Human Genome Center
Laboratory of Genome Database

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 will be opened soon from http://ontology.ims.u-tokyo.ac.jp/.

2. Signal transduction pathways and logical inferences

Ken-ichiro FUKUDA[1] and Toshihisa TAKAGI: [1]Computational Biology Research Center, AIST

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. Development of an information extraction system from biological literatures

Yosiyuki KOBAYASHI and Toshihisa TAKAGI

We developed an information extraction (IE) system, which collects protein interactions from biological literatures. The system achieves higher precision and recall rate than other systems because our system analyzes the syntactic structures of sentences and because we collected many language patterns describing protein interactions. Analyzing sentence structures makes possible to extract information from complex structures, such as relative clauses. Collecting language patterns reduces missed information in literatures. We evaluate the system by extracting protein interactions of *Drosophila melanogaster*'s kinases(DMK). We retrieve 1088 abstracts from PubMed for 144 DMKs. We randomly selected and manually analyzed 300 abstracts, and we found 38 protein interactions of DMK. We use these interactions and abstracts describing the interactions as knowledge source for the system. Then, the system collected 34 interactions from 300 abstracts (i.e. recall rate is 87%) as well as 4 interactions which were not found by manual extraction. The system did not extract any incorrect interactions from 300 abstracts (i.e. precision rate is 100%.) Finally, the system collected 121 interactions from 1088 abstracts. Now, we are evaluating extraction performance for yeast and human's kinases.

4. Evolution and functional conservation of eukaryotic kinases among organisms

Asako KOIKE, Kenta NAKAI, and Toshihisa TAKAGI

Phosphorylation and dephosphorylation catalyzed by protein kinases and phosphatases play a crucial role in the process of tissue growth, cell differentiation, and rapid response to the environmental changes. The knowledge of functional conservation and/or uniqueness of protein kinases among organisms is important for understanding the evolution of signaling pathways. We extracted all eukaryotic protein kinase (ePK) domains of *S. pombe, S. cerevisiae, C. elegans, D. melanogaster, H. sapiens, and A. thaliana*; we then constructed their phylogenetic trees to investigate functional conservation among these organisms. The common ancestor of plants and yeasts is expected to have had at least thirty-one ePKs; and about half of them are believed to have composed intracellular signaling, which plays a key role in signal transduction of multicellular organisms. A kinase database, which contains the description of the functional conservation among organisms and signaling pathway information extracted by natural language processing techniques, orthologue information, and an automatic pathway drawing system, has been developed. The kinase database will be open to the public early in 2002.

5. Nonparametric linkage analysis for complex genetic disease with linked agent loci

Osamu OGASAWARA and Toshihisa TAKAGI

Unraveling the genetic determinants of complex diseases is an important task for healthcare because many diseases which exhibit high prevalence rates are complex diseases. Strategies for complex disease mapping usually involve a combination of linkage and association analysis. Linkage analysis can scan the entire genome by fewer genotyping efforts than association studies. In parametric linkage analysis, the mode of inheritance of the agent loci must be specified exactly. But for complex diseases, it is difficult to determine it. So many researchers use model-free (nonparametric) linkage analysis as the first screening step of genetic mapping. We implemented a new nonparametric linkage analysis program that calculates joint genetic effects of several agent loci. One strength of this program is that it can analyze complex diseases with not only unlinked agent loci but also linked agent loci. Compared with single locus analysis methods, some improvements of the statistical power were observed. Because the improvement of the power was much higher in epistatic model than in heterogeneity model, this program can be used to detect epistasis between agent loci.

6. Lag analysis of genetic networks regulating the cell cycle in budding yeast

Mamoru KATO, Tatsuhiko TSUNODA[2], and Toshihisa TAKAGI:[2]SNP Research Center, RIKEN

The recent emergence of whole-genome expression data requires a novel computational method for efficient extraction of biological information from the large-scale data. Here we propose a computational time-series analysis to infer gene-regulatory networks from whole-genome expression data. This analysis determines a time-lagged correlation between the mRNA expressions of a known transcription factor and a gene. When combined with the upstream sequence analysis of consensus regulatory elements, this analysis allowed us to infer a number of unknown target genes of the transcription factors playing a central role in the mitotic cell cycle of *Saccharomyces cerevisiae.* Furthermore, using this method we discovered interesting "lag" kinetics of gene regulation by the transcription factors. Our computational method will sequentially process a large amount of mRNA expression data and extract interesting information regarding as yet unknown genetic control systems.

7. Database for human UV-regulated genes

Makoto YAMAZAKI, Hiroko AO, Goro TERAI, Nobuo KITAMURA[3], and Toshihisa TAKAGI: [3]Kanebo, Ltd

With an increase of the intensity of ultraviolet (UV) irradiation on the surface of the earth, UV has been considered to be a serious environmental risk factor of skin inflammation, photoaging and cancer. To develop an integrated database for UV-regulated genes, we systematically collected their information, including their function, regulatory sequences and transcription factors as well as their expression data. We made a list of 279 UV-regulated genes. 88 genes were selected by the knowledge of experts in dermatology. 125 genes were listed as UV-regulated genes from published microarray data. 66 genes were taken from PubMed abstracts by an automatic extraction procedure. For each of the genes, we collected SWISS-PROT keywords and carried out motif search using PROSITE and Pfam databases. The results were then converted into Gene Ontology terms for 241 genes. We retrieved information of about 230 *cis*-elements (for 40 genes) from TRANSFAC and literatures. We also obtained information of transcription start sites (for 34 genes) from EPD. 228 Upstream regions of UV-regulated genes were retrieved from the GoldenPath sequence data. We collected the expression data after UV irradiation, manually from 149 PubMed abstracts Now, ATAC-PCR experiments are being performed for 144 genes. The data will be stored in our database. We are developing a method for predicting UV-regulated genes as well as a GUI system for efficient comparison between known information and the ATAC-PCR data.

8. Prediction of co-regulated genes in *Bacillus subtilis* on the basis of upstream elements conserved across three closely related species

Goro TERAI, Toshihisa TAKAGI, and Kenta NAKAI

Transcriptional regulatory networks in an organism play an important role for controlling many biological phenomena, such as development and proliferation. Even in bacteria, elucidation of such networks or identification of co-regulated genes is essential in understanding many cellular processes. Moreover, it provides hints toward identifying gene function because co-regulated genes are likely to function for the same purpose. To identify co-regulated genes, the microarray technique, which enables us to monitor the expression levels of thousands of genes in parallel, appears very powerful. However, even if we can ignore its experimental artifacts, it is not always easy to set experimental conditions to identify differential expression patterns of uncharacterized genes. Thus, it would be desirable to develop some computational methods that can supplement such experimental techniques. Although co-regulated genes should have at least one common sequence element, it is generally difficult to identify these genes from the presence of this element because it is very easily obscured by noises. To overcome this problem, we used the conservation information of three closely related species: *Bacillus subtilis*, *Bacillus halodurans*, and *Bacillus stearothermophilus*. Our method consists of two parts; first, we identified Phylogenetically Conserved Elements (PCEs) in the upstream intergenic regions of *B.subtilis* genes; then, they were clustered according to the similarity of PCEs in their upstream region. By comparing our prediction results with various types of known information, such as experimentally verified co-regulated genes and gene functions, we confirmed that although it is difficult to detect entire set of co-regulated genes using our method, it could identify many known and plausible co-regulated genes. Thus, our approach is promising for exploring potentially co-regulated genes.

9. DBTSS: database for human mRNA transcriptional start sites

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

Although the information on full-length cDNAs is indispensable for analyzing gene function, most of the cDNA sequences stored in current databases are imperfect in the sense that they lack the precise information on 5' end termini. To overcome this difficulty, we have developed the oligo-capping method to obtain full-length or 5'-end cDNAs, and constructed new database DataBase of Transcriptional Start Sites (DBTSS: http://elmo.ims.u-tokyo.ac.jp/dbtss/). Among the obtained 217,402 5'-end sequences, 111,382 (7,889 genes) have been corresponded to cDNA sequences of known genes in a reference sequence database, RefSeq. Sequence comparison between our entries and those of RefSeq sequence indicates that 4,683 (34 %) of RefSeq sequences should be extended towards 5'-ends. We also mapped each sequence on the human draft genome sequence to identify its transcriptional start site, which provides us with more detailed information on distribution patterns of transcriptional start sites and adjacent regulatory regions. We have found that transcriptional start sites are very variable, for single, multiple, tissue specific starts were determined. We believe that our DBTSS would be very useful for future analysis for transcriptional regulation.

10."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

Microarray analysis of gene expression profile is the most systematic experimental approach for identifying of genes, whose transcription is regulated by the same transcription factor interacting with DNA in a sequence-specific manner. The detection of transcription control elements (motifs) in the promoter regions requires a powerful computation method and many programs are available nowadays, thus creating problems for scientists who try to select an appropriate method for their needs. But even if the appropriate program was found the User could not compose the parameter set useful for his/her elucidation task, and after the algorithm failed at once in the *default* mode the User gave up other trials. 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 and many parameter sets can be applied concomitantly to one dataset. *Melina* is being tested on artificially constructed datasets, and we concluded about it capability to solve elucidation tasks of various difficulty when the optimum parameters are defined. *Melina* is opened to the public usage (http://elmo.ims.u-tokyo.ac.jp) and we hope that it can be a helpful and convenient tool for many biological researchers.

11. 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. So the presence of 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 of orthologous genes. We have analysed the co-regulated genes of *Bacillus* genus with such an approach (Terai, G., et al. Genome Biology, 2(11): research0048.1-0048.12 (2001)). In that work, some plausible co-regulated genes were proposed. This time, we applied the same method to five closely related *Chlamydia* genomes. As a result, we obtained 960 motifs in 898 orthologous gene groups. One of these motifs corresponded to sigma54 dependent promoter sequences. Further study will enable us to identify potentially co-regulated *Chlamydia* genes. Our results will complement experiments like DNA microarrays. We also plan to study other bacterial genome sequences with the same approach.

12. Extensive feature detection of N-terminal protein sorting signals

Hideo BANNAI[7], Yoshinori TAMADA[7], Osamu MARUYAMA[8], Kenta NAKAI, and Satoru MIYANO[7]:[7]Laboratory of DNA Information Analysis, [8]Kyushu University

The prediction of localization sites of various proteins is an important and challenging problem in the field of molecular biology. TagetP, by Emanuelsson et al. (2000) is a neural network based system which is currently the best predictor in the literature for N-terminal sorting signals. One drawback of neural networks, however, is that it is generally difficult to understand and interpret how and why they make such predictions. Here, we aim to generate simple and interpretable rules as predictors, and still achieve a practical prediction accuracy. We adopt an approach which consists of an extensive search for simple rules and various attributes which is partially guided by human intuition. We have succeeded in finding rules whose prediction accuracies come close to that of TargetP, while still retaining a very simple and interpretable form. We also discuss and interpret the discovered rules. An web service is currently provided at http://www.hypothesiscreator.net/iPSORT/

13. 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

Fukuda, K. and Takagi, T. Knowledge representation of signal transduction pathways, Bioinformatics 17:829-837, 2001.

