides: Processes and Applications, 295-324, CRC Press, 2002.

Terashima, H., Fukuchi, S., Nakai, K., Arisawa, M., Hamada, K., Yabuki, N., and Kitada, K., Sequence-based approach for identification of cell wall proteins in Saccharomyces cerevisiae, Current Genetics, 40(5), 311-316 , 2002.

Koike, A., Nakai, K., and Takagi, T., The evolution and functional conservation of eukaryotic kinase domains among organisms, Genome Lett., 1(2), 83-104 , 2002.

Koike, A., Nakai, K., and Takagi, T. , The Origin and Evolution of Eukaryotic Protein Kinases, Genome Letters, 1, 2,83-104, 2002.

Ott, S., Tamada, Y., Bannai, H., Nakai, K., and Miyano, S., Intrasplicing: analysis of long intron sequences, Proc. PSB2003, in press.

Poluliakh, N., Takagi, T., and Nakai, K., MELINA: a web server for motif extraction from promoter regions of potentially co-regulated genes, Bioinformatics, in press.

高木利久(企画、著)．生命のシステム的理解に向けたバイオインフォマティクス·序ーパスウェイ情報解析．実験医学．20(13), 1852-1853, 2002.

高木利久(監修), 大藤道衛 , 高井貴子(編集)．これからのバイオインフォマティックスのためのバイオ実験入門．羊土社．2002.

高井貴子、高木利久．Gene Ontology を遺伝子の機能アノテーションに活用する、蛋白質　核酸　酵素．48，1, 2003.

中井謙太．15章後半　Q&A　アルゴリズムって何？隠れマルコフモデルって何？菅原秀明編　あなたにも役立つバイオインフォマティクス．共立出版 , 118-123, 2002.

小池麻子、中井謙太．5 章　構造予測と機能予測．金久実編　ゲノムネットのデータベース利用法（第3版）．共立出版．82-101, 2002.

中尾光輝、中井謙太．PSORT, 中村保一·磯合敦·石川淳編　バイオデータベースとウェブツールの手とり足とり活用法．羊土社．66-73 , 2002.

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 2002.

1. 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 200,000 clones this year. Of these clones, about 80% of them contained already known genes. About 50% of the known clones seemed to be full. With Helix Institute, we also sequenced about 1,300,000 clones. 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 7000 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 5kb. 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.

2. Identification of differentially expressed genes in metastatic site

Junichi Imai, Manabu Watanabe and Sumio Sugano

We have analyzing differetially expressed genes in lung metastatic model using differential display method. Metastasis of a primary tumor is a multistage process, and the interactions of tumor cells with host stromal cells must influence this process. These interactions may regulate the changes of the multiple gene expression in both tumor cells and host stromal cells at the metastasized site. In the course of characterizing these changes, we have identified overexpression of the c-met proto-oncogene at the metastasized lung by using the mRNA differential display technique. Immunohistochemical staining analysis showed that Met protein was overexpressed in tumor cells at the metastasized site. The c-met encodes a transmembrane tyrosine kinase identified as the receptor for hepatocyte growth factor/scatter factor (HGF/SF). HGF/SF was expressed at lung tissue. The Met was phosphorylated at the metastasized lung. Moreover, the overexpression of c-met was a induction process of transcriptional level, not a selection process. Finally, the c-met was also overexpressed at the metastasized lung by injection of both MC-1 fibrosarcoma cells and B16 melanoma cells. These findings suggest that the HGF/SF-Met signaling may be involved in metastasis.

3. Functional analysis of proteins identified by full-length cDNA clones

Takushi Togashi, Masaaki Oyama, Yoshihiro Omori, Munetomo Hida, Yutaka Suzuki, Sumio Sugano

Function of new genes identified by full-length cDNA clones were first analyzed by sequence homology. Many cDNA clones showed some degree of homology with previously known genes. Homolog search revealed that there was significant number of cDNAs which showed similarity to transcription factors. The expression analysis showed that some of them were expressed in the tissue specific fashion. These tissue specific transcription factors will be very interesting targets for the understanding of development and the function of tissues.

In order to facilitate the functional analysis of the proteins, we are now developing a mass expression capacity of the proteins from cDNA. We are also developing the “proteomics” capacity for the high through-put protein identification and interaction analysis.

4. 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.

Publications

Kato H, Tjernberg A, Zhang W, Krutchinsky AN, An W, Takeuchi T, Ohtsuki Y, Sugano S, Chait BT, Roeder RG. SYT associates with human SNF/SWI complexes and the C-terminal region of its fusion partner SSX1 targets histones. J Biol Chem. 277: 5498-5505, 2002.

Watanabe J, Sasaki M, Suzuki Y, Sugano S. Analysis of transcriptomes of human malaria parasite Plasmodium falciparum using full-length enriched library: identification of novel genes and diverse transcription start sites of messenger RNAs. Gene. 2002 291:105-113, 2002.

Shiroki K, Ohsawa C, Sugi N, Wakiyama M, Miura K, Watanabe M, Suzuki Y, Sugano S. Internal ribosome entry site-mediated translation of Smad5 *in vivo*: requirement for a nuclear event. Nucleic Acids Res. 30: 2851-2861 2002.

Sugiyama T, Ishii S, Yamamoto J, Irie R, Saito K, Otuki T, Wakamatsu A, Suzuki Y, Hio Y, Ota T, Nishikawa T, Sugano S, Masuho Y, Isogai T. cDNA macroarray analysis of gene expression in synoviocytes stimulated with TNFalpha. FEBS Lett. 517: 121-128, 2002.

Omori Y, Imai J, Suzuki Y, Watanabe S, Tanigami A, Sugano S. OASIS is a transcriptional activator of CREB/ATF family with a transmembrane domain. Biochem Biophys Res Commun. 293: 470-477, 2002.

Osada N, Kusuda J, Hirata M, Tanuma R, Hida M, Sugano S, Hirai M, Hashimoto K. Search for genes positively selected during primate evolution by 5'-end-sequence screening of cynomolgus monkey cDNAs. Genomics. 2002 79: 657-662, 2002.

Azuma T, Takei M, Yoshikawa T, Nagasugi Y, Kato M, Otsuka M, Shiraiwa H, Sugano S, Mitamura K, Sawada S, Masuho Y, Seki N. Identification of candidate genes for Sjogren’s syndrome using MRL/lpr mouse model of Sjogren’s syndrome and cDNA microarray analysis. Immunol Lett. 81: 171-176, 2002.

Nishikawa T, Ota T, Kawai Y, Ishii S, Saito K, Yamamoto JI, Wakamatsu A, Ozawa M, Suzuki Y, Sugano S, Isocal T. Comparison of sequences of cDNA clones obtained from oligo-capping cDNA libraries with those from unigene. DNA Res. 8: 255-262, 2001.

Nishikawa T, Ota T, Kawai Y, Ishii S, Saito K, Yamamoto J, Wakamatsu A, Ozawa M, Suzuki Y, Sugano S, Isogai T. Database and analysis system for cDNA clones obtained from full-length enriched cDNA libraries. In Silico Biol. 2: 5-18, 2002.

Osada N, Hida M, Kusuda J, Tanuma R, Hirata M, Hirai M, Terao K, Suzuki Y, Sugano S, Hashimoto K. Prediction of unidentified human genes on the basis of sequence similarity to novel cDNAs from cynomolgus monkey brain. Genome Biol. 3: RE-

SEARCH0006, 2002.

Suzuki Y, Yamashita R, Nakai K, Sugano S. DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. Nucleic Acids Res. 30: 328-331, 2002.

Ueda R. H, Chen W, Adachi A, Wakamatsu H, Hayashi S, Takasugi T, Nagano M, Nakahama K, Suzuki Y, Sugano S, Iino M, Shigeyoshi Y, Hashimoto S. A Transcription Factor Response Element for Gene Expression During Circadian Night. Nature 418: 534-539, 2002.

Takamatsu K, Maekawa K, Togashi T, Choi DK, Suzuki Y, Taylor TD, Toyoda A, Sugano S, Fujiyama A, Hattori M, Sakaki Y, Takeda T. Identification of two novel primate-specific genes in DSCR. DNA Res. 2002 9: 89-97.

Watanabe J, Sasaki M, Suzuki Y, Sugano S. Analysis of transcriptomes of human malaria parasite Plasmodium falciparum using full-length enriched library: identification of novel genes and diverse transcription start sites of messenger RNAs. Gene. 2002 291:105-113.

Shiroki K, Ohsawa C, Sugi N, Wakiyama M, Miura K, Watanabe M, Suzuki Y, Sugano S. Internal ribosome entry site-mediated translation of Smad5 in vivo: requirement for a nuclear event. Nucleic Acids Res. 2002 30: 2851-2861.

Human Genome Center

Laboratory of DNA Information Analysis

DNA 情報解析分野

Professor	Satoru Miyano, Ph.D.	教　授	理学博士	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

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. Computational Strategies for Gene Network Analysis and Gene Expression Profile Analysis

a. Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network

Seiya Imoto, Sunyong Kim, Takao Goto, Sachiyo Aburatani[1], Kousuke Tashiro[1], Satoru Kuhara[1], Satoru Miyano: [1]Graduate School of Genetic Resources Technology, Kyushu University

We propose a new statistical method for constructing a genetic network from microarray gene expression data by using a Bayesian network. An essential point of Bayesian network construction is in the estimation of the conditional distribution of each random variable. We consider fitting nonparametric regression models with heterogeneous error variances to the microarray gene expression data to capture the nonlinear structures between genes. A problem still remains to be solved in selecting an optimal graph, which gives the best representation of the system among genes. We theoretically derive a new graph selection criterion from Bayes approach in general situations. The proposed method includes previous methods based on Bayesian networks. We demonstrate the effectiveness of the proposed method through the analysis of *Saccharomyces cerevisiae* gene expression data newly obtained by disrupting 100 genes.

b. Inferring gene regulatory networks from time-ordered gene expression data of Bacillus subtilis using differential equations

Michiel J.L. de Hoon, Seiya Imoto, Kazuko Kobayashi[2], Naotake Ogasawara[2], Satoru Miyao: [2]Graduate School of Biological Science, Nara Institute of Science and Technology

We describe a new method to infer a gene regulatory network, in terms of a linear system of differential equations, from time course gene expression data. As biologically the gene regulatory network is known to be sparse, we expect most coefficients in such a linear system of differential equations to be zero. In previously proposed methods, the number of nonzero coefficients in the system was limited based on ad hoc assumptions. Instead, we propose to infer the degree of sparseness of the gene regulatory network from the data, where we use Akaike's Information Criterion to determine which coefficients are nonzero. We apply our method to MMGE time course data of *Bacillus subtilis*.

c. Statistical analysis of a small set of time-ordered gene expression data using linear splines

Michiel de Hoon, Seiya Imoto, Satoru Miyano

Recently, the temporal response of genes to changes in their environment has been investigated using cDNA microarray technology by measuring the gene expression levels at a small number of time points. Conventional techniques for time series analysis are not suitable for such a short series of time-ordered data. The analysis of gene expression data has therefore usually been limited to a fold-change analysis, instead of a systematic statistical approach. We use the maximum likelihood method together with Akaike's Information Criterion to fit linear splines to a small set of time-ordered gene expression data in order to infer statistically meaningful information from the measurements. The significance of measured gene expression data is assessed using Student's t-test. Previous gene expression measurements of the cyanobacterium *Synechocystis* sp. PCC6803 were reanalysed using linear splines. The temporal response was identified of many genes that had been missed by a fold-change analysis. Based on our statistical analysis, we found that about four gene expression measurements or more are needed at each time point.