Fukuda, K. and Takagi,T. Signal transduction pathways and logical inferences, The 2001 International Conference on Mathematics and Engineering (METMBS2001), 297-303, 2001.

Ono, T., Hishigaki, H., Tanigami, A., and Takagi, T. Automated extraction of information on protein-protein interactions from the biological literature, Bioinformatics, 17:155-161, 2001.

Kihara, C., Tsunoda, T., Tanaka, T., Yamana, H., Furukawa, Y., Ono, K., Kitahara, O., Zembutsu, H., Yanagawa, R., Hirata, K., Takagi, T., and Nakamura, Y. Prediction of sensitivity of esophageal tumors to adjuvant chemotherapy by cDNA microarray analysis of gene- expression profiles, Cancer Res. 61:6474-6479, 2001.

Okuno, S., Watanabe, T., Ono, T., Oga, K., Miyataka, A., Yamasaki, Y., Goto, Y., Shinomiya, H., Momota, H., Miyao, H., Hayashi, Isamu., Asai, T., Suzuki, M., Harada, Y., Hishigaki, H., Wakitani, S., Takagi, T., Nakamura, Y., and Tanigami, A. Effects of Dmo1 on obesity, dyslipidaemia and hyperglycaemia in the Otsuka Long Evans Tokuhsima Fatty strain, Genet. Res., Camb. 77:183-190, 2001.

Watanabe, K., Okuno, S., Ono, T., Yamasaki, Y., Oga, K., Miyakita, A., Miyano, H., Suzuki, M., Momota, H., Goto, Y., Shinomiya, H., Hishigaki, H., Hayashi, I., Asai, T., Wakitani, S., Takagi, T., Nakamura, Y., and Tanigami, A. Single-allele correction of the Dmo1 locus in congenic animals substantially attenuates obesity, dyslipidaemia and diabetes phenotypes of the Oletf rat, Clinical and Experimental Pharmacology and Physiology 28:28-42, 2001.

Kitamura, O., Furukawa, Y., Tanaka, T., Kihara, C., Ono, K., Yanagawa, R., Nita, M. E., Takagi, T., Nakamura, Y., nad Tsunoda, T. Alterations of gene expression during colorectal carcinogenesis revealed by cDNA microarrays after laser-capture microdissection of tumor tissues and normal epithelia, Cancer Res. 61:3544-3549, 2001.

Terai, G., Takagi, T., Nakai, K. Prediction of co-regulated genes in *Bacillus subtilis* on the basis of upstream elements conserved across three closely related species, Genome Biol. 2(11):RESEARCH 0048.1-12, 2001.

Ishii, T., Yoshida, K., Terai, G., Fujita, Y., and Nakai, K. DBTBS: A database of *Bacillus subtilis* promoters and transcription factors, Nucleic Acids Res., 29:278-280, 2001.

Mizuno, H., Tanaka, Y., Nakai, K., and Sarai, A. ORIGENE: gene classification based on the evolutionary tree, Bioinformatics, 17:167-173, 2001.

Hishigaki, H., Nakai, K., Ono, T., Tanigami, A., and Takagi, T. Assessment of prediction accuracy of protein function from protein-protein interaction data, Yeast, 18:523-531, 2001.

Bannai, H., Tamada, Y., Maruyama, O., Nakai, K., and Miyano, S. Views: fundamental building blocks in the process of knowledge discovery, Proc. 14th Int. FLAIRS Conf., 233-238, AAAI Press, 2001.

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

Miura, F., Yada, T., Nakai, K., Sakaki, Y., and Ito, T. Differential display analysis of mutants for the transcription factor Pdr1p regulating multidrug resistance in the budding yeast, FEBS Lett., 505:103-108, 2001.

Nakai, K., Prediction of *in vivo* fates of proteins in the era of genomics and proteomics, J. Struct. Biol., 134:103-116, 2001.

高木利久, ゲノム情報科学の目指すもの, 数理科学 No.458, 5-7, 2001.

高木利久, テキストからの情報抽出と辞書構築・機能データベースとオントロジーの構築に向けて, 榊・小原・大木・金久・高木・菅野・小笠原編 ゲノムサイエンスの新たなる挑戦, 蛋白質核酸酵素増刊号, 46:2526-2531, 2001.

小笠原理, 高木利久, SNPs データベース, 高木利久編 ゲノム医科学と基礎からのバイオインフォマティクス, 実験医学増刊号 19:1322-1328, 2001

高井貴子, シグナル伝達パスウェイの情報解析, 計算工学6: 218-221, 2001.

高井貴子, 高木利久, 生命科学のためのオントロジー, 高木利久編 ゲノム医科学と基礎からのバイオインフォマティクス, 実験医学増刊号 19:1337-1343, 2001.

中井謙太, 蛋白質科学におけるバイオインフォマティクス私論, 中村・森川・鈴木・三浦編 新世紀における蛋白質科学の進展, 蛋白質核酸酵素増刊号, 46:1488-1495, 2001.

中井謙太， ゲノムに書き込まれたシグナル情報のコンピュータ解析, 榊・小原・大木・金久・高木・菅野・小笠原編 ゲノムサイエンスの新たなる挑戦, 蛋白質核酸酵素増刊号, 46:2544-2549, 2001.

Human Genome Center

Laboratory of Genome Structure Analysis

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 2001.

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 170,000 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 1,000,000 clones. Now, we have about 20,000 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 for 12300 clones out of 20,000. 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. Furthermore, 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 been 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 in lung tissue. Met proteins were phosphorylated in metastasized lung. Moreover, c-met was overexpressed at transcriptional level, not by a selection process. Finally, c-met was also overexpressed in 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

Yoshihiro Omori, Takushi Togashi, Masaaki Oyama, 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. Homology search revealed that there were significant number of cDNAs which showed similarity to transcription factors. Expression analysis showed that some of them were expressed in a tissue specific manner. 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. Monky 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 macaque brain and more than 40,000 cDNA clones were sequenced at their 5' end and the comparison between human data is in progress.

Publications

Watanabe, J., Sasaki, M., Suzuki Y., Sugano S. FULL-malaria: a database for a full-length enriched cDNA library from human malaria parasite, Plasmodium falciparum. Nucl. Acid Res. 29: 70-71, 2001.

Watanabe, M., Sugano, S, Imai, J., Yoshida, K., Onodera, R., Amin, M. R., Uchida, K., Yamaguchi, R., Tateyama, S. Inhibition of cell proliferation, suppression of tumourigenecity, and induction of differentiation of the canine mammary tumour cell line by sodium phenylacetate. Res. Vet. Sci. 70: 27-32, 2001

Yudate, H. T., Suwa, M., Irie, R., Matsui, H., Nishikawa, T., Nakamura, Y., Yamaguchi, D., Peng, Z. Z., Yamamoto, T., Nagai, K., Hayashi, K., Otsuki, T., Sugiyama, T., Ota, T., Suzuki, Y., Sugano, S., Isogai, T., Masuho, Y. HUNT: launch of a full-length cDNA database from the helix research institute. Nucl. Acid Res. 29: 185-188, 2001.

Moria, S., Tanaka, K., Ohkuma, M., Sugano, S., Kudo, T. Diversification of the microtubule system in the early stage of eukaryote evolution: elongation factor 1 alpha and alpha-tubulin protein phylogeny of termite symbiotic oxymonad and hypermastigote protist. J. Mol. Evol. 52: 6-16, 2001

Shibui, A., Sasaki, M., Sugano, S., Watanabe, J., DNA vaccine using a full-length cDNA library had adverse effects in murine malaria. J. Med. In press

Harada H, Nagai H, Tsuneizumi M, Mikami I, Sugano S, Emi M. Identification of DMC1, a novel gene in the TOC region on 17q25.1 that shows loss of expression in multiple human cancers. J Hum Genet. 46: 90-95, 2001.

Choi DK, Suzuki Y, Yoshimura S, Togashi T, Hida M, Taylor TD, Wang Y, Sugano S, Hattori M, Sakaki Y. Molecular cloning and characterization of a gene expressed in mouse developing tongue, mDscr5 gene, a homolog of human DSCR5 (Down syndrome Critical Region gene 5). Mamm Genome. 12: 347-351, 2001.

Mikami I, Harada H, Nagai H, Tsuneizumi M, Nobe Y, Koizumi K, Sugano S, Tanaka S, Emi Down-regulation in multiple human cancers of a novel gene, DMHC, from 17q25.1 that encodes an integral membrane protein. Jpn J Cancer Res. 92: 417-422, 2001.

Omori Y, Imai J, Watanabe M, Komatsu T, Suzuki Y, Kataoka K, Watanabe S, Tanigami A, Sugano S. CREB-H: a novel mammalian transcription factor belonging to the CREB/ATF family and functioning via the box-B element with a liver-specific expression. Nucleic Acids Res. 29: 2154-2162, 2001.

Suzuki Y, Tsunoda T, Sese J, Taira H, Mizushima-Sugano J, Hata H, Ota T, Isogai T, Tanaka T, Nakamura Y, Suyama A, Sakaki Y, Morishita S, Okubo K, Sugano S. Identification and characterization of the potential promoter regions of 1031 kinds of human genes. Genome Res. 11: 677-684, 2001.

Suzuki Y, Taira H, Tsunoda T, Mizushima-Sugano J, Sese J, Hata H, Ota T, Isogai T, Tanaka T, Morishita S, Okubo K, Sakaki Y, Nakamura Y, Suyama A, Sugano S. Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. EMBO Rep. 2: 388-393, 2001.

Yamaguchi R, Tanimoto N, Tateyama S, Uchida K, Hirano N, Tsuchiya K, Shimizu H, Sugano S. Immunohistochemical study of age-dependent brain lesions in mice infected intracerebrally with Kasba (Chuzan) virus. J Comp Pathol. 124: 36-45, 2001.

Suzuki Y, Sugano S. Construction of full-length-enriched cDNA libraries. The oligo-capping method. Methods Mol Biol. 175: 143-153, 2001.

Nishimoto M, Fukushima A, Miyagi S, Suzuki Y, Sugano S, Matsuda Y, Hori T, Muramatsu M, Okuda A. Structural analyses of the utf1 gene encoding a transcriptional coactivator expressed in pluripotent embryonic stem cells. Biochem Biophys Res Commun. 285: 945-953, 2001.

Hida, M., Suzuki, Y., Sugano, S. Full-length cDNA libraries: Reagents for functional studies of the nervous system. In Method in genomic neuroscience. Edited by Chen, H. R. and Moldin, S. O. (CRC press FL). Pp207-227, 2001.

Date H, Onodera O, Tanaka H, Iwabuchi K, Uekawa K, Igarashi S, Koike R, Hiroi T, Yuasa T, Awaya Y, Sakai T, Takahashi T, Nagatomo H, Sekijima Y, Kawachi I, Takiyama Y, Nishizawa M, Fukuhara N, Saito K, Sugano S, Tsuji S. Early-onset ataxia with ocular motor apraxia and hypoalbuminemia is caused by mutations in a new HIT superfamily gene. Nature Genet. 29: 184-188, 2001.

Osada N, Hida M, Kususda J, Tanuma R, Iseki K, Hirata M, Suto Y, Hirai M, Terao K, Suzuki Y, Sugano S, Hashimoto K. Assignment of 118 novel cDNAs of cynomolgus monkey brain to human chromosomes. Gene. 275: 31-37, 2001.

Kato M, Seki N, Sugano S, Hashimoto K, Masuho Y, Muramatsu M, Kaibuchi K, Nakafuku M. Identification of sonic hedgehog-responsive genes using cDNA microarray. Biochem Biophys Res Commun. 289: 472-478, 2001.

Kurotaki N, Harada N, Yoshiura K, Sugano S, Niikawa N, Matsumoto N. Molecular characterization of NSD1, a human homologue of the mouse Nsd1 gene. Gene. 279: 197-204, 2001.

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. In press.

Human Genome Center

Laboratory of DNA Information Analysis

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 Analyzing Gene Expression Profiles

a. Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression

Seiya Imoto, Takao Goto, Satoru Miyano

We propose a new method for constructing genetic network from gene expression data by using Bayesian networks. We use nonparametric regression for capturing nonlinear relationships between genes and derive a new criterion for choosing the network in general situations. In a theoretical sense, our proposed theory and methodology include previous methods based on Bayes approach. We applied the proposed method to the *S. cerevisiae* cell cycle data and showed the effectiveness of our method by comparing with previous methods.

b. Selecting informative genes for cancer classification using gene expression data

Tatsuya Akutsu, Satoru Miyano

Recently, several methods have been proposed for classification of cancer cells based on gene expression monitoring by DNA microarrays. In these methods, not all genes were used for classification, but several tens of genes that were relevant to class distinction were selected and used. In this paper, this selection problem is formalized using threshold functions for Boolean variables. A simple greedy algorithm is also proposed for the selection problem. This greedy algorithm was compared with two other algorithms using real gene expression data obtained from human acute leukemia patients by Golub *et al.* The results of comparison show that the greedy algorithm is as good as the other two algorithms for the test data set and is much better for the training data set.

2. Knowledge Discovery Systems

a. Extensive feature detection of N-terminal protein sorting signals

Hideo Bannai, Yoshinori Tamada[1], Osamu Maruyama[2], Kenta Nakai, Satoru Miyano: [1]Department of Mathematics, Tokai University and [2]Department of Mathematics, Kyushu University

The prediction of localization sites of various proteins is an important and challenging problem in the field of molecular biology. TargetP, by Emanuelsson *et al.* (2000) is a neural network based system which is currently the best predictor in the literature for N-terminal sorting signals. One drawback of neural networks, however, is that it is generally difficult to understand and interpret how and why they make such predictions. In this paper, we aim to generate simple and interpretable rules as predictors, and still achieve a practical prediction accuracy. We adopt an approach which consists of an extensive search for simple rules and various attributes which is partially guided by human intuition. We have succeeded in finding rules whose prediction accuracies come close to that of TargetP, while still retaining a very simple and interpretable form. We also discuss and interpret the discovered rules. An (experimental) web service using rules obtained by our method is provided at http://hypothesiscreator.net/iPSORT/.

b. VML: a view modeling language for computational knowledge discovery

Hideo Bannai, Yoshinori Tamada[1], Osamu Maruyama[2], Satoru Miyano

We present the concept of a functional programming language called VML (View Modeling Language), providing facilities to increase the efficiency of the iterative, trial-and-error cycle which frequently appears in any knowledge discovery process. In VML, functions can be specified so that returning values implicitly "remember", with a special internal representation, that it was calculated from the corresponding function. VML also provides facilities for "matching" the remembered representation so that one can easily obtain, from a given value, the functions and/or parameters used to create the value. Further, we describe, as VML programs, successful knowledge discovery tasks which we have actually experienced in the biological domain, and argue that computational knowledge discovery experiments can be efficiently developed and conducted using this language.

c. Views: fundamental building blocks in the process of knowledge discovery

Hideo Bannai, Yoshinori Tamada[1], Osamu Maruyama[2], Satoru Miyano

We present a novel approach to describe the knowledge discovery process, focusing on a generalized form of attribute called view. It is observed that the process of knowledge discovery can, in fact, be modeled as the design, generation, use, and evaluation of views, asserting that views are the fundamental building blocks in the discovery process. We realize these concepts as an object oriented class library and conduct computational knowledge discovery experiments on biological data, namely the characterization of N-terminal protein sorting signals, yielding significant results.

d. HypothesisCreator: concepts for accelerating the computational knowledge discovery process

Hideo Bannai, Yoshinori Tamada[1], Osamu Maruyama[2], Satoru Miyano

We summarize and discuss the work accomplished through the HypothesisCreator project, an ongoing effort whose aim is to develop systematic methods to accelerate the computational knowledge discovery process. A novel approach to describe the knowledge discovery process, focusing on a generalized form of attribute called view, is presented. It is observed that the process of knowledge discovery can, in fact, be modeled as the design, generation, use, and evaluation of views, asserting that views are the fundamental building blocks in the discovery process. Also, we note that the trial-and-error cycle inherent in any knowledge discovery process, can be formulated as the composing and decomposing of views. To assist this cycle, a programming language called VML (View Modeling Language) was designed, providing facilities for this purpose. Following this view oriented perspective, we describe our approach to the problem of characterizing N-terminal protein sorting signals in which we have obtained significant results.

e. Foundations of designing computational knowledge discovery processes

Yoshinori Tamada[1], Hideo Bannai, Osamu Maruyama[2], Satoru Miyano

We propose a new paradigm for computational knowledge discovery, called VOX (View Oriented eXploration). Recent research has revealed that actual discoveries cannot be achieved using only component technologies such as machine learning theory or data mining algorithms. Recognizing how the computer can assist the actual discovery tasks, we developed a solution to this problem. Our aim is to construct a principle of computational knowledge discovery, which will be used for building actual applications or discovery systems, and for accelerating such entire processes. VOX is a mathematical abstraction of knowledge discovery processes, and provides a unified description method for the discovery processes. We present advantages obtained by using VOX. Through an actual computational experiment, we show the usefulness of this new paradigm. We also designed a programming language based on this concept. The language is called VML (View Modeling Language), which is defined as an extension of a functional language ML. Finally, we present the future plans and directions in this research.

f. More Speed and more pattern variations for knowledge discovery system BONSAI

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

BONSAI is a machine learning system for knowledge acquisition from positive and negative examples of strings. A hypothesis generated by the system is a pair of a classification of symbols called an alphabet indexing, and a decision tree over regular patterns, which classifies given examples (strings) to either positive or negative. The algorithm of the system consists of two parts: a learning algorithm for constructing a decision tree over regular patterns, and a local search algorithm for finding a good alphabet indexing for the production of the decision tree. Our focus here is in the improvement of the former, increasing both the speed of hypothesis construction, and the descriptional strength of the generated hypotheses. It has been reported that the system has discovered knowledge which can classify amino acid sequences of trans-membrane domains and randomly chosen amino acid sequences located in other parts of the PIR database, with over 90% accuracy. However, in the current implementation, only substring patterns (i.e. whether or not a string pattern appears as a substring of the data string) are searched for, and such patterns may not be powerful enough for distinguishing between positive and negative data of a more complex nature. In this paper, we present a new version of the BONSAI system which implements several, more powerful variations of patterns, namely, subsequence patterns, approximate patterns, and episode patterns. We also implement an efficient branch-and-bound algorithm for finding the best pattern which distinguishes between the positive and negative data sets.

g. Learning conformation rules

Osamu Maruyama[2], Takayoshi Shoudai[3], Emiko Furuichi[4], Satoru Kuhara[5], Satoru Miyano: [4]Fukuoka Women's Jinior College, [5]Graduate School of Genetic Resources Technology, Kyushu University

Protein conformation problem, one of the hard and important problems, is to identify conformation rules which transform sequences to their tertiary structures, called conformations. Our aim of this work is to give a concrete theoretical foundation for graph-theoretic approach for the protein conformation problem in the framework of a probabilistic learning model. We propose the conformation problem as a learning problem from hypergraphs capturing the conformations of proteins in a loose way. We consider several classes of functions based on conformation rules, and show the PAC-learnability of them. The refutable PAC-learnability of functions is discussed, which would be helpful when a target function is not in the class of functions under consideration. We also report the conformation rules learned in our preliminary computational experiments.

3. Biopathway Simulations and Systems Biology

a. XML documentation of biopathways and their simulations in Genomic Object Net

Hiroshi Matsuno[6], Atsushi Doi[6], Yuichi Hirata, Satoru Miyano:[6]Faculty of Science, Yamaguchi University

Genomic Object Net is a software tool for modeling and simulating biopathways which employs the notion of hybrid functional net as its basic architechture. This paper shows how to integrate this basic architecture with XML documents for biopathway representations, simulations, and visualizations for creating a tailor-made simulation environment.

b. Simulation of the pattern formation in multicellular organism by Genomic Object Net

Atsushi Doi[6], Rie Yamane[6], Naoyuki Yamasaki[6], Haruka Yoshimori[6], Ryutaro Murakami[6], Hiroshi Matsuno[6], Satoru Miyano

By using Genomic Object Net, we constructed Delta-Notch mechanism, which is working in a pattern formation of multicellular organism. In the simulation, the number of the neural precursor cells increases when the value of Notch signal decreases, which is thought to represent the situation of Notch mutation.

c. Biopathway model conversion from E-Cell to Genomic Object Net

Mika Matsui[6], Atsushi Doi[6], Hiroshi Matsuno[6], Yuichi Hirata, Satoru Miyano

E-Cell is a cenceputually attractive biosimulation tool for representing and simulating biopathways. With E-Cell, Tomita *et al*. have modeled several biopathways including biochemical reactions in human erythrocyte, signal transduction for bacterial chemotaxis, energy metabolism in mitochondria and lytic-lysogenic switch network of λ phage. On the other hand, we developed a tool for representing and simulating biopathways called Genomic Object Net which uses hybrid functional Petri net as a basic mechanism for representing biopathways, where XML documentation of biopathways and their simulatons is also realized as another tool in Genomic

Object Net. The purpose of this paper is to show a procedure for converting biopathway models with E-Cell to the ones executable on Genomic Object Net. Thus E-Cell can be regarded as a subset of Genomic Object Net. A conversion program of E-Cell to Genomic Object Net is being developed.

d. Genomic Object Net: Petri net enhancement for multi-cellular processes

Yuichi Hirata, Hiroshi Matsuno[6], Makiko Sasaki, Satoru Miyano

Recent study of cell lineage of multi-cellular organism such as *C. elegans* shows us that cell-cell interaction and localization of gene products take very important role in the development of early embryo. In order to describe such multi-cellular processes, we extend the software tool Genomic Object Net in which we enhanced the hybrid parametrized functional Petri net model to extend the parameter space of elements of Petri net and carry out the dynamical change of Petri net.

e. Automatic acquisition of cell lineage through 4D micorscopy and analysis of early *C. elegans* embryogenesis

Shiichi Onami[7], Shugo Hamahashi[7], Masao Nagasaki, Satoru Miyano, Hiroaki Kitano[7]:[7]Kitano Sybiotic Systems Project, ERATO, JST

Cell lineage analysis is an important technique for studying the development of multicellular organisms. We have developed a system that automatically acquires cell lineages of *C. elegans* from the 1-cell stage up to approximately the 30-cell stage. The system utilizes a set of 4D Nomarski DIC microsope images of *C. elegans* embryo consisting of more than 50 focal plane images at each minute for about 2 hours. The system detects the region of cell nuleus in each of the images, and makes 3D nucleus regions, each of which is a complete set of nucleus regions that represent the same nucleus at the same time point. Each pair of 3D nucleus regions is then connected, if they represent the same nucleus and their time points are consecutive, and the cell lineage is created based on these connections. The resulting cell lineage consists of the three-dimensional positions of nuclei at each time point and their lineage. Encouraged by the performance of our system, we have started systematic cell lineage analysis of *C. elegans*, which will produce a large amount of quantitative data essential for system-level understanding of *C. elegans* embryogenesis.