A Dynamic-Link Library (DLL) containing the Fortran routines to fit a linear spline function to data is available at http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/, together with an example Excel spreadsheet calling the routine using Visual Basic. This software package (with patent pending) is free of charge for academic use only.

d. A visualization tool for gene network discovery - G.NET

Ken Aoshima[5], Masayuki Ikawa[5], Satoshi Tanaka[5], Koji Yanagisawa[5], SunYong Kim, Naoki Nariai, Seiya Imoto, Satoru Miyano: [5]Mitsui Knowledge Industry Co., Ltd.

For solving the whole aspect of gene regulation mechanism, the analysis of a gene network attracts considerable attention in the field of molecular biology and bioinformatics. Various methodologies have been developed for inferring a gene network from cDNA microarray gene expression data. However, after constructing a gene network, there are still some problems to be solved in how to extract valuable information from such large-scale network, for example, finding the complex interactions among genes, the evaluation of the estimated gene pathways and so on. For a solution of these problems, we have developed a computer software, named G.NET, for visualizing and analyzing the large-scale gene network. We have also developed the gene network layout algorithms, named GNL algorithm, in order to display the large-scale gene network in two and three dimensional spaces effectively.

2. Knowledge Discovery Systems

a. A string pattern regression algorithm and its application to pattern discovery in long introns

Hideo Bannai, Shunsuke Inenaga[3], Ayumi Shinohara[3], Masayuki Takeda[3], Satoru Miyano: [3]Department of Informatics, Kyushu University

We present a new approach to pattern discovery called string pattern regression, where we are given a data set that consists of a string attribute and an objective numerical attribute. The problem is to find the best string pattern that divides the data set in such a way that the distribution of the numerical attribute values of the set for which the pattern matches the string attribute, is most distinct, with respect to some appropriate measure, from the distribution of the numerical attribute values of the set for which the pattern does not match the string attribute. By solving this problem, we are able to discover, at the same time, a subset of the data whose objective numerical attributes are significantly different from rest of the data, as well as the splitting rule in the form of a string pattern that is conserved in the subset. Although the problem can be solved in linear time for the substring pattern class, the problem is NP-hard in the general case (i.e. more complex patterns), and we present an exact but efficient branch-and-bound algorithm which is applicable to various pattern classes. We apply our algorithm to intron sequences of human, mouse, fly, and zebrafish, and show the practicality of our approach and algorithm. We also discuss possible extensions of our algorithm, as well as promising applications, such as microarray gene expression data.

b. Toward drawing an atlas of hypothesis classes: approximating a hypothesis via another hypothesis model

Osamu Maruyama[4], Takayoshi Shoudai[3], Satoru Miyano: [4]Faculty of Mathematics, Kyushu University

Computational knowledge discovery can be considered to be a complicated human activity concerned with searching for something new from data with computer systems. The optimization of the entire process of computational knowledge discovery is a big challenge in computer science. If we had an atlas of hypothesis classes which describes prior and basic knowledge on relative relationship between the hypothesis classes, it would be helpful in selecting hypothesis classes to be searched in discovery processes. In this paper, to give a foundation for an atlas of various classes of hypotheses, we have defined a measure of approximation of a hypothesis class C_1 to another class C_2. The hypotheses we consider here are restricted to m-ary Boolean functions. For $0 \leq \varepsilon \leq 1$, we say that C_1 is (1-ε)-approximated to C_2 if, for every distribution D over $\{0,1\}^m$, and for each hypothesis h1 in C_1, there exists a hypothesis h_2 in C_2 such that, with the probability at most ε, we have $h_1(x) \neq h_2(x)$ where x in $\{0,1\}^m$ is drawn randomly and independently according to D. Thus, we can use the approximation ratio of C_1 to C_2 as an index of how similar C_1 is to C_2. We discuss lower bounds of the approximation ratios among representative classes of hypotheses like decision lists, decision trees, linear discriminant functions and so on. This prior knowledge would come in useful when selecting hypothesis classes in the initial stage and the sequential stages involved in the entire discovery process.

c. On the complexity of deriving position specific score matrices from examples

Tatsuya Akutsu[6], Hideo Bannai, Satoru Miyano, Sascha Ott: [6]Bioinformatics Center, Institute for Chemical Research, Kyoto University

PSSMs (Position-Specific Score Matrices) have been applied to various problems in Bioinformatics. We study the following problem: given positive examples (sequences) and negative examples (sequences), find a PSSM which correctly discriminates between positive and negative examples. We prove that this problem is solved in polynomial time if the size of a PSSM is bounded by a constant. On the other hand, we prove that this problem is NP-hard if the size is not bounded. We also prove similar results on deriving a mixture of PSSMs.

d. Intrasplicing - analysis of long intron sequences

Sascha Ott, Yoshinori Tamada[8], Hideo Bannai, Kenta Nakai, Satoru Miyano: [8]Department of Mathematical Sciences, Tokai University

We propose a new model for the splicing of long introns, which we call intrasplicing. The basic idea of this model is that the splicing of long introns may be facilitated by the splicing of inner parts of the intron prior to the splicing of the long intron itself. Since long introns have up to about 100,000 bases, this model seems to be a likely explanation of their splicing. To investigate the possibility of this model, we develop a new computational method for the analysis of DNA sequences with respect to splicing. We analyze the genomic sequence of four species with our method and derive several results indicating that intrasplicing may be an appropriate model for the splicing of at least part of the long intron sequences.

3. Biopathway Simulations and Systems Biology

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

Masao Nagasaki, Atsushi Doi, Hiroshi Matsuno[7], Satoru Miyano: [7]Faculty of Science, Yamaguchi University

One of the key issues for exploring systems biology is development of computational tools and capabilities which enable us to understand complex biological systems. In particular, a platform is strongly expected with which biological scientists (users) can comfortably model and simulate dynamic causal interactions and processes in the cell such as gene regulations, metabolic pathways, and signal transduction cascades. Genomic Object Net (GON) was developed aiming at this important mission in systems biology.

There have been pioneering attempts and accumulations of knowledge for this direction, e.g. Gepasi, E-Cell, BioSPICE, etc. for simulation tools, and KEGG, BioCyc, etc. for biopathway databases. In appreciation of these efforts, the architecture of GON was designed so that users can get involved with modeling and simulation biologically intuitively with their profound knowledge and insights and can also be benefited from such biopathway databases.

For modeling a biopathway in a mathematical way, GON employs the notion of hybrid functional Petri net with extension (HFPNe). HFPNe was defined for GON by enhancing some functions to hybrid Petri net, functional Petri net, and hybrid object net so that various aspects in biopathways can be intuitively modeled. Informally speaking, a Petri net is a network of places and transitions connected with arcs. In biopathway modeling with GON, a biological object in a biopathway, e.g. protein, mRNA, functional protein, UV, etc., is represented with a place and the content of the place shows its amount measure. A reaction or interaction among objects such as an enzyme reaction is represented with a transition together with arcs and it defines how the contents of places are consumed or produced. HFPNe can define a hybrid system of continuous and discrete events together with hierarchization of objects for intuitive creation of complex objects. Furthermore, HFPNe allows more "types" for places (integer, real, boolean, string, vector) with which complex information such as localization, etc., can be handled.

We defined XML tags for HFPNe and GON has a graphical drawing editor for HFPNe with which a biopathway model can be simulated. Different from conventional Petri net editors/simulators, GON has functions to include any pictures to represent the basic elements (place, transition, arc) of HFPNe and its background picture for supplement biological meanings, and to attach any text comments and URL links. This helps us understand biopathway models more intuitively. GON also provides a core biopathway library which consists of biopathway components frequently used in modeling. With this library, users can make a model without paying unnecessary attentions to the details of HFPNe.

With this platform, we have succeeded in creating models for gene switch mechanism of λ phage, glycolytic pathway of *E. coli*, boundary formation by notch signaling in *Drosophila*, circadian rhythms in {Drosophila}, and apoptosis induced by Fas ligand, etc.

In the model editor of GON, simulation can be only viewed as a 2D time-course graph. Enhancing visualization, GON equips a tool called visualizer. By writing an XML file for visualization, users can evaluate and tune the model by realizing a personalized visualization of simulation.

Communication between the model editor and the visualizer is established with CORBA that is one of distributed object technologies. Thus, simulation and visualization can be performed on multiple computers. This allows a large scale simulation and visualization of biopathways. For example, users can make visualized comparative analysis of system behaviors among biopathways, such as a wild-type biopathway and mutant ones.

KEGG and BioCyc compile a large number of static biopathway models. The benefits from the HFPNe architecture for biopathway modeling and the flexible features in the model editor of GON have opened a way to recreate dynamic models from these databases. We have developed a tool which transforms biopathway models in KEGG and BioCyc to the GON XML files that can be re-modeled and simulated with GON. This tool can also be extended to cope with another biopathway databases. The software and all XML files of biopathway models are available from http://genomicobject.net/.

b. Boundary formation by Notch signaling in *Drosophila* multicellular systems: experimental observations and gene network modeling by Genomic Object Net

Hiroshi Matsuno[7], Ryutaro Murakami[7], Rie Yamane[7], Naoyuki Yamasaki[7], Sachie Fujita[7], Haruka Yoshimori[7], Satoru Miyano

The Delta-Notch signaling system plays an essential role in various morphogenetic systems of multicellular animal development. Here we analyzed the mechanism of Notch-dependent boundary formation in the Drosophila large intestine, by experimental manipulation of Delta expression and computational modeling and simulation by Genomic Object Net. Boundary formation representing the situation in normal large intestine was shown by the simulation. By manipulating Delta expression in the large intestine, a few types of disorder in boundary cell differentiation were observed, and similar abnormal patterns were generated by the simulation. Simulation results suggest that parameter values representing the strength of cell-autonomous suppression of Notch signaling by Delta are essential for generating two different modes of patterning: lateral inhibition and boundary formation, which could explain how a common gene regulatory network results in two different patterning modes in vivo. Genomic Object Net proved to be a useful and flexible biosimulation system that is suitable for analyzing complex biological phenomena such as patternings of multicellular systems as well as intracellular changes in cell states including metabolic activities, gene regulation, and enzyme reactions.