Publications

Akutsu, T., Miyano, S. Selecting informative genes for cancer classification using

gene expression data. Proc. IEEE-EURASIP Workshop on Non-linear Signal and Image Processing - NSIP-01, 1-6, 2001.

Bannai, H., Tamada, Y., Maruyama, O., Nakai, K., Miyano, S. Extensive feature detection of N-terminal protein sorting signals. Bioinformatics, in press.

Bannai, H., Tamada, Y., Maruyama, O., Miyano, S. VML: a view modeling language for computational knowledge discovery. Lecture Notes in Artificial Intelligence 2226:30-44, 2001.

Bannai, H., Tamada, Y., Maruyama, O., Nakai, N., Miyano, S. Views: fundamental building blocks in the process of knowledge discovery. Proceedings of the 14th International FLAIRS Conference, 233-238, AAAI Press, 2001.

Bannai, H., Tamada, Y., Maruyama, O., Miyano, S. HypothesisCreator: concepts for accelerating the computational knowledge discovery process. Linkoping Electronic Articles in Computer and Information Science Vol. 6, in press.

Bannai, H., Iida, K., Shinohara, A., Takeda, M., Miyano, S. More speed and more pattern variations for knowledge discovery system BONSAI. Genome Informatics 12:454-455, 2001

Doi, A., Yamane, R., Yamasaki, N., Yoshimori, H., Murakami, R., Matsuno, H., Miyano, S. Simulation of the pattern formation in multicellular organism by Genomic Object Net. Genome Informatics 12:288-289, 2001.

Hirata, Y., Matsuno, H., Sasaki, M., Miyano, S. Genomic Object Net: Petri net enhancement for multi-cellular processes. Genome Informatics 12:292-293, 2001.

Imoto, S., Goto, T., Miyano, S. Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression. PSB 2002, in press.

Maruyama, O., Shoudai, T., Furuichi, E., Kuhara, S., Miyano, S. Learning conformation rules. Lecture Notes in Artificial Intelligence 2226:243-257, 2001.

Matsuda, H., Wong, L., Miyano, S., Takagi, T. (eds.). Genome Informatics 2001, Universal Academy Press, 2001.

Matsui, M., Doi, A., Matsuno, H., Hirata, Y., Miyano, S. Biopathway model conversion from E-Cell to Genomic Object Net. Genome Informatics 12:290-291, 2001.

Matsuno, H., Doi, A., Drath, R., Miyano, S. Genomic Object Net: hybrid Petri net for describing biological systems. Currents in Computational Molecular Biology 2001, El-Mabrouk, N., Lengauer, T., Sankof, D. (eds., Les Publications CRM, Montreal, 233-234, 2001.

Matsuno, H., Doi, A., Fujita, S., Sasaki, M., Hirata, Y., Miyano, S. Genomic Object Net: XML visualization of simulation results from

biological modeling on hybrid functional Petri net. Genome Informatics 12:239-240, 2001.

Matsuno, H., Doi, A., Hirata, Y., Miyano, S. XML documentation of biopathways and their simulations in Genomic Object Net. Genome Informatics 12:54-62, 2001.

Miyano, S., Rangahathan, S. The Asia-Pacific Regional Perspective on Bioinformatics. IEEE Intelligent Systems 16(6):19-20, 2001.

Nakayashiki, T., Ebihara, K., Bannai, H., Nakamura, Y. Yeast [PSI+] "prions" that are crosstransmissible and susceptible beyond a species barrier through a quasi-prion state. Molecular Cell 7(6):1121-1130, 2001.

Onami, S., Hamahashi, S., Nagasaki, M., Miyano, S., Kitano, H., Automatic acquisition of cell lineage through 4D micorscopy and analysis of early *C. elegans* embryogenesis. Foundations of Systems Biology, Kitano, H. (ed.), MIT Press, 39-55, 2001.

Sim, K.L., Uchida, T., Miyano, S. ProDDO: a database of disordered proteins form the Protein Data Bank (PDB). Bioinformatics 17(4):379-380, 2001.

Tamada, Y. Bannai, H., Maruyama, O., Miyano, S. Foundations of designing computational knowledge discovery processes. Progresses in Discovery Science (Lecture Notes in Artificial Intelligence), Springer-Verlag, in press.

Waddell, P.J., Kishino, H., Ota, R. A phylogenetic foundation for comparative mammalian genomics. Genome Informatics 12:141-154, 2001.

安道知寛，井元清哉，小西貞則：動径基底関数ネットワークに基づく非線形回帰モデルとその推定．応用統計学，30(1):19-35, 2001.

大浪修一，長崎正朗，宮野悟，北野宏明：Bio-calculus: 生物シミュレーションのための知識表現形式．システムバイオロジーの展開，北野宏明編，シュプリンガー･フェアラーク東京．47-56, 2001.

松野浩嗣，Drath, R.，宮野悟：遺伝子制御ネットワークのハイブリッドペトリネットによるシミュレーション．システムバイオロジーの展開，北野宏明編，シュプリンガー・フェアラーク東京．165-176, 2001.

松野浩嗣，宮野悟：ゲノムデータからの知識発見と遺伝子オブジェクトのシミュレーション．電子情報通信学会誌，84(5):356-358, 2001.

坂内英夫，玉田嘉紀，丸山修，宮野悟：ゲノムデータからの知識発見支援システムーソフトウエアライブラリHypothesis Creator．蛋白質核酸酵素，46(16):2555-2560, 2001.

Human Genome Center

Laboratory of Molecular Medicine
Laboratory of Genome Technology

The major goal of the Human Genome Project is to identify genes predisposing to diseases, and 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, and some allergic diseases. By means of technologies developed through the genome project including high-resolution genetic maps, a large scale DNA sequencing, and the differential display 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, Hiroshi Nakanishi, Kyong-Ah Yoon, Kouji Yoshida, Takashi Kimura, Ching C. Ng, Megumi Iiizumi, Chizu Tanikawa, Keiko Yamazaki, Yoshio Anazawa 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. That is, p53 appears to participate directly in DNA repair by inducing p53R2 in response to DNA damage. Thus inactivation of p53 could directly interfere with damage-induced transcription of p53R2, enhance mis-incorporation of dNTPs and increase the frequency of mutations, resulting in genomic instability in cancers where p53 has undergone mutation. 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.

Through direct cloning of p53-binding sequences from human genomic DNA, we have isolated a novel gene, designated p53AIP1 (p53-regulated Apoptosis Inducing Protein 1), whose expression is inducible by wild-type p53. Ectopically expressed p53AIP1, which is localized within mitochondria, leads to apoptotic cell death through dissipation of mitochondrial ΔΨm. We have found that upon severe DNA damage, Ser46 on p53 is phosphorylated and apoptosis is induced. In addition, substitution of Ser46 inhibits the ability of p53 to induce apoptosis, and selectively blocks expression of p53AIP1. Our results suggest that p53AIP1 is likely to play an important role in mediating p53-dependent apoptosis, and phosphorylation of Ser46 regulates the transcriptional activation of this apoptosis-inducing gene. Ectopic expression of p53AIP1 induced down-regulation of mitochondrial ΔΨm and release of cytochrome c from mitochondria in human cells. Immunoprecipitation and immunostaining experiments indicated interaction between p53AIP1 and bcl-2 proteins at mitochondria. Over-expression of bcl-2 blocked the down-regulation of mitochondrial ΔΨm and the pro-apoptotic activity of p53AIP1. Our results implicate p53AIP1 as a pivotal mediator of the p53-dependent mitochondrial apoptotic pathway, and suggest that gene therapy involving p53AIP1 may be useful for treatment of p53-resistant tumors.

We also isolated a novel p53-inducible gene, termed p53DINP1 (p53-dependent Damage Inducible Nuclear Protein 1). Cell death induced by DSBs, as well as Ser46-phosphorylation of p53 and induction of p53AIP1, were blocked when we inhibited expression of p53DINP1 by means of an antisense oligonucleotide. Over-expression of p53DINP1 and DNA damage by DSBs synergistically enhanced Ser46-phosphorylation of p53, induction of p53AIP1 expression, and apoptotic cell death. Furthermore, the protein-complex interacting with p53DINP1 was shown to have a function to phosphorylate Ser46 of p53. Our results suggest that p53DINP1 may regulate p53-dependent apoptosis through phosphorylation of p53 at Ser46, serving as a cofactor for the putative p53-Ser46-kinase.

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, U373MG and SW480 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 characterized a gene, termed p53ABC1L (p53-inducible Activity of bc1 complex 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 *p53ABC1L* product was located in mitochondria, and colony-formation assays with cancer-cell lines indicated its ability to suppress cell growth. Inhibition of p53ABC1L expression by transfection with antisense oligonucleotide significantly reduced the apoptotic response induced by wild-type p53. These results suggest that p53ABC1L may play an important role in mediating p53-inducible apoptosis through the mitochondrial pathway.

A cDNA-microarray analysis indicated that *semaphorin3B*, a gene whose product is involved in axon guidance and axonal repulsion, is inducible by p53. Introduction of exogenous p53 into a glioblastoma cell line lacking wild-type p53 (U373MG) dramatically induced expression of *Semaphorin3B* mRNA, and introduction of semaphorin3B into other p53-defective cells reduced the number of colonies in colony-formation assays. 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 p53+/+ cells, but not in p53-/- cells, in response to genotoxic stresses caused by adriamycin treatment or UV irradiation. These results suggest that Semaphorin3B might play some role in regulating cell growth, as a mediator of p53 tumor-suppressor activity.

b. APC, β-catenin, and Axin in human cancer

Yoichi Furukawa, Ryuji Hamamoto, Limeihua, Meiko Takahashi, Takashi Shimokawa, Tatsushi Katoh, Nobutomo Miwa, Ryuichiro Yagyu, Suguru Hasegawa, Takeshi Watanabe, Daisuke Yuki, and Yusuke Nakamura

Axin, an important regulator of β-catenin, is frequently mutated in human hepatocellular carcinomas (HCCs), and transduction of the wild-type Axin gene (*AXIN1*) induces apoptosis in HCC cells as well as in colon cancer cells. To investigate the detailed biological function of Axin, we searched on a cDNA microarray for genes whose expression was altered by transfer of wild-type *AXIN1* into colon-cancer cell line LoVo. Among the genes showing altered expression, we focused on one, termed *AXUD1* (AXIN1 up-regulated), that revealed enhanced expression in response to exogeneously expressed *AXIN1* but not to *LacZ*, a control gene. The *AXUD1* gene consists of 5 exons and encodes a transcript with an open reading frame of 1767 bp. A 3.2-kb transcript of *AXUD1* was expressed in all human tissues examined, most abundantly in lung, placenta, skeletal muscle, pancreas and leukocyte. By radiation-hybrid mapping we assigned its chromosomal location at 3p22, a region where frequent loss of heterozygosity has been reported in lung, renal, prostate, breast and cervical cancers. *AXUD1* was frequently down-regulated in lung, kidney, liver and colon cancers compared with their corresponding normal tissues, suggesting that *AXUD1* may have a tumor-suppressor function in those organs.

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.

The ectodermal-neural cortex 1 (*ENC1*) gene was down-regulated in response to Ad-Axin. The promoter activity of *ENC1* was elevated approximately three-fold by transfection of an activated form of β-catenin as well as by that of the wild type Tcf4 in HeLa cells. Semiquantitative RT-PCR experiments revealed that expression of *ENC1* was increased in more than two-thirds of 24 primary colon-cancer tissues we examined compared with corresponding non-cancerous mucosae. Introduction of exogenous *ENC1* increased the growth rate of HCT116 colon-cancer cells in serum-depleted medium. In other experiments, over-expression of *ENC1* in HT-29 colon-cancer cells suppressed the usual increase of two differentiation markers, in response to treatment with sodium butyrate, a differentiation-inducible agent. These data suggest that *ENC1* is regulated by the β-catenin/Tcf pathway and that its altered expression may contribute to colorectal carcinogenesis by suppressing differentiation of colonic cells.

The Claudin-1 (*CLDN1*) gene is also shown to be one of the genes regulated by β-catenin. Not only did expression of *CLDN1* in SW480 colon-cancer cell decrease significantly in response to reduction of intracellular β-catenin by adenovirus-mediated transfer of wild-type *APC* into the APC-deficient cancer cells, but two putative Tcf4-binding elements in the 5′ flanking region of *CLDN1* were confirmed to be responsible for activating its transcription. We documented increased expression of *CLDN1* in all 16 primary colorectal cancers we examined, compared with adjacent non-cancerous mucosae. Our results imply that Claudin-1 is involved in the β-catenin-Tcf/LEF signaling pathway, and that increased expression of Claudin-1 may have some role in colorectal tumorigenesis.

The complex signaling pathways leading to remodeling of the actin cytoskeleton, a process that plays a critical role in cell adhesion and migration, involve many different components. Although members of the Rho family of small guanosine triphosphatases (Rho-GTPases) have emerged as key coordinators of these pathways, the precise regulatory mechanisms remain to be resolved. Here we report isolation of a novel human gene, *ARHGAP9*, which encodes a protein containing a Rho-GTPase activating protein (Rho-GAP) domain, a src-homology 3 (SH3) domain, a pleckstrin homology (PH) region, and a WW domain. *In vitro*, the recombinant protein revealed substantial GAP activity toward Cdc42Hs and Rac1, and less toward RhoA. The transcript was predominantly expressed in peripheral blood leukocytes, spleen and thymus. Exogenous expression of the entire coding region of *ARHGAP9* into human leukemia KG-1 cells repressed adhesion of the cells to fibronectin, without changing the amounts of integrins a4, a5 or b1. Our results indicate that *ARHGAP9* is involved in regulating adhesion of hematopoietic cells to extracellular matrix by way of Rho-GTPase-mediated signal transduction.

c. Growth suppression of human ovarian cancer cells by adenovirus-mediated transfer of the PTEN Gene

Motoko Unoki and Yusuke Nakamura

Defects in *PTEN*, a tumor suppressor, have been found in cancers arising in a variety of human tissues. To elucidate the tumor-suppressive function of this gene, we have been analyzing expression profiles of cancer cells after introduction of exogenous *PTEN*. Those experiments identified 99 candidate genes that were transcriptionally transactivated. Among them, we report here the further analyses of eight genes, *EGR2/Krox-20, BPOZ, APS, HCLS1/HS1, DUSP1/MKP1, NDRG1/Drg1/RTP, NFIL3/E4BP4*, and a novel gene (*PINK1, PTEN-induced putative kinase*). Expression of six of them (*PINK1, EGR2, HCLS1, DUSP1, BPOZ and NFIL3*) was decreased in ovarian tumors compared with corresponding normal tissues. Colony-formation assays using plasmid clones designed to express each gene indicated that

EGR2 and *BPOZ* were able to suppress growth of cancer cells significantly; in particular, cancer-cell lines stably expressing BPOZ grew more slowly than control cells containing mock vector. Flow cytometry suggested that over-expression of BPOZ inhibited progression of the cell cycle at the G_1/S transition. Anti-sense oligonucleotides for *BPOZ* or *EGR2* effectively inhibited their expression, and cell growth was accelerated. Therefore both genes appear to be novel candidates as mediators of the PTEN growth-suppressive signaling pathway.

d. cDNA microarray analysis of cancer

Toyomasa Katagiri, Hitoshi Zenbutsu, Yoichi Furukawa, Toshihiro Tanaka, Tatsuhiko Tsunoda, Takehumi Kikuchi, Satoshi Kakiuchi, Toru Nakamura, Koichi Okada, Kensuke Ochi, Yasuyuki Kaneta, Kenichi Horikoshi, Satoshi Nagayama, Toshihisa Takagi, and Yusuke Nakamura

To disclose detailed genetic mechanisms in hepatocellular carcinoma (HCC) with a view toward development of novel therapeutic targets, we analyzed expression profiles of 20 primary HCCs and their corresponding non-cancerous tissues by means of cDNA microarrays consisting of 23,040 genes. Up-regulation of mitosis-promoting genes was observed in the majority of the tumors examined. Some genes showed expression patterns in HBV-positive HCCs that were different from those in HCV-positive HCCs; most of them encoded enzymes that metabolize carcinogens and/or anti-cancer agents. Furthermore, we identified a number of genes associated with malignant histological type or invasive phenotype. Accumulation of such data will make it possible to define the nature of individual tumors, to provide clues for identifying new therapeutic targets, and ultimately to optimize treatment of each patient. From among the transcripts that were commonly up-regulated in these tumors we identified a novel human gene at chromosomal band 1p36.13, termed *DDEFL1* (development and differentiation enhancing factor-like 1), encoding a product that shared structural features with centaurin-family proteins. The deduced 903-amino-acid sequence showed 46% homology to DDEF/ASAP1 (development and differentiation enhancing factor), and contained an Arf GTPase-activating protein (ArfGAP) domain and two ankyrin repeats. Gene transfer of *DDEFL1* promoted proliferation of cells that lacked endogenous expression of this gene. Furthermore, reduction of *DDEFL1* expression by transfection of anti-sense S-oligonucleotides inhibited the growth of SNU475 cancer cells, in which *DDEFL1* expression was highly up-regulated. Our results provide novel insight into hepatocarcinogenesis and may contribute to development of new strategies for diagnosis and treatment of HCC.

We also isolated a novel human gene, termed *MARKL1* (*MAP/microtubule affinity-regulating kinase like 1*), whose expression was down-regulated in response to decreased Tcf/LEF1 activity. The transcript expressed in liver consisted of 3,529 nucleotides that contained an open reading frame of 2,256 nucleotides, encoding 752 amino acids homologous to human *MARK3*. Expression levels of *MARKL1* were markedly elevated in 8 of 9 HCCs in which nuclear accumulation of β-catenin were observed, which may suggest that MARKL1 plays some role in hepatocelular carcinogenesis.

In spite of intensive and increasingly successful attempts to determine the multiple steps involved in colorectal carcinogenesis, the mechanisms responsible for metastasis of colorectal tumors to the liver remain to be clarified. To identify genes that are candidates for involvement in the metastatic process, we analyzed genome-wide expression profiles of ten primary colorectal cancers (CRCs) and their corresponding metastatic lesions by means of a cDNA microarray consisting of 9121 human genes. This analysis identified 40 genes whose expression was commonly up-regulated in metastatic lesions, and seven that were commonly down-regulated. The up-regulated genes encoded proteins involved in cell adhesion, or remodeling of the actin cytoskeleton. Investigation of the functions of more of the altered genes should improve our understanding of metastasis and may identify diagnostic markers and/or novel molecular targets for prevention or therapy of metastatic lesions.

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 anti-cancer drugs for individual patients. To identify genes that might be associated with chemosensitivity, we used a cDNA microarray representing 23,040 genes to analyze expression profiles in a panel of 85 cancer xenografts derived from nine human organs. The xenografts, implanted into nude mice, were examined for sensitivity to nine anti-cancer drugs (5FU, ACNU, ADR, CPM, DDP, MMC, MTX, VCR, and VLB). 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.

We further applied cDNA microarray analyses of 9,216 genes to establish a genetic method for predicting the outcome of adjuvant chemotherapy to esophageal cancers. We analyzed expression profiles of 20 esophageal-cancer tissues from patients who were treated with the same adjuvant chemotherapy after removal operation of tumor, and attempted to find genes associated with the duration of survival after surgery. By comparing expression profiles of those cancer tissues, we identified by statistical analysis 52 genes that were likely to be correlated with prognosis and possibly with sensitivity/resistance to the anti-cancer drugs. We further developed a "drug response score (DRS)" on the basis of differential expression of these genes and found a significant correlation between DRS and individual patients' prognoses. Our results indicated that this scoring system, based on microarray analysis of selected genes, is likely to have great potential for predicting the prognosis of individual cancer patients with the adjuvant chemotherapy.