Publications

Akiyama, F., Tanaka, T., Yamada, R., Ohnishi, Y., Tsunoda, T., Maeda, S., Takei, T., Obara, W., Ito, K., Honda, K., Uchida, K., Tsuchiya, K., Nitta, K., Yumura, W., Nihei, H., Ujiie, T., Nagane, Y., Miyano, S., Suzuki, Y., Fujioka, T., Narita, I., Gejyo, F., Nakamura, Y. Single-nucleotide polymorphisms in the class II region of the major histocompatibility complex in Japanese patients with immunoglobulin A nephropathy. J. Hum. Genet. 47(10):532-538, 2002.

Akutsu, T., Bannai, H., Miyano, S., Ott, S. On the complexity of deriving position specific score matrices from examples. Proc. 13th Annual Symposium on Combinatorial Pattern Matching (CPM 2002), Lecture Notes in Computer Science, 2373:168-177, 2002.

Akutsu, T., Miyano, S. Selecting informative genes for cancer classification using gene expression data. Computational and Statistical Approaches to Genomics (W Zhang & I Shmulevich eds.), Kluwer Academic Pub., Boston, 79-92, 2002.

Akutsu, T., Ott, S. Inferring a union of halfspaces from examples. Proc. 8th Annual International Conference Computing and Combinatorics (COCOON 2002), Lecture Notes in Computer Science, 2387:117-126, 2002.

Aoshima, K., Ikawa, M., Tanaka, S., Yanagisawa, K., Kim, S., Nariai, N., Imoto, S., Miyano, S. A visualization tool for gene network discovery - G.NET. Genome Informatics, 13:445-446, 2002.

Bannai, H. Inenaga, S., Shinohara, A., Takeda, M., Miyano, S. A string pattern regression algorithm and its application to pattern discovery in long introns. Genome Informatics, 13:3-11, 2002.

Bannai, H., Tamada, Y., Maruyama, O., Nakai, K., Miyano, S. Extensive feature detection of N-terminal protein sorting signals. Bioinformatics, 18(2):298-305, 2002.

Doi, A., Matsuno, H., Matsui, M., Hirata, Y., Miyano, S. Simulation of biological systems by hybrid Petri net with an enhancement. Proc. International Conference on Fundamentals of Electronics, Communications and Computer Science 2002, S5, 13-18, 2002.

De Hoon, M., Imoto, S., Miyano, S. Open source clustering software. Genome Informatics, 13:250-251, 2002.

De Hoon, M., Imoto, S., Miyano, S. Inferring gene regulatory networks from time-ordered gene expression data using differential equations. Proc. Fifth International Conference on Discovery Science, Lecture Notes in Artificial Intelligence, 2534:267-274, 2002.

De Hoon, M.J.L., Imoto, S., Miyano, S. Statistical analysis of a small set of time-ordered gene expression data using linear splines. Bioinformatics, 18:1477-1485, 2002.

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, in press.

De Hoon, M.J.L., Lee, E.P., Barnard, J.J., Friedman, A. Cold phase fluid model of the longitudinal dynamics of space-charge dominated beams. Physics of Plasmas, in press.

Imoto, S., Goto, T., Miyano, S. Estimation of genetic networks and functional structures between genes by using Bayesian network and nonparametric regression. Pacific Symposium on Biocomputing. 7:175-186, 2002.

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. Proc. IEEE Computer Society Bioinformatics Conference, 219-227, 2002.

Imoto, S., Kim, S., Shimodaira, H., Aburatani, S., Tashiro, K., Kuhara, S., Miyano, S. Bootstrap analysis of gene networks based on Bayesian netowrks and nonparamatric regression. Genome Informatics, 13:369-370, 2002.

Inenaga, S., Bannai, H., Shinohara, A., Takeda, M., Arikawa, S. Discovering best variable-length-don't-care patterns. Proc. 5th International Conference on Discovery Science (DS2002), Lecture Notes in Computer Science, 2534:86-97, 2002.

Inenaga, S., Shinohara, A., Takeda, M., Bannai, H., Arikawa, S. Space-economical construction of index structures for all suffixes of a string. Proc. 27th International Symposium on Mathematical Foundations of Computer Science (MFCS2002), Lecture Notes in Computer Science, 2420:341-352, 2002.

Kim, S., Imoto, S., Miyano, S. Dynamic Bayesian network and nonparametric regression model for inferring gene networks. Genome Informatics, 13:371-372, 2002.

Kim, S., Imoto, S., Miyano, S. Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data. Proc. Computational Methods in Systems Biology, Lecture Note in Computer Science, Springer-Verlag. in press.

Lathrop, R., Nakai, K., Miyano, S., Takagi, T., Kanehisa, M. (eds.). Genome Informatics 2002, Universal Academy Press, Tokyo, 2002.

Matsuno, H., Hirata, Y., Doi, A., Miyano, S. Genomic Object Net: on-going report on biopathway modeling and simulation. Currents in Computational Molecular Biology, 132-133, 2002.

Matsuno, H., Murakami, R., Yamane, R., Yamasaki, N., Fujita, S., Yoshimori, H., Miyano, S. Experimental observations and simulations by Genomic Object Net of Notch signaling in Drosophila multicellular systems. Genome Informatics, 13:453-454, 2002.

Maruyama, O., Bannai, H., Tamada, Y., Kuhara, S., Miyano, S. Fast algorithm for extracting multiple unordered short motifs using bit operations. Information Sciences, 146(1-4):115-126, 2002.

Maruyama, O., Shoudai, T., Miyano, S. Toward drawing an atlas of hypothesis classes: approximating a hypothesis via another hypothesis model. Proc. 5th International Conference on Discovery Science (DS2002), Lecture Notes in Computer Science, 2534:220-232, 2002.

Matsuno, H., Murakani, 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, inpress.

Miyano, S. Inference, Modeling and Simulation of Gene Networks. Proc. International workshop on Computational Methods in Systems Biology, Lecture Notes in Computer Science, in press.

Nagasaki, M., Doi, A., Sasaki, M., Savoie, C.J., Matsuno, M., Miyano, S. Genomic Object Net in JAVA: a platform for biopathway modeling and simulation, Genome Informatics, 13:252-253, 2002.

Nagasaki, M., Doi, A., Matsuno, H., Miyano, S. Recreating biopathway databases towards simulation. Proc. International workshop on Computational Methods in Systems Biology, Lecture Notes in Computer Science, in press.

Nakano, M., Kitakaze, H., Matsuno, H., Miyano, S. XML pathway file conversion between Genomic Object Net and SBML. Genome Informatics, 13:457-458, 2002.

Ott, S., Tamada, Y., Bannai, H., Nakai, K., Miyano, S. Intrasplicing - analysis of long intron sequences. Pacific Symposium on Biocomputing, 8, in press.

Sumii, E., Bannai, H. VM lambda: a functional calculusfor scientific discovery. Proc. 6th International Symposium on Functional and Logic Programming (FLOPS 2002), Lecture Notes in Computer Science, 2441:290-304, 2002.

Takei, T., Iida, A., Nitta, K., Tanaka, T., Ohnishi, Y., Yamada, R., Maeda, S., Tsunoda, T., Takeoka, S., Ito, K., Honda, K., Uchida, K., Tsuchiya, K., Suzuki, Y., Fujioka, T., Ujiie, T., Nagane, Y., Miyano, S., Narita, I., Gejyo, F., Nihei, H., Nakamura, Y. Association between single-nucleotide polymorphisms in selectin genes and immunoglobulin A nephropathy. Am. J. Hum. Genet., 70(3):781-786, 2002.

Tamada, Y., Bannai, H., Maruyama, O., Miyano, S. Foundations of designing computational knowledge discovery processes. Progress in Discovery Science, Lecture Notes in Computer Science, 2281:459-470, 2002.

Tamada, Y., Ott, S., Bannai, H., Kim, S., Nakai, K, Miyano, S. Analysis of aberrant splicing data. Genome Informatics, 13:424-425, 2002.

Tanaka, T., Kitakaze, H., Matsuno, H., Miyano, S. Development of Genomic Object Net builder for supporting XML design for visualization. Genome Informatics, 13:455-456, 2002.

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、in press.

松野浩嗣、宮野悟．バイオシミュレーションツールGenomic Object Net ～生命システムをわかりやすくモデル化・視覚化できる～．実験医学、20(13):1873-1878, 2002.

松野浩嗣、宮野悟．パスウェイをデジタル化する．蛋白質　核酸　酵素、47(15):2062-2070, 2002.

宮野悟、Christopher J. Savoie．バイオインフォマティクスの創薬応用．実験医学、20(18):2632-2637,2002.

宮野悟．ゲノム情報学とシステムバイオロジー．分子生物学イラストレイティッド（羊土社）

Human Genome Center

Laboratory of Molecular Medicine
Laboratory of Genome Technology

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.
Associate Professor	Hirofumi Arakawa, M.D., Ph.D.
Associate Professor	Youichi Furukawa, M.D., Ph.D.
Research Associate	Toyomasa Katagiri, Ph.D.
Research Associate	Ryuji Hamamoto, Ph.D.
Research Associate	Yataro Daigo, M.D.

教授	医学博士	中村祐輔
助教授	医学博士	荒川博文
助教授	医学博士	古川洋一
助手	医学博士	片桐豊雅
助手	理学博士	浜本隆二
助手	医学博士	醍醐弥太郎

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, bone 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 interesting genes.