2. Genes responsible for other diseases

a. Bronchial asthma

Sachiyo Takeoka, Motoko Unoki , Yoshihiro Onouchi , Satoru Doi[2], Hiroshi Fujiwara[2], Akihiko Miyatake[3], Kimie Fujita[4], Ituro Inoue[1], Yusuke Nakamura, and Mayumi Tamari:[1]Division of Genetic Diagnosis, Institute of Medical Science, University of Tokyo, Minato-ku, Tokyo, Japan;[2]Osaka Prefectural Habikino Hospital, Osaka, Japan;[3]Miyatake Asthma Clinic, Osaka, Japan;[4]College of Nursing, University of Shiga, Japan

The complex etiology of bronchial asthma (BA), one of the most common inflammatory diseases throughout the world, involves a combination of various genetic and environmental factors. A number of linkage and association studies have been performed to shed light on the genetic background of BA, but the genetic aspects are still poorly understood. In the course of a project to screen the entire human genome for single nucleotide polymorphisms (SNPs) that might represent useful markers for large-scale association analyses of common diseases and pharmacogenetic traits, we identified six SNPs within the gene encoding IkB-associated protein (IKAP), a regulator of the NF-kB signal pathway. Most of the SNPs were in linkage disequilibrium each other. A strong allelic association between BA in childhood and two SNPs, T3214A (Cys1072Ser) and C3473T (Pro1157Leu), was observed (P=0.000004 for T3214A and P=0.0009 for C3473T). T3214A was also associated with BA in adult (P=0.000002), while C3473T was not (P=0.056). To confirm the above results, haplotype frequencies with six SNPs were estimated and compared between BA patients and controls. A strong association with the BA in childhood and a specific haplotype, 819T, 2295G, 2446A, 2490A, 3214A, and 3473T (haplotype TGAAAT), (P=0.00004, Odds ratio 2.94, 95%CI=2.48-3.4), where two amino-acid substitutions are present. Interestingly, the other haplotype TACGTC, in which the last five nucleotides were different from the haplotype TGAAAT, was inversely correlated with the BA phenotype (P=0.002, Odds ratio 9.83, 95%CI=8.35-11.31). These results indicated that specific variants of the *IKAP* or a variant in linkage disequilibrium with the specific haplotype might be associated with mechanisms responsible for early-onset BA.

b. Rheumatoid arthritis

Ryo Yamada[1], Toshihiro Tanaka, Motoko Unoki, Tatsuo Nagai[2], Tetsuji Sawada[2], Yozo Ohnishi, Tatsuhiko Tsunoda[1], Masao Yukioka[3], Akira Maeda[4], Kenji Suzuki[2], Hiroomi Tateishi[4], Takahiro Ochi[5], Yusuke Nakamura, Kazuhiko Yamamoto[2]:[1]SNP Research Center, Institute of Physical and Chemical Research (RIKEN), Tokyo, Japan;[2]Department of Allergy and Rheumatology, Graduate School of Medicine, University of Tokyo;[3]Department of Orthopedic Surgery, Yukioka Hospital;[4]Sasayama-Hospital, Hyogo College of Medicine;[5]Department of Orthopedics, Osaka University Medical School

Genetic variants of interleukin-3 (IL-3), a well-studied cytokine, may have a role in the pathophysiology of rheumatoid arthritis (RA) but reports on this association sometimes conflict. We designed a case-control study to investigate association between RA and a single nucleotide polymorphism (SNP) in the IL-3 promoter region. Comparison of RA cases with controls yielded a chi square value of 14.28 (p=0.0002) with a genotype relative risk of 2.24 [95% CI, 1.44, 3.49]. When we compared younger-onset female cases with female controls, the SNP revealed an even more significant correlation, with a chi square value of 21.75 (p=0.000004) and a genotype relative risk of 7.27 [2.80, 18.89]. The stronger association we observed in this clinically distinct subgroup (females, early onset) within a region where linkage disequilibrium was not significantly extended suggested that the genuine RA locus should locate within or close to the IL-3 gene. We combined genotypic data for SNPs on eight other candidate genes with our IL-3 results, to estimate relationships between pairs of loci and RA by maximum-likelihood analysis. We discuss here the utility of combining genotypic data in this way to identify possible contributions of various genes to this disease.

c. IgA nephropathy

Takashi Takei, Aritoshi Iida[2], Kosaku Nitta[1], Toshihiro Tanaka[3], Yozo Ohnishi[3], Ryo Yamada[4], Shiro Maeda[2], Tatsuhiko Tsunoda[5], Sachiyo Takeoka, Kyoko Ito, Kazuho Honda[1], Keiko Uchida[1], Ken Tsuchiya[1], Yasushi Suzuki[6], Tomoaki Fujioka[6], Takashi Ujiie[7], Yutaka Nagane[8], Satoru Miyano, Ichiei Narita[9], Fumitake Gejyo[9], Hiroshi Nihei[1], Yusuke Nakamura:[1]Department of Medicine, Kidney Center, Tokyo Women's Medical University;[2]Laboratory for Genotyping, SNP Research Center, The Institute of Physical and Chemical Research (RIKEN);[3]Laboratory for Cardiovascular Disease, SNP Research Center, The Institute of Physical and Chemical Research (RIKEN);[4]Laboratory for Rheumatic Disease, SNP Research Center, The Institute of Physical and Chemical Research (RIKEN);[5]Laboratory for Medical Informatics, SNP Research Center, The Institute of Physical and Chemical Research (RIKEN);[6]Department of Urology, Iwate Medical University;[7]Department of Urology, Iwate Prefectural Ofunato Hospital;[8]Department of Urology, Sanai Hospital;[9]Department of Medicine, Niigata University School of Medicine, 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. We designed a case-control association study on the basis of linkage disequibrium among single-nucleotide polymorphisms (SNPs) in the selectin gene cluster on chromosome 1q24-25, and found two SNPs in the E-selectin gene (SELE8, SELE13) and six in the L-selectin gene (SELL1, SELL4, SELL5, SELL6, SELL10 and SELL11) that were significantly associated with IgAN in Japanese patients. All eight SNPs were in almost complete linkage disequilibrium. SELE8 and SELL10 caused amino-acid substitutions from Tyr to His and from Pro to Ser (χ^2=9.02, *p*=0.0026, odds ratio=2.73 (95%CI; 1.38-5.38) and χ^2=17.4, *p*=0.000031, odds ratio=3.61 (95%CI; 1.91-6.83)), respectively, and SELL1 could affect promoter activity of the L-selectin gene (χ^2=19.5, *p*=0.000010, odds ratio=3.77 (95%CI; 2.02-7.05)); the TGT haplotype at these three loci was associated significantly with IgAN (χ^2=18.67, *p*=0.000016, odds ratio=1.88 (95%CI; 1.41-2.51)). Our results suggest that these eight SNPs in selectin genes may be useful for screening populations susceptible to the IgAN phenotype that involves interstitial infiltration.

Publications

A. Hirano, Y. Utada, S. Haga, T. Kajiwara, G. Sakamoto, F. Kasumi, Y. Nakamura, and M. Emi: Allelic losses as prognostic markers for breast cancers. Int J Clin Oncol 6:6-12, 2001

K. Kyo, T. Muto, H. Nagawa, G. M. Lathrop, and Y. Nakamura: Associations of distinct variants of the intestinal mucin gene MUC3A with ulcerative colitis and Crohn's disease. J. Human Genetics 46:5-20, 2001

H. Iwasa , M. Kurabayashi, R. Nagai, Y. Nakamura, and T. Tanaka: Multiple single-nucleotide polymorphisms (SNPs) in the Japanese population among six candidate genes for long QT syndrome. J Human Genetics 46:158-162, 2001

O. Kitahara, Y. Furukawa, T. Tanaka, C. Kihara, K. Ono, R. Yanagawa, E. M. Nita, H. Ogasawara, J. Okutsu, H. Zenbutsu, N. Shiraishi, T. Takagi, Y. Nakamura, and T. Tsunoda: Alterations of gene expression during colorectal carcinogenesis revealed by cDNA microarrays after laser-capture microdissection of tumor tissues and normal epithelia. Cancer Research 61:3544-3549, 2001

T. Itoh, K. Kikuchi, Y. Odagawa, S. Takata, K. Yano, S. Okada, N. Haneda, S. Ogawa, O. Nakano, Y. Kawahara, H. Kasai, T. Nakayama, T. Fukutomi, H. Sakurada, A. Shimizu, Y. Yazaki, R. Nagai, Y. Nakamura, and T. Tanaka: Correlation of genetic etiology with response to _-adrenergic blockade among symptomatic patients with familial long-QT syndrome. J. Human Genetics 46:38-40, 2001

S. Takeoka, M. Unoki, Y. Onouchi, S. Doi, H. Fujiwara, A. Miyatake, K. Fujita, Ituro Inoue, Y.

Nakamura, and M. Tamari: Amino-acid substitutions in the *IKAP* significantly increase a risk of childhood bronchial asthma. J. Human Genetics, 46:57-63, 2001

R. Yamada, T. Tanaka, M. Unoki T. Nagai, T. Sawada, Y. Ohnishi, T, Tsunoda, M. Yukioka, A. Maeda, K. Suzuki, H. Tateishi, T. Ochi, Y. Nakamura, and K. Yamamoto: Association between a single-nucleotide polymorphism in the promoter of the human interleukin-3 gene and rheumatoid arthritis in Japanese patients, and maximum-likelihood estimation of combinatorial effect that two genetic loci have on susceptibility to the disease. Am. J. Human Genetics 68:674-685, 2001

M. Matsushima-Nishiu, M. Unoki, K. Ono, T. Tsunoda, T. Minaguchi, H. Kuramoto, M. Nishida, T. Satoh, T. Tanaka, and Y. Nakamura: Microarray analysis of gene-expression profiles in endometrial cancer cells expressing exogenous PTEN. Cancer Research 61:3741-3749, 2001

T. Kato, S. Satoh, H. Okabe, O. Kitahara, K. Ono, C. Kihara, T. Tanaka, T. Tsunoda, Y. Yamaoka, Y. Nakamura, and Y. Furukawa: Isolation of a novel human gene, *MARKL1*, homologous to *MARK3* and its involvement in hepatocellular carcinogenesis. Neoplasia 3:4-9, 2001

C. Suzuki, M. Unoki, and Y. Nakamura: Identification and allelic frequencies of novel single-nucleotide polymorphisms in the DUSP1 and BTG1 genes. J Human Genetics 46:155-157, 2001

H. Okabe, S. Satoh, T. Kato, O. Kitahara, R. Yanagawa, Y. Yamaoka, T. Tsunoda, Y. Furukawa and Y. Nakamura: Genome-wide analysis of gene expression in human hepatocellular carcinomas. Cancer Research, 6:2129-2137, 2001

T. Watanabe, S. Okuno, T. Ono, Y. Yamasaki, K. Oga, A. Mizoguchi-Miyakita, H. Miyano, M. Suzuki, H. Momota, Y. Goto, H. Shinomiya, H. Hishigaki, I. Hayashi, T. Asai, S. Wakitani, T. Takagi, Y. Nakamura, and A. Tanigami: Single-allele correction of the Dmo1 locus in congenic animals substantially attenuates obesity, dyslipidaemia and diabetes phenotypes of the OLETF rat. Clin. and Exp. Pharmacology and Psysiology 28:28-42, 2001

A. Iida, A. Sekine, S. Saito, Y. Kitamura, T. Kitamoto, S. Osawa, C. Mishima and Y. Nakamura: Catalog of 320 single nucleotide polymorphisms (SNPs) in 20 quinone oxidoreductase and sulfotransferase genes. J. Human Genetics 46:225-240, 2001

O. Watanabe, M. Tamari, K. Natori, Y. Onouchi, Y. Shiomoto, I. Hiraoka and Y. Nakamura: Loci on murine chromosomes 7 and 13 that modify the phenotype of the NOA mouse, an animal model of atopic dermatitis. J. Human Genetics 46:221-4, 2001

A. Sekine, S. Saito, A. Iida, Y. Mitsunobu, S. Higuchi, S. Harigae, and Y. Nakamura: Identification of single-nucleotide polymorphisms (SNPs) of human N-acetyltransferase genes NAT1, NAT2, AANAT, ARD1 and L1CAM in the Japanese population. J. Human Genetics, 46:314-319, 2001

S. Saito, A. Iida, A. Sekine, C. Eguchi, Y. Miura, and Y. Nakamura: Seventy genetic variations in human microsomal and soluble epoxide hydrolase genes (EPHX1 and EPHX2) in the Japanese population J. Human Genetics 46:325-329, 2001

K. Kobayashi, J. Sasaki, E. Kondo-Iida, Y. Fukuda, M. Kinoshita, Y. Sunada, Y. Nakamura, and T. Toda: Structural organization, complete genomic sequences and mutational analyses of the Fukuyama-type congenital muscular dystrophy gene, fukutin. FEBS Letters 489:192-196, 2001

A. Hirano, M. Emi, M. Tsuneizumi, Y. Ueda, M. Yoshimoto, F. Kasumi, F. Akiyama, G. Sakamoto, S. Haga, T. Kajiwara, and Y. Nakamura: Allelic losses of loci at 3p25.1, 8p22, 13q12, 17p13.3 and 22q13 correlate with postoperative recurrence in breast cancer. Clinical Cancer Research 7:876-882, 2001

Y. Nakamura: Isolation of disease-associated genes through genome analysis and their clinical application. The Keio Journal of Medicine 50:13-140, 2001

A. Iida, S. Saito, A. Sekine, T. Kitamoto, Y. Kitamura, C. Mishima, S. Osawa, K. Kondo, S. Harigae, and Y. Nakamura: Catalog of 434 single nucleotide polymorphisms (SNPs) in genes of the alcohol dehydrogenase, glutathione S-transferase, and NADH ubiquinone oxidoreductase families. J Human Genetics 46:385-407, 2001

H. Ishiguro, T. Tsunoda, T. Tanaka, Y. Fujii, Y. Nakamura, and Y. Furukawa: Identification of *AXUD1*, a novel human gene induced by *AXIN1* and its reduced expression in human carcinomas of the lung, liver, colon and kidney. Oncogene 20:5062-5066, 2001

M. Unoki and Y. Nakamura: Growth-suppressive effects of *BPOZ* and *EGR2*, two genes involved in the PTEN signaling pathway. Oncogene 20:4457-4465, 2001

C. Kihara, T. Tsunoda, T. Tanaka, H. Yamana, Y. Furukawa, K. Ono, O. Kitahara, H. Zenbutsu, R. Yanagawa, K. Hirata, T. Takagi, and Y. Nakamura: Prediction of sensitivity of esophageal tumors to adjuvant chemotherapy by cDNA microarray analysis of gene-expression profiles. Cancer Research, 61:6474-6479, 2001

Y. Furukawa, T. Kawasoe, Y. Daigo, T. Nishiwaki, H. Ishiguro, M. Takahashi, J. Kitayama, and Y. Nakamura: Isolation of *ARHGAP9*, a novel member of the Rho-GAP gene family, that regulates integrin-mediated adhesion of hematopoietic cells to extracellular matrix. B. B. R. C. 284:643-649, 2001

Y.-M. Lin, K. Ono, S. Satoh, H. Ishiguro, M. Fujita, N. Miwa, T. Tanaka, T. Tsunoda, Y. Nakamura, and Y. Furukawa: Identification of AF17 as a downstream gene of the β-catenin/Tcf pathway, and its involvement in colorectal carcinogenesis. Cancer Research, 61:6345-6349, 2001

M. Furuhashi, K. Yagi, H. Yamamoto, Y. Furukawa, S. Shimada, Y. Nakamura, A. Kikuchi, K. Miyazono, and M. Kato: Axin facilitates Smad3 activation in the transforming growth factor-β signaling pathway. Mol. Cell. Biology, 21:5132-5141, 2001

R. Yanagawa, Y. Furukawa, T. Tsunoda, O. Kitahara, K. Murata, O. Ishikawa, and Y. Nakamura: Genome-wide screening of genes showing altered expression in liver metastases of human colorectal cancers by cDNA microarray. Neoplasia 3:395-401, 2001

Y. Ohnishi, T. Tanaka, K. Ozaki, R. Yamada, H. Suzuki and Y. Nakamura: A high-throughput SNP typing system for genome-wide association studies. Journal of Human Genetics 46:471-477, 2001

H. Iwasa, M. Kurabayashi, R. Nagai, Y. Nakamura, T. Tanaka: Genetic variations in five genes involved in the excitement of cardiomyocytes. Journal of Human Genetics 46:549-552, 2001

S. Saito, A. Iida, A. Sekine, Y. Miura, T. Sakamoto, C. Ogawa, S. Kawauchi, S. Higuchi, and Y. Nakamura: Identification of 197 genetic variations in six human methyltransferase genes in the Japanese population. Journal of Human Genetics 46:529-537, 2001

S. Okamura, H. Arakawa, T. Tanaka, H. Nakanishi, C. C. Ng, Y. Taya, M. Monden, and Y. Nakamura: p53DINP1, a p53-inducible gene, regulates p53-dependent apoptosis. Molecular Cell, 8:85-94, 2001

A. Iida, S. Saito, A. Sekine, Y. Kitamura, K. Kondo, C. Mishima, S. Osawa, S. Harigae, and Y. Nakamura: High-density SNP (single-nucleotide polymorphism) map of the 150-kb region corresponding to the human ATP binding cassette transporter A1 (ABCA1) gene. Journal of Human Genetics 46:522-528, 2001

M. Fujita, Y. Furukawa, T. Tsunoda, T. Tanaka, M. Ogawa and Y. Nakamura: Up-regulation of the *ENC1* (ectodermal-neural cortex 1) gene, a downstream target of the β-catenin/Tcf complex, in colorectal carcinomas. Cancer Research 61:7722-7726, 2001

Y. Suzuki, H. Taira, T. Tsunoda, J. Mizushima-Sugano, J. Sese, H. Hata, T. Ota, T. Isogai, T. Tanaka, S. Morishita, K. Okubo, Y. Sakaki, Y. Nakamura, A. Suyama, and S. Sugano: Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. EMBO reports 2:388-393, 2001

S. Okuno, T. K. Watanabe, T. Ono, K. Oga, A. Mizoguchi-Miyakita, Y. Yamasaki, Y. Goto, H. Shinomiya, H. Momota, H. Miyano, I. Hayashi, T. Asai, M. Suzuki, Y. Harada, H. Hishigaki, S. Wakitani, T. Takagi, Y. Nakamura, and A. Tanigami: Dffects of Dmo1 on obesity, dyslipiaemia and hyperglycemia in the Otsuka Long Evans Tokushima Fatty strain. Genet. Res., Comb 77:183-190, 2001

Y. Suzuki, T. Tsunoda, J. Sese, H. Taira, J. Mizushima-Sugano, H. Hata, T. ota, T. Isogai, T. tanaka, Y. Nakamura, A. Suyama, Y. Sakaki, S. Morishita, K. Okubo, and S. Sugano: Identification and characterization of the potential promoter regions of 1031 kinds of human genes. Genome Research 11:677-684, 2001

K. Yasui, I. Imoto, Y. Fukuda, A. Pimkhaokham, Z.Q. Yang, T. Naruto, Y. Shimada, Y. Nakamura, J. Inazawa: Identification of target genes within an amplicon at 14q12-q13 in esophageal squamous cell carcinoma. Genes Chromosomes Cancer 32:112-118, 2001

C. Sakakura, A. Hagiwara, R. Yasuoka, Y. Fujita, M. Nakanishi, K. Masuda, K. Shimomura, Y. Nakamura, J. Inazawa, T. Abe, H. Yamagishi: Tumour-amplified kinase BTAK is amplified and overexpressed in gastric cancers with possible involvement in aneuploid formation. Br J Cancer 84:824-31, 2001

A. Iida, S. Saito, A. Sekine, S. Harigae, S. Osawa, C. Mishima, K. Kondo, Y. Kitamura, and Y. Nakamura: Catalog of 46 single nucleotide polymorphisms (SNPs) in the microsomal glutathione S-transferase 1 (MGST1) gene. J Human Genet 46:590-591, 2001

A. Iida, Y. Ohnishi, K. Ozaki, Y. Ariji, A. Inose, M. Takahashi, Y. Nakamura, and T. Tanaka: High-density single-nucleotide polymorphism (SNP) map in the 96-kb region containing the entire human DiGeorge syndrome critical region 2 (DGCR2) gene at 22q11.2. J Human Genet 46:604-608, 2001

Y. Yamanaka, M. Tamar, T. Nakahata and Y. Nakamura: Gene-expression profiles of human small airway epithelial cells treated with low doses of 14- and 16-membered macrolides. B.B.R.C., 287:198-203, 2001

A. Iida, S. Saito, A. Sekine, C. Mishima, K. Kondo, Y. Kitamura, S. Harigae, S. Osawa, and Y. Nakamura: Catalog of 258 single nucleotide polymorphisms (SNPs) in genes encoding three organic anion transporters, three organic anion-transporting polypeptides, and three NADH:ubiquinone oxidoreductase flavoproteins. Journal of Human Genetics 46:668-683, 2001

M. Murata,Y. Miyoshi, M. Ohshima, K. Shibata, T. Ohta, Y. Imai, M. Nishikawa, K. Iwao, H. Tateishi, T. Shimano, T. Kobayashi, and Y. Nakamura: Accumulation of β-catenin in the cytoplasm and the nuclei during the early hepatic tumorigenesis. Hepatology Research 21:126-135, 2001

M. Murata,Y. Miyoshi, K. Iwao, H. Wada, K. Shibata, H. Tateishi, T. Shimano, M. Ohsawa, Y. Imai, M. Nishikawa, T. Kobayashi, and Y. Nakamura:

Combined hepatocellular/cholangiocellular carcinoma with sarcomatoid features: genetic analysis for histogenesis. Hepatology Research 21:220-227, 2001