1. Genes playing significant roles in human cancer

a. Genes that are inducible by p53

Hirofumi Arakawa, Koichi Matsuda, Kyong-Ah Yoon, Kouji Yoshida, Takashi Kimura, Ching C. Ng, Megumi Iiizumi, Chizu Tanikawa, Yoshio Anazawa, Yasuyuki Nakamura, Hiroshi Nakanishi, Koji Ueda, Meng-Luy Lin, Park Woong Ryeon and Yusuke Nakamura

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. We examined the role of this p53R2-dependent pathway for DNA synthesis in a p53-regulated cell-cycle checkpoint, comparing it to R2-dependent DNA synthesis. The elevation of DNA synthesis activity through RR in response to gamma irradiation was closely correlated with the level of expression of p53R2, but not of R2. The p53R2 product accumulated in nuclei, while R2 levels in cytoplasm decreased. We found a point mutation of p53R2 in cancer-cell line HCT116, which resulted in loss of RR activity. In those cells, DNA damage-inducible apoptotic cell death was enhanced through transcriptional activation of *p53AIP1*. The results suggest that p53R2-dependent DNA synthesis plays a pivotal role in cell survival by repairing damaged

DNA in the nucleus, and that dysfunction of this pathway might result in activation of p53-dependent apoptosis to eliminate dangerous cells.

Cyclin K, a newly recognized member of the "transcription" cyclin family, may play a dual role by regulating CDK and transcription. Using cDNA microarray technology, we found that *cyclin K* mRNA was dramatically increased in U373MG, a glioblastoma cell line deficient in wild-type p53, in the presence of exogenous p53. An electrophoretic mobility-shift assay (EMSA) showed that a potential p53-binding site (p53BS) in intron 1 of the cyclin K gene could indeed bind to p53 protein. Moreover, a heterologous reporter assay revealed that the p53BS possessed p53-dependent transcriptional activity. Colony-formation assays indicated that over-expression of *cyclin K* suppressed growth of T98G and U373MG cells. The results suggested that cyclin K may play a role in regulating the cell cycle after being targeted for transcription by p53.

Interferon regulatory factors (IRFs) regulate transcription of interferon genes through DNA sequence-specific binding to these targets. Using a differential display method for examining gene expression in p53-defective cells infected with adenovirus containing wild-type p53, we found that expression of interferon regulatory factor 5 (*IRF5*) mRNA was increased in the presence of exogenous p53. An electrophoretic mobility-shift assay showed that a potential p53-binding site (p53BS) detected in exon 2 of the *IRF5* gene could in fact bind to p53 protein. Moreover, a heterologous reporter assay revealed that the p53BS possessed p53-dependent transcriptional activity. Expression of *IRF5* was induced in normal human dermal fibroblast (NHDF4042) cells when DNA was damaged by adriamycin or g-irradiation, in a wild-type p53-dependent manner. These results suggest that *IRF5* is a novel p53-target, and that it might mediate the p53-dependent immune response.

We also 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. p53RDL1 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. Therefore 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.

Furthermore, we characterized one gene, termed CABC1 (chaperone-ABC1 (activity of bc1 complex in *S. pombe*)-like) that encodes a 647-amino-acid peptide with significant sequence similarity to ABC1 (Activity of bc1 complex) in Arabidopsis thaliana and *Schizosaccharomyces pombe*. The *CABC1* product was located in mitochondria, and colony-formation assays with cancer-cell lines indicated its ability to suppress cell growth. Inhibition of CABC1 expression by transfection with antisense oligonucleotide significantly reduced the apoptotic response induced by wild-type p53. These results suggest that CABC1 may play an important role in mediating p53-inducible apoptosis through the mitochondrial pathway.

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.

b. Colorectal cancer

Yoichi Furukawa, Ryuji Hamamoto, Li Meihua, Meiko Takahashi, Takashi Shimokawa, Ryuichiro Yagyu, Suguru Hasegawa, Takeshi Watanabe, Daisuke Yuki, Kazutaka, Obama, Michihiro Sakai, Pittella Fabio, Natini Jinawath and Yusuke Nakamura

To investigate the mechanisms of colorectal carcinogenesis, we searched for genes regulated by adenomatous polyposis coli gene product (APC) and identified a novel gene, termed *HELAD1* (helicase, APC down-regulated 1), which was expressed in fetal tissues, but not in any of the adult tissues examined. A recombinant polypeptide representing the ATPases associated with cellular activities (AAA) domain of the *HELAD1* product showed 3' to 5' helicase activity and exonuclease activity *in vitro*. *HELAD1* was abundantly expressed in 16 of 20 colon cancers examined but hardly detectable in corresponding non-cancerous mucosae. Treatment of colon-cancer cells with antisense oligonucleotides suppressed its expression and induced apoptosis. These data revealed an importance of *HELAD1* in colorectal carcinogenesis and suggest that suppression of *HELAD1* may be a promising therapeutic strategy.

We also identified a novel human gene, termed

APCDD1, which was down-regulated in the cancer cells by exogenous wild-type *APC*; its expression was also reduced in response to transduction of *AXIN1*. Moreover, we documented elevated expression of *APCDD1* in 18 of 27 primary colon-cancer tissues compared with corresponding non-cancerous mucosae. A reporter-gene assay using the 5' flanking region of *APCDD1* indicated that transfection of b-catenin together with wild-type Tcf4 into HeLa cells increased the reporter activity through two putative Tcf/LEF-binding motifs upstream of the transcription start site, indicating that *APCDD1* is one of the direct targets of this transcription complex. Exogenous APCDD1 promoted growth of colon-cancer cells both *in vitro* and *in vivo*, while transfection with antisense S-oligodeoxynucleotides decreased cell/tumor growth. These data suggest that *APCDD1* is directly regulated by the b-catenin/Tcf complex and that its elevated expression is likely to contribute to colorectal tumorigenesis.

Analysis by cDNA microarray also indicated that *AF17*, a fusion partner of the *MLL* gene in acute leukemias with t(11;17)(q23;q21), was transactivated according to accumulation of β-catenin. Expression of *AF17* was significantly enhanced in 8 of the 12 colorectal-cancer tissues examined. Introduction of a plasmid designed to express AF17 stimulated growth of NIH3T3 cells, and FACS analysis indicated that the AF17 regulation of cell-cycle progression was occurring mainly at the G2/M transition. Our results suggest that the *AF17* gene product is likely to be involved in the β-catenin-Tcf/LEF signaling pathway and to function as a growth-promoting, oncogenic protein. These findings should aid development of new strategies for diagnosis, treatment and prevention of colon cancers and acute leukemias, by clarifying the pathogenesis of these conditions.

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.

d. cDNA microarray analysis of cancer

Toyomasa Katagiri, Yataro Daigo, Yoichi Furukawa, Hitoshi Zembutsu, Takehumi Kikuchi, Soji Kakiuchi, Toru Nakamura, Koichi Okada, Yasuyuki Kaneta, Satoshi Nagayama, Suguru Hasegawa, Takahide Arimoto, Shingo Ashida, Toshihiro Nishidate, Kensuke Ochi, Chie Suzuki, Nobuhisa Ishikawa, Seiji Tachiiri, Tatsuya Kato, Akira, Togashi, Toshihisa Takagi, and Yusuke Nakamura

i Gastric cancer

To shed light on mechanisms that underlie development and/or progression of intestinal-type gastric cancer, we compared expression profiles of cancer cells obtained by laser-capture microdissection of 20 intestinal-type gastric tumors with expression of genes in corresponding non-cancerous mucosae, by means of a cDNA microarray consisting of 23,040 genes. We identified 62 genes that were commonly up-regulated and 63 that were commonly down-regulated in the cancer tissues. Altered expression of 12 of those genes was associated with lymph-node metastasis. A "predictive score," based on expression profiles of five of the genes that were able to distinguish tumors with metastasis from node-negative tumors in our panel, correctly diagnosed the lymph-node status of nine additional gastric cancers. This genome-wide information contributes to an improved understanding of molecular changes during development of intestinal-type gastric cancers. It may help clinicians predict metastasis to lymph nodes and assist researchers in identifying novel therapeutic targets for this type of cancer.

ii 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.

iii Radiosensitivity

To identify a set of genes related to radiosensitivity of cervical squamous-cell carcinomas and to establish a predictive method, we compared expression profiles of nine radiosensitive and ten radioresistant tumors obtained by biopsy prior to treatment, on a cDNA microarray consisting of 23,040 human genes. We identified 121 genes whose expression was significantly greater in radiosensitive cells than in radio-resistant cells, and 50 genes that showed higher levels of expression in radio-resistant cells than in radiosensitive cells. Some of these genes had already been associated with the radiation response, such as *ALDH1* and *XRCC5* ($P<0.05$, Mann-Whitney test). We selected 62 genes on the basis of a clustering analysis and used expression data from them to establish a predictive scoring (PS) system for discriminating radiation-sensitive from radiation-resistant biopsy samples according to expression profiles of those 62 genes. This PS system successfully and unequivocally discriminated the radiosensitive phenotype from the radioresistant phenotype in our test panel of 19 cervical carcinomas. The extensive list of genes identified in these experiments provides a large body of potentially valuable information for studying the mechanism(s) of radiosensitivity, and our novel PS system opens the possibility of providing appropriate and effective radiotherapy to cancer patients.

iv Chemosensitivity

To explore genes that determine the sensitivity of cancer cells to anticancer drugs, we investigated using cDNA microarrays the expression of 9,216 genes in 39 human cancer cell lines pharmacologically characterized upon treatment with various anticancer drugs. A bioinformatical approach was then exploited to identify genes related to anticancer-drug sensitivity. An integrated database of gene expression and drug sensitivity profiles was constructed and used to identify genes with expression patterns that showed significant correlation to patterns of drug responsiveness. As a result, sets of genes were extracted for each of the 55 anticancer drugs examined. While some genes commonly correlated with various classes of anticancer drug, other genes correlated only with specific drugs with similar mechanisms of action. This latter group of genes may encode molecules that are key determinants in the intrinsic susceptibility of cancer cells to particular drugs. Therefore, the integrated database approach of gene expression and chemosensitivity profiles may be useful in the development of systems to predict anticancer drug susceptibility, as well as be a powerful tool in the discovery of novel targets for cancer chemotherapy.

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 anticancer 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.

To identify genes involved in the sensitivity of acute myeloid leukemia (AML) cells to chemotherapy, we monitored gene-expression profiles of cancer cells from 76 AML patients using a cDNA microarray consisting of 23,040 genes. We identified 63 genes that were commonly over-expressed and 372 genes suppressed in AML. As these genes represent key molecules for disclosing the molecular mechanisms of AML, they may be potential targets for drug development. We also found 28 that revealed different expression levels between good and poor responders to chemotherapy, and appeared to be associated with chemosensitivity. On that basis we developed a "Drug Response Scoring" system that was correlated well with individual sensitivity to an anti-cancer drug regimen. Among the 44 cases with positive drug-response scores by our definition, 40 achieved complete remission after treatment while the only three of the 20 cases with negative scores responded well to the treatment. An ability to predict chemosensitivity should eventually lead to achievement of our goal of "personalized therapy".

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.

v 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.

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 mutations 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 μ-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 (K314T) 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 K314T 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.

b. Osteoporosis and cardiac sudden death

Bone remodeling, *i.e.*, formation and absorption of bone, is under precise regulation; osteoblasts deposit calcified bone matrix and osteoclasts absorb it. Deregulation of this process leads to a variety of metabolic bone diseases, one of them being osteoporosis, a condition in which an increase of bone catabolism over anabolism increases the risk of bone fracture. Osteoporotic fracture is one of the most common and the most serious complications affecting elderly people with the number estimated to increase three-fold by the middle of this century, from 1.7 million in 1990 to 6.3 million by 2050. However, molecular mechanisms involved in its pathogenesis are not largely revealed partly because of its genetic heterogeneity as well as a number of environmental factors in relation to its pathology in human. Hence, a mammalian model will be of great help for better molecular understanding of this disease.We isolated a mammalian gene, whose expression transiently increased in response to intimal denudation of rabbit aorta. It was identical to a gene encoding a zinc transporter, *ZNT5*, reported very recently by others. Mice deficient for this gene showed poor growth and a decrease in bone density due to impairment of osteoblast maturation to osteocyte. More than 60% of male null-mice died suddenly because of the bradyarrhythmias. Analysis of gene-expression profiles in murine hearts by means of an oligonucleotide microarray disclosed that a subset of genes encoding immediate-early response factors (IEGs) and heat shock proteins (HSPs) were down-regulated in *Znt5*-null mice. These results indicate that Znt5 protein plays an important role in maturation of osteoblasts and in maintenance of the cells involved in the cardiac conduction system, partly owing to dysregulated expression of IEGs and HSPs.

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, [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

Although intensive efforts have been undertaken to elucidate the genetic background of IgA nephropathy (IgAN), genetic factors associated with the pathogenesis of this disease are still not well understood. As a first step in investigating a possible relationship between HLA class II genes and IgAN, we analyzed the extent of linkage disequilibrium (LD) in this region of chromosome 6p21.3 in a Japanese test population and found extended LD blocks within the class II locus. We designed a case-control association study of single-nucleotide polymorphisms (SNPs) in each of those LD blocks, and determined that SNPs located in the HLA-DRA gene were significantly associated with an increased risk of IgAN ($P = 0.000001$, odds ratio = 1.91 [95% confidence interval [95% CI] 1.46-2.49]); SNPs in other LD blocks were not. Our data imply that some haplotype of the HLA-DRA locus has an important role in development of IgAN in Japanese patients.

d. Arteriosclerosis

Shuichi Tsukada, Toshihiro Tanaka, Yusuke Nakamura

Vascular restenosis due to intimal thickening remains a major problem after percutaneous transluminal coronary angioplasty (PTCA). Through differential-display analysis we have identified a novel gene whose expression was increased after catheter injury of rabbit aorta. The gene which is expressed predominantly in vascular smooth muscle cells encodes a novel protein with seven transmembrane domains, and we termed it *ITR* (intimal thickness related receptor). The *ITR* sequence contains a motif common to the Rhodopsin-like GPCR (G-protein-coupled receptor) superfamily. *In vivo* analyses of this gene revealed that expression of ITR protein increased with intimal thickening induced by cuff placement around murine femoral artery. Furthermore, *ITR*-knockout mice were found to be resistant to this experimental intimal thickening. *ITR* thus appears to be a novel receptor that may play a role in vascular remodeling and that may represent a good target for development of drugs in the prevention of vascular restenosis.

e. Crohn's Disease

Keiko Yamazaki, Masakazu Takazoe[1], Torao Tanaka[1], Toshiki Ichimori[1], and Yusuke Nakamura: [1]Department of Medicine, Division of Gastroenterology, Social Insurance Chuo General Hospital, Tokyo, Japan

Chronic inflammatory bowel diseases (IBDs), specifically Crohn's disease (CD) and ulcerative colitis (UC), have increased significantly in western countries and Japan over the last decade, but very little is known about their pathogenesis. A candidate-gene approach recently identified *NOD2/CARD15* as one susceptibility gene from the *IBD1* locus on chromosome 16. Alterations in this gene were found in many Caucasian patients with CD; in particular, two nonsynonymous substitutions (R702W and G908R) and a frame-shift mutation (1007fs) were shown to be independent risk factors for CD. We investigated DNA from 483 Japanese CD patients to detect those three mutations in *NOD2/CARD15* by appropriate genotyping techniques, but found only an R702Q substitution in a single patient. Direct sequencing of DNA from 96 of our patients in the regions containing the three reported major mutations detected no sequence alterations of consequence. Our findings indicate that the *NOD2/CARD15* gene is not a major contributor to CD susceptibility in the Japanese population.

Publications

T. Mori, Y. Anazawa, M. Iiizumi, S. Fukuda, Y. Nakamura, and H. Arakawa: Identification of the interferon regulatory factor 5gene (IRF-5) as a direct target for p53. Oncogene 21:2914-2918, 2002.

K. Ochi, T. Mori, Y. Toyama, Y. Nakamura and H. Arakawa: Identification of *semaphorin3B* as a direct target of p53. Neoplasia, 4:82-87, 2002

K. Matsuda, K. Yoshida, Y. Taya, K. Nakamura, Y. Nakamura, H. Arakawa: p53AIP1 regulates the mitochondrial apoptotic pathway. Cancer Res., 62:2883-2889, 2002

M. Iiizumi, H. Arakawa,. T Mori, A. Ando, and Y. Nakamura: Isolation of a novel human p53-target gene encoding a mitochondrial protein, p53ABC1L, that is highly homologous to yeast activity of bc1 complex. Cancer Research, 62:1246-1250, 2002

A. Iida, S. Saito, A. Sekine, K. Kondo, C. Mishima, Y. Kitamura, S. Harigae, S. Osawa, and Y. Nakamura: Thirteen single nucleotide polymorphisms (SNPs) in the alcohol dehydrogenase 4 (ADH4) gene locus. Journal of Human Genetics 47:74-76, 2002

S. Dan, T. Tsunoda, O. Kitahara, R. Yanagawa, H. Zembutsu, T. Katagiri, K. Yamazaki, Y. Nakamura, and T. Yamori: An integrated database of chemosensitivity to 55 anticancer drugs and gene expression profiles of 39 human cancer cell lines. Cancer Research, 62:1139-1147, 2002

A. Iida, S. Saito, A. Sekine, C. Mishima, Y. Kitamura, K. Kondo, S. Harigae, S. Osawa, and Y. Nakamura: Catalog of 77 single nucleotide polymorphisms (SNPs) in the carbohydrate sulfotransferase 1 (CHST1) and carbohydrate sulfotransferase 3 (CHST3) genes. Journal of Human Genetics 47:14-19, 2002

H. Zembutsu, Y. Ohnishi, T. Tsunoda, Y. Furukawa, T. Katagiri, Y. Ueyama, N. Tamaoki, T. Nomura, O. Kitahara, R. Yanagawa, K. Hirata, and Y. Nakamura: Genome-wide cDNA microarray screening to correlate gene-expression profiles with sensitivity of 85 human-cancer xenografts to anticancer drugs. Cancer Research, 62:518-527, 2002

M. Hirakawa, T. Tanaka, Y. Hashimoto, M. Kuroda, T. Takagi, and Y. Nakamura: JSNP: a database of common gene variations in the Japanese population. Nucleic Acid Research, 30:158-162, 2002

S. Saito, A. Iida, A. Sekine, Y. Miura, C. Ogawa, S. Kawauchi, S. Higuchi, and Y. Nakamura: 326 genetic variations in genes encoding nine members of ATP-binding cassette, sub-family B (*ABCB/MDR/TAP*) in the Japanese population. Journal of Human Genetics 47:38-50, 2002

T. Takei, A. Iida, K. Nitta, T. Tanaka, Y. Ohnishi, R. Yamada, S. Maeda, T. Tsunoda, S. Takeoka, K. It o, K. Honda, K. Uchida, K. Tsuchiya, Y. Suzuki, T. Fujioka, T. Ujiie, Y. Nagane, S. Miyano, I. Narita, F. Gejyo, H. Nihei, Y. Nakamura: Association between single-nucleotide polymorphisms in selectin genes and IgA nephropathy. Am. J. Human Genetics, 70:781-786, 2002

M. Doi, M. Nagano and Y. Nakamura: Genome-wide screening by cDNA microarray of genes associated H. Yamanaka, N. Hashimoto, K. Koyama, H. Nakagawa, Y. Nakamura, and K. Noguchi: Expression of Apc2 during mouse development. Gene Expression Patterns 1:107-114, 2002

M. Tsuneizumi, M. Emi, A. Hirano, Y. Utada, K. Tsumagari, K. Takahashi, F. Kasumi, F. Akiyama, G. Sakamoto, T. Kazui and Y. Nakamura: Assocaition of allelic loss at 8p22 with poor prognosis breast cancer cases treated with high-dose adjuvant chemotherapy. Cancer Letters 180:75-82, 2002

M. Tachikawa, Y. Nagai, K .Nakamura, K. Kobayashi, T. Fujiwara, H-J. Han, Y. Nakabayashi, Y. Ichikawa, J. Goto, I. Kanazawa, Y. Nakamura, and T. Toda: Identification of CAG repeat-containing genes expressedin human brain as candidate genes for autosomal dominant spinocerebellar ataxias and other neurodegenerative diseases. J Hum Genet, 47:275-278, 2002.

T. Nagahata, A. Hirano, Y. Utada, S. Tsuchiya, K. Takahashi, T. Tada, M. Makita, F. Kasumi, F. Akiyama, G. Sakamoto, Y. Nakamura, and M. Emi: Correlation of allelic losses and clinicopathological factors in 504 primary breast cancers, Breast Cancer 9:208-215, 2002.

D.G. Duda, M. Sunamura, L. Lozonshi, T. Yokoyama, T. Yatsuoka, A. Horii, K. Tani, S. Asano, Y. Nakamura, and S. Matsuno: Overexpression of the p53-inducible brain angiogenesis inhibitor 1 suppresses efficiently tumour angiogenesis. British Journal of Cancer, 86:490-496, 2002

Y. Sasaki, S. Ishida, I. Morimoto, T. Yamashita, T. Kojima, C. Kihara, T. Tanaka, K. Imai, Y.

Nakamura, and T. Tokino: The p53 family member genes are involved in the notch signal pathway. Journal of Biological Chemistry 277:719-724, 2002

M. Iizaka, Y. Furukawa, T. Tsunoda, H. Akashi, M. Ogawa, and Y. Nakamura: Expression profile analysis of colon cancer cells in response to sulindac or aspirin. B.B.R.C., 292:498-512, 2002

K. Miura, E. D. Bowman, R. Simon, A. C. Peng, A. I. Robles, R. T. Jones, T. Katagiri, P. He, H. Mizukami, L. Charboneau, T. Kikuchi, L. A. Liotta, Y. Nakamura, and C. C. Harris: Laser capture microdissection and microarray expression analysis of lungadenocarcinoma reveals tobacco smoking- and prognosis-related molecular profiles. Cancer Research 62:3244–3250, 2002

J. Kamogawa, M.Terada, S. Mizuki, M, Nishihara, H. Yamamoto, S, Mori, Y. Abe, K. Morimoto, S. Nakatsuru, Y. Nakamura, and M. Nose: Arthritis in MRL/lpr Mice Is Under the Control of Multiple Gene Loci With an Allelic Combination Derived From the Original Inbred Strains. Arthritis & Rheumatism, 46:1067-1074, 2002.

T. Mori, Y. Anazawa, K. Matsui, S. Fukuda, Y. Nakamura, and H. Arakawa: Cyclin K as a direct transcriptional target of the p53 tumor suppressor. Neoplasia, 4:268-274, 2002.

S. Saito, A. Iida, A. Sekine, Y. Miura, C. Ogawa, S. Kawauchi, S. Higuchi, and Y. Nakamura: 779 genetic variations in eight genes encoding members of ATP-binding cassette, subfamily C (ABCC/MRP/CFTR). J. Human Genetics, 47:147-171, 2002.

S. Higuchi, Y. Nakamura and S. Saito: Characterization of a VNTR polymorphism in the coding region of the *CEL* gene. Journal of Human Genetics, 47:213-215, 2002.

Y.-M. Lin, Y. Furukawa, T. Tsunoda, C.-T. Yue, K.-C. Yang, and Y. Nakamura: Molecular diagnosis of colorectal tumors by expression profiles of 50 genes expressed differentially in adenomas and carcinomas. Oncogene, 21:4120-4128, 2002.

H. Iwasa, T. Itoh, R. Nagai, Y. Nakamura, T. Tanaka: Twenty single nucleotide polymorphisms (SNPs) and their allelic frequencies in four genes that are responsible for familial long QT syndrome in the Japanese population. Journal of Human Genetics, 45:182-183, 2002

T. Kayashima, M. Katahira, N. Harada, N. Miwa, T. Ohta, K. Yoshiura, N. Matsumoto, Y. Nakane, Y. Nakamura, T. Kajii, N. Niikawa, and T. Kishino; Maternal Isodisomy for 14q21-24 in a man with daiabetes mellitus. Amercan Journal of Medical Genetics 111:38-42, 2002

M. Nishiu, R. Yanagawa,S. Nakatsuka, M. Yao, T. Tsunoda, Y. Nakamura, and K. Aozasa: Microarray analysis of gene-expression profiles in diffuse large B-cell lymphoma: Identification of genes related to disease progressionÅDJpn. J Cancer Research, 93:894-901, 2002

A. Iida, S. Saito, A. Sekine, C. Mishima, Y. Kitamura, K. Kondo, S. Harigae, S. Osawa, and Y. Nakamura: Catalog of 605 single-nucleotide polymorphisms (SNPs) among 13 genes encoding human ATP-binding cassette transporters: ABCA4, ABCA7, ABCA8, ABCD1, ABCD3, ABCD4, ABCE1, ABCF1, ABCG1, ABCG2, ABCG4, ABCG5, and ABCG8. Journal of Human Genetics, 47:285-310, 2002

O. Kitahara, T. Katagiri, T. Tsunoda, Y. Harima, and Y. Nakamura: Classification of sensitivity or resistance of cervical cancers to ionizing radiation according to expression profiles of 62 genes selected by cDNA microarray analysis. Neoplasia, 4:295-303, 2002.

R. Yagyu, R. Hamamoto, Y. Furukawa, H. Okabe, T. Yamamura, and Y. Nakamura: Isolation and characterization of a novel human gene, *VANGL1*, as a therapeutic target for hepatocellular carcinoma. Int. J Oncology, 20:1173-1178, 2002.

A. Saito-Hisaminato, T. Katagiri, S. Kakiuchi, T. Nakamura, T. Tsunoda, and Y. Nakamura: Genome-wide profiling of gene expression in 29 normal human tissues with a cDNA microarray. DNA Research, 9:35-45, 2002.

F. Saito-Ohara, Y. Fukuda, M. Ito, K.L. Agarwala, M. Hayashi, M. Matsuo, I. Imoto, K. Yamakawa, Y. Nakamura, and J. Inazawa: The Xq22 inversion breakpoint interrupted a novl ras-like GTPase gene in a patient with Duchenne muscular dystrophy and profound mental reterdation. Am. J. Human Genetics 71:637-645, 2002

S. Nagayama, T. Katagiri, T. Tsunoda, T. Hosaka, Y. Nakashima, N. Araki, K. Kusuzaki, T. Nakayama, T. Tsuboyama, T. Nakamura, M. Imamura, Y. Nakamura, and J. Toguchida: Genome-wide analysis of gene expression in synovial sarcomas using a cDNA microarray. Cancer Research, 62:5859-5866, 2002

M. Takahashi, T. Tsunoda, M. Seiki, Y. Nakamura, and Y. Furukawa: Identification of membrane-type matrix metalloproteinase-1 as a target of the _-catenin/Tcf4 complex in human colorectal cancers. Oncogene 21:5861-5867, 2002

S. Saito, A. Iida, A. Sekine, C. Ogawa, S. Kawauchi, S. Higuchi, M. Ohno, and Y. Nakamura: 906 variations among 27 genes encoding cytochrome P450 (*CYP*) enzymes and aldehyde dehydrogenases (*ALDH*) in the Japanese population. Journal of Human Genetics 47:419-444, 2002

K. Inoue, K. Matsuda, M. Itoh, H. Tomoike, T. Aoyagi, R. Nagai, M. Hori, Y. Nakamura, and T. Tanaka: Osteopenia and male-specific sudden cardiac death in mice with deficiency of *COT1*, a novel gene containing a cation-efflux domain. Human Molecular Genetics 11:1775-1784, 2002

Ishiguro, T. Shimokawa, T. Tsunoda, T. Tanaka, Y. Fujii, Y. Nakamura and Y. Furukawa: Isolation of HELAD1, a novel human helicase gene up-regulated in colorectal carcinomas. Oncogene 21:6387-6394, 2002

K. Yamazaki, M. Takazoe, T. Tanaka, T. Ichimori, and Y. Nakamura: Absence of mutation in the NOD2/CARD15 gene among 483 Japanese patients with Crohn's Disease. Journal of Human Genetics 47:469-472, 2002

A. Iida, S. Saito, A. Sekine, C. Mishima, Y. Kitamura, K. Kondo, S. Harigae, S. Osawa, and Y. Nakamura: Catalog of 86 single-nucleotide polymorphisms (SNPs) in three uridine diphosphate glycosyltransferase genes: UGT2A1, UGT2B15, and UGT8. Journal of Human Genetics 47:505-510, 2002

F. Akiyama, T. Tanaka, R. Yamada, Y. Ohnishi, T. Tsunoda, S. Maeda, T. Takei, W. Obara, K. Ito, K. Honda, K. Uchida, K. Tsuchiya, K. Nitta, K. Yumura, H. Nihei, T. Ujiie, Y. Nagane, Y. Suzuki, T. Fujioka, I. Narita, F. Gejyo, and Y. Nakamura: Single-nucleotide polymorphisms in the class II region of the major histocompatibility complex in Japanese patients with immunoglobulin A nephropathy. Journal of Human Genetics 47:532-538 2002

S. Saito, A. Iida, A. Sekine, C. Ogawa, S. Kawauchi, S. Higuchi, and Y. Nakamura: Catalog of 238 variations among six human genes encoding solute carriers (*hSLCs*) in the Japanese population Journal of Human Genetics 47:576-584, 2002

M. Takahashi, M. Fujita, Y. Furukawa, R. Hamamoto, T. Shimokawa, N. Miwa, and Y. Nakamura: Isolation of a novel human gene, *APCDD1*, as a direct target of the β-catenin/T-cell factor 4 complex with probable involvement in colorectal carcinogenesis. Cancer Research 62:5651-5656, 2002

Y. Yamanaka, Y. Hamazaki, Y. Sato, K. Ito, K. Watanabe, T. Heike, T. Nakahata and Y. Nakamura: Maturational sequence of neuroblastoma revealed by molecular analysis on cDNA microarrays. Int. J. Oncology 21:803-807, 2002

Y. Kaneta, Y. Kagami, T. Katagiri, T. Tsunoda, I. Jinnai, H. Taguchi, H. Hirai, K. Ohnishi, T. Ueda, N. Emi, A. Tomida, T. Tsuruo, Y. Nakamura, and R. Ohno: Prediction of sensitivity to STI571 among chronic myeloid leukemia patients by genome-wide cDNA microarray analysis. Jpn J Cancer Research 93:849-856, 2002

H. Haga, R. Yamada, Y. Ohnishi, Y. Nakamura, and T. Tanaka; The summary of the gene-based SNP discovery project in the Japanese Millennium Genome Project; Identification of 190,562 genetic variations in the human genome. Journal of Human Genetics 47:605-610, 2002

M. Horie M, K. Kobayashi, S. Takeda, Y. Nakamura, G. E. Lyons, and T. Toda: Isolation and characterization of the mouse ortholog of the Fukuyama-type congenital musclular dystrophy gene. Genomics 80:482-486, 2002

S. Hasegawa, Y. Furukawa, M. Li, S. Satoh, T. Kato, T. Watanabe, T. Katagiri, T. Tsunoda, Y. Yamaoka, and Y. Nakamura: Genome-wide analysis of gene expression in intestinal-type gastric cancers using a cDNA microarray representing 23,040 genes. Cancer Research, 62: 7012-7017, 2002

J. Okutsu, T. Tsunoda, Y. Kaneta, T. Katagiri, O. Kitahara, H. Zembutsu, R. Yanagawa, S. Miyawaki, K. Kuriyama, N. Kubota, Y. Kimura, K. Kubo, F. Yagasaki, T. Higa, H.Taguchi, T. Tobita, H. Akiyama, A. Takeshita, Y.-H. Wang, T. Motoji, R. Ohno, and Y. Nakamura: Prediction of chemosensitivity for patients with acute myeloid leukemia, according to expression levels of 28 genes selected by genome-wide cDNA microarray analysis. Clinical Cancer Research, in press

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 Research, in press

K. Ozaki, Y. Ohnishi, A. Iida, A. Sekine, R. Yamada, T. Tsunoda, H. Sato, H. Sato, M. Hori, Y. Nakamura, and T. Tanaka: Functional SNPs in the lymphotoxin-_ gene that contribute to susceptibility to myocardial infarction. Nature Genetics, in press

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

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. Oncogene, in press

T. Arimoto, T. Katagiri, K. Oda, T. Tsunoda, T. Yasugi, Y. Osuga, H. Yoshikawa, O. Nishii, T. Yano, Y. Taketani and Yusuke Nakamura: Genome-wide cDNA microarray analysis of gene-expression profiles involved in ovarian endometriosis. Int. J. Oncology. In press

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. J. Human Genetics, in press

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

C. Tanikawa, K. Matsuda, S. Fukuda, Y. Nakamura, and H. Arakawa: *p53RDL1*^{UNC5B} as a positive and negative regulator of p53-dependent apoptosis. Nature Cell Biology in press

A. Iida and Y. Nakamura: High-resolution SNP map in the 55-kb region containing the selectin gene family on chromosome 1q24-q25. J. Human Genetics, in press

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

S. Abe, K. Koyama, S. Usami, Y. Nakamura: Construction and characterization of a vestibular-specific cDNA library using T7-based RNA amplification. J. Human Genetics, in press

A. Iida, T. Tanaka, and Y. Nakamura: High-density SNP map of human *ITR*, a gene associated with vascular remodeling. Journal of Human Genetics, in press

Human Genome Center

Laboratory of Sequence Analysis
シークエンスデータ情報処理分野

Associate Professor Tetsushi Yada, Ph.D.
Research Associate Natsuhiro Ichinose, Ph.D.

助教授　理学博士　矢　田　哲　士
助　手　工学博士　市　瀬　夏　洋

With the rapid growth of sequencing techniques for the genomes of a great number of species, it becomes more important to understand coding principles of biological information in the genome sequences. The main mission of our laboratory is to develop the techniques to extract such coding principles by using the data-mining techniques, the theory of deterministic dynamics and the statistical analysis.

1. DIGIT: a novel gene finding program by combining gene-finders

Tetsushi Yada, Yasushi Totoki, Yoshio Takaeda, Yoshiyuki Sakaki, Toshihisa Takagi

We have developed a general purpose algorithm which finds genes by combining plural existing gene-finders. The algorithm has been implemented into a novel gene-finder named DIGIT. An outline of the algorithm is as follows. First, existing gene-finders are applied to an uncharacterized genomic sequence (input sequence). Next, DIGIT produces all possible exons from the results of gene-finders, and assigns them their exon types, reading frames and exon scores. Finally, DIGIT searches a set of exons whose additive score is maximized under their reading frame constraints. Bayesian procedure and a hidden Markov model are used to infer exon scores and search the exon set, respectively. We have designed DIGIT so as to combine the results of FGENESH, GENSCAN and HMMgene, and have assessed its prediction accuracy by using recently compiled benchmark data sets. For all data sets, DIGIT successfully discarded many false-positive exons predicted by individual gene-finders and yielded remarkable improvements in sensitivity and specificity at the gene level compared with the best gene level accuracies achieved by any single gene-finder.

2. DIGITized Gene: A putative human gene set with abundance of novel genes

Tetsushi Yada, Yasushi Totoki, Takehiko Itoh, Toshihisa Takagi

DIGITized Gene is a data set of putative human genes which abundantly contains novel genes. We have applied our novel gene finding program (gene-finder), called DIGIT, to the human genome sequences and have identified 8,151 genes, DIGITized genes, whose genomic regions do not overlap with known genes. DIGIT finds genes by combining plural ab initio gene-finders. Our previous work clearly showed that DIGIT discarded many false-positive exons predicted by ab initio gene-finders and yielded remarkable improvements in sensitivity and specificity at the gene level compared with the best gene level accuracies achieved by any single gene-finder. We have checked the quality of DIGITized Gene and have successfully accumulated the evidences to show that DIGITized Gene is highly reliable. DIGIT and DIGITized Gene are made available upon request to the authors.

3. A novel index which precisely derives protein coding regions from cross-species genome alignments

Hideki Noguchi[1], Tetsushi Yada, Yoshiyuki Sakaki: [1]RIKEN Genomic Sciences Center

We introduce here a novel index which precisely derives protein coding regions from cross-species genome alignments. The index is deeply related to frame recovery observed in coding sequence alignments, that is, if insertions or deletions of nucleotides causes frame shifts in coding regions, other in-dels which recover the reading frames will be often observed in the vicinity. In contrast, such frame recoveries are not observed in other conserved regions. We prepared two gene models: a model which finds gene by using sequence similarity and intrinsic gene measures (basic model), and the other model which finds gene by using frame recovery index in addition to sequence similarity and intrinsic gene measures (frame recovery model). We evaluated the prediction accuracies of the two models, and our benchmark test revealed that frame recovery model significantly improved the prediction accuracy in comparison with basic model.

4. Quadtree representation of oligonucleotides

Natsuhiro ICHINOSE, Tetsushi YADA and Toshihisa TAKAGI

One of the most fundamental characteristics of DNA sequences is oligonucleotide frequencies. Since we can access large-scale genome sequences today, it becomes more important to analyze such frequency characteristics in genome-wide sequences.

There are two problems to perform such analysis: One is the scale problem, and another is the sensitivity problem. Since the genome sequence has mega-giga-order bases, the algorithm should overcome such large scale. Additionally, the distribution of oligonucleotide frequencies is not uniform generally. This causes to lose the sensitivity to extract the frequency characteristics.

The quadtree representation is a framework to represent oligonucleotide frequencies in the regular square hierarchically. The basic algorithm to analyze oligonucleotide frequencies is to count oligonucleotides. Since the each oligonucleotide is represented as a pair of integer values in the quadtree representation, the count algorithm can become simpler. For example, to count 9-mer oligonucleotides of the human chromosome 21 (which has about 30 Mb) it takes only 10-20 seconds.

In order to solve the sensitivity problem, we introduce the log-odds between an oligonucleotide frequency and a background frequency represented as the N-th order Markov chain. As results, we have shown that the representation of the dinucleotide repeats with mutations has skewness corresponding to the transition-transversion skewness of mutations, which is not observed in the conventional representation. It is interesting that the characteristics of mutations can be observed in a single sequence without comparing with other genomes.

5. A motif extraction method by using Gray code

Natsuhiro ICHINOSE, Tetsushi YADA and Toshihisa TAKAGI

One of causes of difficulties to analyze DNA sequences is that only the distance can be defined between a pair of oligonucleotides generally (ex. Hamming distance). Then the difficulty arises when we attempt to find relations among numerous oligonucleotides such as motif finding. Namely, since we should check all combinations of the oligonucleotides and their distances to identify the motif candidates, the computations increases exponentially with increasing the number of sequences.

If we can assume an order among oligonucleotides with preserving the distance, the problem is more simplified. Namely, we can search for each oligonucleotide in the order locally. In this case, the computations of searching takes only a linear order of the number of the motif candidates.

As one of such ordering methods, we introduce the Gray code. The Gray code is an order of oligonucleotides in which adjacent ones have only a single nucleotide difference (namely, have the Hamming distance 1). Therefore, by searching oligonucleotides in the Gray code order locally, we may find the motif which corresponds to a set of oligonucleotides with small substitutions.

We are now developing the method to extract the motif by using the Gray code. Since the Gray code method has locality for searching as we mentioned above, it is hopeful that it will be developed as a large-scale motif finding system.

6. An expression-based design of model gene networks

Natsuhiro ICHINOSE and Kazuyuki AIHARA[1]: [1]Graduate school of frontier science, the university of Tokyo

Gene networks play an important role of biological functions in human and other species. In order to understand its biological dynamics and apply them to genetic engineering, we study a method to design model gene networks.

When we consider to design gene networks, a direct method is that we determine the gene interaction directly. However, since the gene network system has nonlinearity, then we can not predict its dynamics from that of its partial system always. Therefore we adopt an indirect method in which we determine the gene interaction indirectly from gene expression time series by using an optimization method.

The gene expression time series is essentially continuous in time and its quantity. However, it may be difficult for a designer to present such continuous time series. Therefore we developed the method to determine gene interaction represented by differential equations from binary expression time series.

We have shown that the we can construct oscillatory gene networks from cyclic expression patterns. We also have shown that we can implement two different cyclic patterns to a single gene networks.

7. Identifying local periodic patterns in genomic DNA sequences

Hiroo MURAKAMI, Natsuhiro Ichinose, Tetsushi YADA and Toshihisa TAKAGI

It is known that the local secondary structure of genomic DNA is concerned with transcriptional regulation of genes. Further, sequence periodicity is observed in the corresponding genomic regions. Therefore, it is expected that exhaustive identification of periodic sequences unevenly distributed in genomic DNA brings novel knowledge of the relationship between transcriptional regulation and sequence pattern.

We have developed a computer program named STEPSTONE. STEPSTONE can identify a set of oligonucleotides which are periodically observed in local genomic regions. The periodicity and locality of oligonucleotide sequences are evaluated on the bases of their auto-correlations and CV(coefficient of variation) values, respectively.

We have applied STEPSTONE to genomic sequences of human chromosome 22q and 21. As results, we successfully identified a set of oligonucleotides which are periodically observed in local genomic regions. These oligonucleotides tend to be observed in the vicinity of genes. However, their sequences identified in chromosome 22q are quite different from those identified in chromosome 21. Further, their periodicities are also different. We are now analyzing human chromosome 19, where several local secondary structures of genomic DNA concerned with transcriptional regulation of genes were already identified experimentally.

Publications

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

H. Noguchi, T. Yada, and Y. Sakaki. A novel index which precisely derives protein coding regions from cross-species genome alignments. In Proc. of Genome Informatics Workshop 13, 183-191, 2002.

Y. Watanabe, A. Fujiyama, Y. Ichiba, M. Hattori, T. Yada, Y. Sakaki, and T. Ike-mura. Chromosome-wide assessment of replication timing for human chromosomes 11q and 21q: disease-related genes in timingswitch regions. Human Molecular Genetics, 11, 13-21, 2002.

T. Yada, Y. Totoki, T. Takagi, and K. Nakai. A novel bacterial gene-finding system with improved accuracy in locating start codons. DNA Res., 8, 97-106, 2001.

F. Miura, T. Yada, K. Nakai, Y. Sakaki, and T. Ito. Differential display analysis of mutants for the transcription factor pdr1p regulating multidrug resistance in the budding yeast. FEBS Letters, 505, 103-108, 2001.

International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature, 409:860-921, 2001.

The International Human Genome Mapping Consortium. A physical map of the human genome. Nature, 409:934-941, 2001.

中井 謙太、矢田 哲士．シグナル同定、遺伝子同定技術．生物情報工学の基礎．東京化学同人、印刷中．

阿久津 達也、浅井 潔、矢田 哲士．バイオインフォマティクス：確率モデルによる遺伝子配列解析．医学出版、2001.

N. Ichinose and K. Aihara, A design method of model gene networks, Proceedings of International Symposium on Artificial Life and Robotics (AROB 7th), S1-4, 66-69 (2002)

Human Genome Center

Laboratory of Functional Genomics

ゲノム機能解析分野

Professor	Yoshiyuki Sakaki, Ph.D.
Associate Professor	Hajime Tei, Ph.D.
Research Associate	Yuriko Hagiwara-Takeuti, Ph.D.

教　授	理学博士	榊　　佳　之
助教授	農学博士	程　　　　肇
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 hunting of genes with unique expression patterns, and molecular mechanism regulating mammalian circadian rhythms.

1. Comparative genomics of the human genome

Kunihiko Takamatsu, Kouhei Maekawa, Tomoyo Shirakawa, Kohji Okamura, Tadayuki Takeda1, 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 is reaching the final goal. However, the knowledge obtained from human genome sequence alone (even if completely determined) is limited. One powerful approaches to zoom up important regions. Important regions (sequences) of the genome is 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 rmay eveal "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 MMU16 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% but there exist considerable number of indels in the genomes.

2. A comprehensive analysis of allelic methylation status of CpG islands on human chromsome 21

Yoichi Yamada[2], Tomoyo Shirakawa, Hidemi Watanabe[1], Takashi Ito[2], Yoshiyuki Sakaki: [1]RIKEN, Genomic Sciences Center, Yokohama, [2]Cancer Research Institute, Kanazawa Univ.

Approximately half of the human genes have CpG islands (CGIs) around their promoter regions. While CGIs usually escape methylation, those on chromosome X in females and those in the vicinity of imprinted genes are exceptions: they have both methylated and unmethylated alleles to display a "composite" pattern in methylation analysis. In addition, aberrant methylation of CGIs is known to often occur in cancer cells. Here we developed a simple HpaII-McrBC PCR method for discrimination of full, null, incomplete, and composite methylation patterns, and applied it to all CGIs on human chromosome 21q. This comprehensive analysis revealed that, although most CGIs (103 out of 149) escape methylation, a sizable fraction (31 out of 149) are fully methylated even in normal peripheral blood cells. Furthermore, we identified seven CGIs showing the composite methylation, and demonstrated that three of them are indeed methylated mono-allelically. Further analyses using informative pedigrees revealed that two of the three are subject to maternal allele-specific methylation. Intriguingly, the other CGI is methylated in an allele-specific but parental-origin-independent manner. Thus the cell seems to have a broader repertoire of methylating CGIs than previously thought, and our approach may contribute to uncover novel modes of allelic methylation. We are currently applying this strategy to mouse chromosome 16 and chimpanzee chromosome 22, both corresponding to human chromosome 21, as a comparative epigenomics study.

3. Allelic message display screening and comparative genome analysis for imprinted genes

Yuriko Hagiwara, Kohji Okamura, Aya Nakayama, Takashi Ito[1] and Yoshiyuki Sakaki: [1]Cancer Research Institute, Kanazawa Univ.

Various human diseases are known to have the feature of differential expression of the phenotype, depending on the parent of origin. Such diseases include not only well-defined genetic disorders like Prader-Willi/Angelman syndrome but also unstable triplet repeat diseases and common diseases such as insulin-dependent diabetes mellitus, atopy, bipolar affective disorder and various malignant tumors. Thus the systematic screening for imprinted genes would accelerate the identification of genes involved in these diseases. We developed a unique Allelic Message Display (AMD) screening for imprinted genes and identified a novel paternally expressed gene Impact and some known genes *Snrnp Peg3, Igf2r and Necdin*. In additon to them, 24 maternally expressed genes and 18 paternally expressed genes identified by improved-AMD using nuclear-transplanted mice are now being analyzed.

We also conducted a comparative genome analysis of paternally-expressed gene *Impact* and its non-imprinted human homolog *IMPACT* to reveal a CpG island unique to the mouse gene. Intriguingly, the island is subject to parent-of-origin-dependent methylation, thereby serving as a candidate region to control inprinting. We are trying to develop mice deleted for this differentially methylated CpG island to reveal its role in genomic imprinting. In addition, we identified genes flanking *Impact* and *IMPACT*, and revealed their biallelic expression. Thus, *Impact* is an isolated imprinted gene, which is not embedded in so-called imprinted chromosomal domains.

4. Molecular mechanisms regulating mammalian circadian clock

Hajime Tei, Akiko Hida, Rika Numano, Nobuya Koike, Shihoko Kojima, Yoko Sato, Satomi Shiozuka, Soshi, Kawaguchi, Atsuki Shinozaki, Matsumi Hirose, Miyuki Shimada 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*, a mammalian clock gene, 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* in the SCN is induced immediately by a light pulse even at night. Function of the circadian expression of the mammalian *Per1* gene 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 *Per1* genes in addition to their transcriptional start sites have been determined. Both of the genomic sequences consist 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. Second, We generated a *Per1::luc* transgenic rat line in which luciferase is rhythmically expressed under the control of the mouse *Per1* promoter, and have used it to study mammalian circadian organization. Light emission from cultured suprachiasmatic nuclei (SCN) of these rats was invariably and robustly rhythmic. Circadian rhythm light emission from the SCN followed light cycle shifts 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. Third, we constructed transgenic rat lines with constitutive expression of the mouse *Per1* gene using *Elongation 1 alpha* or *Neural specific enolase* promoters. Both the circadian period of locomoter activity and entrainment to light-dark cycles were severely affected in several transgenic lines. In addition, we measured the expression of the native (rat) *Per1* and *Per2* genes in the SCN and retina of the transgenic lines under DD conditions. The circadian expression of endogenous *Per1* and *Per2* genes was diminished in the transgenic lines. These results clearly indicate that the circadian expression and light induction of the mammalian *Per1* gene is involved in rhythm generation and/or entrainment of the circadian clock.

Publications

Watanabe, Y., Fujiyama, A., Ichiba, Y., Hattori, M., Yada, T., Sakaki, Y. and Ikemura, T. Chromosome-wide assessment of replication timing for human chromosomes 11q and 21q: disease-related genes in timing-switch regions. Human Molecular Genetics 11:13-21, 2002.

Fujiyama, A., Watanabe, H., Toyoda, A., Taylor, T., Itoh, T., Tsai, S-F., Park, H-S., Yaspo, M-L., Lehrach, H., Chen, Z., Fu, G., Saitou, N., Osoegawa, K., de Jong, P., Suto, Y., Hattori, M. and Sakaki, Y. Construction and Analysis of a Human-Chimpanzee Comparative Clone Map. Science 295: 1-220, 2002.

Takamatsu, K., Maekawa, K., Togashi, T., Choi, D-K., Suzuki, Y., Taylor, T., Toyoda, A., Sugano, S., Fujiyama, A., Hattori, M., Sakaki, Y. and Takeda, T. Isentification of Two Novel Primate-Specific Genes in DSCR. DNA Res. 9:89-97, 2002.

Iijima, Y., Okuda, K., Tojo, A., Khanh Tri, N., Setoyama, M., Sakaki, Y., Asano, S., Tokunaga, K., Kruh, G. and Sato, Y. Transformation of Ba/F3 cells and Rat-1 cells by ETV6/ARG. Oncogene 21: 4374-4383, 2002.

Takeuchi, M., Yamamoto, M., Tatematsu, M., Miki, K., Sakaki, Y. and Furihata, C. Dendritic Cell Appearance and Differentiation during Early and Late Stages of Rat Stomach Carcinogenesis. Jpn. J. Caner Res. 93: 925-934, 2002.

Maeng, H-Y., Song, S-B., Choi, D-K., Kim, K-E., Jeong, H-Y., Sakaki, Y. and Furihata, C. Osteonectin-expressing cells in human stomach cancer and their possible clinical significance. Cancer Letters. 184: 117-121, 2002.

Maeng, H-Y., Choi, D-K., Takeuchi, M., Yamamoto, M., Tominaga, M., Tsukamoto, T., Tatematsu, M., Ito, T., Sakaki, Y. and Furihata, C. Appearance of Osteonectin-expressing Fibroblastic Cells in Early Rat Stomach Carcinogenesis and Stomach Tumors Induced with N-Methyl-N'-nito-N-nitrosoguanidine. Jpn. J. Cancer Res. 93: 960-967, 2002.

Toyoda, A., Noguchi, H., Taylor, T., Ito, T., Pletcher, M., Sakaki, Y., Reeves, R. and Hattori, M. Comparative Genomic Sequence Analysis of the Human Chromosome 21 Down Syndrome Critical Region. Genome Res. 12: 1323-1332, 2002.

Wakui, K., Toyoda, A., Kubota, T., Hidaka, E., Ishikawa, M., Katsuyama, T., Sakaki, Y., Hattori, M. and Fukushima, Y. Familial 14-Mb deletion at 21q11.2-q21.3 and variable phenotypic expression. J Hum Genet 47: 511-516, 2002.

Tri, N-K., Xinh, P-T., Nagao, H., Izumi, T., Ozawa, K., Toyoda, A., Hattori, M., Sakaki, Y., Tokunaga, K. and Sato, Y. Identification of the Breakpoints at 1p36.2 and 3p21.3 in an AML (M3) Patient Who Had t(1;3)(p36.2;p21.3) at the Third Relapse. GENES, CHROMOSOMES & CANCER 35: 365-367, 2002.

Constance, C.M., Green, C.B., Tei, H., Block, G.D. Bulla gouldiana period exhibits unique regulation at the mRNA and protein levels. J. Biol. Rhythms. 17:413-427, 2002.

Yamazaki, S., Straume, M., Tei, H., Sakaki, Y., Menaker, M., Block, G.D. Effects of aging on central and peripheral mammalian clocks. Proc Natl Acad Sci U S A. 99:10801-10806, 2002.

Abe, M., Herzog, E.D., Yamazaki, S., Straume, M., Tei, H., Sakaki, Y., Menaker, M., Block, G.D. Circadian rhythms in isolated brain regions. J. Neurosci. 22:350-356, 2002.

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