N. Miwa, M. Furuse, S. Tsukita, N. Niikawa, Y. Nakamura and Y. Furukawa: Involvement of Claudin-1 in the β-catenin /Tcf signaling pathway and its frequent up-regulation in human colorectal cancers. Onclogy Res., in press

T. Yamaguchi, K. Matsuda, Y. Sagiya, M. Iwadate, M. A. Fujino, Y. Nakamura, and H. Arakawa p53R2-dependent pathway for DNA synthesis in a p53-regulated cell-cycle checkpoint. Cancer Research 61:8256-8262, 2001

S. Haga, M. Emi, A. Hirano, Y. Utada, T. Kajiwara, F. Akiyama, G. Sakamoto, K. Takahashi, T. Tada, F. Kasumi, and Y. Nakamura: Association of allelic losses at 3p25.1, 13q12 or 17p13.3 woth poor prognosis in breast cancers with lymphnode metastasis. Jpn J Can Research 92:1199-1206 2001

M. Tsuneizumi, M. Emi, H. Nagai, H. Harada, G. Sakamoto, F. Kasumi,S. Inoue, T. Kazui and Y. Nakamura: Overexpression of the EBAG9 gene at 8q23 associated with early-stage breast cancer. Clinical Cancer Research 7:3526-3532, 2001

S. Iwashita, K. Koyama, and Y. Nakamura: VNTR sequence on human chromosome 11p15 that affects transcriptional activity. J Human Genetics in press

T. Mori, Y. Nakamura, and H. Arakawa: Identification of the interferon regulatory factor 5 gene (*IRF5*) as a direct target for p53. Oncogene in press

K. Ochi, Y. Nakamura and H. Arakawa: Identification of *semaphorin3B* as a direct target of p53. Neoplasia, in press

K. Matsuda, K. Yoshida, S. Onoue, K. Nakamura, Y. Nakamura and H. Arakawa: Interplay of p53AIP1 and bcl-2 in the mitochondrial apoptotic pathway. Cancer Research, in press

K. Matsuda, H. Arakawa and Y. Nakamura: Adenovirus-mediated *PTEN* gene therapy for endometrial cancers. Gene Therapy, in press

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 ABC1 (activity of bc1 complex), Cancer Research, in press

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, in press

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, in press

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, in press

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 anti-cancer drugs. Cancer Research, in press

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, in press

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, in press

T. Takei, A. Iida, K. Nitta, T. Tanaka, Y. Ohnishi, R. Yamada, S. Maeda, T. Tsunoda, S. Takeoka, K. Ito, 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, in press

M. Doi, M. Nagano and Y. Nakamura: Genome-wide screening by cDNA microarray of genes associated with matrix mineralization by human mesenchymal stem cells *in vitro*. B. B. R. C., 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. 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, in press

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, in press

Human Genome Center

Laboratory of Sequence Analysis

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. HGREP: Human Genome REconstruction Project

Tetsushi YADA, Yasushi TOTOKI, Yoshiyuki SAKAKI and Toshihisa TAKAGI

On February 2001, working draft sequences which are estimated to cover 90% of the human genome were released, and the time was ripe for tracing the outline of the genome and for finding genes. However, since these sequences are too fragmentary, we cannot use them in raw format for the above purposes. Therefore, we launched a new project named HGREP (Human Genome REconstruction Project). In HGREP, (1) we sort the draft sequences by chromosome and assemble them, and (2) we make annotations such as genes and repeats on the sequences. These results are distributed via internet (http://hgrep.ims.u-tokyo.ac.jp/) and are continuously updated.

Although other research groups are doing similar attempts, such as Ensembl (http://www.ensembl.org/), HGREP is unique in the following points: (1) Ensembl assembles the draft sequences based on DNA fingerprint data, while HGREP assembles them based on sequence similarity; Therefore, HGREP realizes the reliable reconstruction of human chromosomes; (2) HGREP annotates genes by using a novel gene finding program named DIGIT; DIGIT enables us to predict exact gene structures with 10% higher accuracy compared to existing gene finding programs.

HGREP is a joint project between the Laboratory of Genome Database (Human Genome Center, IMS, the University of Tokyo) and the Human Genome Research Group (Genomic Sciences Center, RIKEN).

2. DIGIT: Digit Integrates Gene Identification Tools

Tetsushi YADA, Yasushi TOTOKI, Yoshio TAKAEDA, Yoshiyuki SAKAKI and 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 (Digit Integrates Gene Identification Tools).

An outline of the algorithm is as follows. First, existing gene-finders are applied to an uncharacterized genome sequence (input sequence). Next, DIGIT produces all possible exons from the results of gene-finders, and assigns them their reading frames and scores. Finally, DIGIT searches a set of exons whose additive score is maximized under their reading frame constraints. Bayesian procedure and hidden Markov model are used to infer scores and search exon set, respectively.

We have designed DIGIT so as to combine FGENESH, GENSCAN and HMMgene, and have assessed its prediction accuracy by using recently compiled data sets. For all data sets, it has been revealed that DIGIT successfully discarded many false positive exons predicted by 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.

3. Quadtree Representation of DNA Sequences

Natsuhiro ICHINOSE, Tetsushi YADA and Toshihisa TAKAGI

Since we can access large-size genome sequences today, it becomes possible to analyze them not only by a deductive approach but also by an inductive approach. In the deductive approach, we first assume a hypothesis of biological dynamics, then we verify whether or not it can explain the sequence data. On the other hand, we try to extract the biological information from the sequence data heuristically by observing them directly in the inductive approach.

In order to realize a tool for the inductive approach, we study a method of quadtree representation. The quadtree representation is a visualization tool of oligonucleotide frequencies, which provides us a method to find qualitative characteristics in sequence data heuristically.

In the quadtree representation, the oligonucletide frequencies are arranged in a regular square hierarchically. Then the position of each oligonucletide sequence is topologically preserved for the measure space of its distance. Therefore we can obtain not only a property of a single oligonucleotide sequence but also the relation between sequences whose characteristics are represented in the measure space. In this work, we developed the method by which we can extract the characteristics of longer oligonucleotide sequence, by using the Nth-order Markov model as the expectation value. We showed the following results for the human chromosome 21: A type of the dinucleotide repeats with mutations, which is expressed as ([AG][CT])+ in regular expressions, is over-represented; Another type ([AC][GT])+, however, is not over-represented. This is because the mutations have the transition-transversion skewness. It is interesting that the characteristics of mutations can be observed in a single sequence.

As future works, we will develop the statistical comparison method between a pair of distinct sequences such as those of distinct species, distinct sequence regions (coding region and promoter region) and so on. We will also introduce statistical tests into our method.

4. A Design Method 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 it to genetic engineering such as gene therapy, we study a method to design model gene networks.

Because gene networks should be directly controlled in the gene therapy, it is important to understand their characteristics. Furthermore, we think that it is necessary that the gene networks are controlled by themselves, because the object of the therapy is tiny cells and the autonomous controls are required. Therefore our purpose in this work is to develop the method to design model gene networks from given gene expression data in order to obtain the fundamental knowledge of such genetic engineering.

Using continuous-time switching networks as the model, we showed that the networks can be constructed from binary switching data corresponding to gene expressions in continuous time by using the optimization methods. We also showed that even if there is a known partial structure in the networks, the construction can be done with preserving its structure.

Publications

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

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.

N. Ichinose, T. Yada and T. Takagi. Quadtree Representation of DNA Sequences, Proceedings of the 12th International Conference on Genome Informatics, 2001, in press.

N. Ichinose and K. Aihara. A Design Method of Model Gene Networks, Proceedings of International Symposium on Artificial Life and Robotics

AROB7th, 2002, in press.

矢田哲士：ヒト遺伝子発見プログラム DIGIT．蛋白質・核酸・酵素 , 46:2580-2585, 2001.

矢田哲士：ヒトゲノム再構築プロジェクト: HGREP．ゲノム医学 , 1:57-63, 2001.

矢田哲士：ヒトゲノムドラフト配列データの情報処理．情報処理 , 42:600-605, 2001.

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.

市瀬夏洋，小林徹也，合原一幸：遺伝子相互作用系の非線形数理解析．蛋白質・核酸・酵素 , 46:2598-2602, 2001.

Human Genome Center

Laboratory of Functional Genomics

We are interested in sequence-based analysis of human genome, functional analysis of the yeast genome, molecular mechanism regulating mammalian circadian rhythms, and the hunting of genes with unique expression patterns.

1. Sequence analysis of human chromosome 21

Kunihiko Takamatsu, Kouhei Maekawa, Tadayuki Takeda[1], Masahira Hattori[1], Todd Taylor[1] and Yoshiyuki Sakaki:[1]RIKEN, Genomic Sciences Center, Yokohama

Comprehensive knowledge of the gene content of human Chromosome 21 (HSA21) is essential for understanding the etiology of Down syndrome (DS). We have done the largest comparison of finished mouse and human sequence to date for a region of 1.35 Mb mouse Chromosome 16 (MMU16) that is conserved with human Chr 21q22.2. This is corresponds to a portion of a commonly described "DS critical region," thought to contain a gene or genes whose dosage imbalance contributes to a number of DS phenotypes. We utilized comparative sequence analysis to construct a "DNA features" map of this region which includes 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 provides the first annotation indicating 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*, do not have counterparts at the corresponding positions on MMU16 nor elsewhere in the mouse genome. Our findings have implications for evolution and for modeling the genetic basis of DS in mice.

2. Identification of methylation imprints on human chromosome 21 through a comprehensive analysis of CpG islands

Yoichi Yamada, Hidemi Watanabe[1], Takashi Ito[2], Yoshiyuki Sakaki:[1]RIKEN, GSC;[2]Cancer Research Institute, Kanazawa Univ.

Imprinted genes are often associated with DNA regions subject to allele-specific methylation, termed differentially methylated region (DMR) or methylation imprints, which often share structural features such as tandemly repeats and CpG island-like base composition. Because CpG islands generally lie near promoter regions and escape methylation, monoallelically methylated DMRs and the islands on inactivated X chromosome in female stand for exceptions: cells thus bear both methylated and unmethylated alleles for these islands and hence display composite pattern upon methylation analysis. It is thus conceivable that screening for CpG islands with composite methylation pattern serves as a novel method to identify DMRs and associated imprinted genes. To screen for such islands, we developed a novel HpaII-McrBC PCR method by exploiting the complementary nature of HpaII, which cuts the unmethylated DNA, and McrBC, which digests the methylated one. We applied the method to a comprehensive methylation analysis of 146 CpG islands on human chromosome 21. While most islands escape methylation as expected, 21 and 15 CpG islands display complete and composite methylation patterns, respectively. Intriguingly, a positive correlation was observed between these methylated islands and the repeated structure. From the 15 islands showing composite methylation pattern, we have so far confirmed that 3 islands are indeed subject to allele-specific methylation. Furthermore, for one of such islands, we proved that the methylated allele is derived from the maternal lineage using a SNP and an informative pedigree. These DMRs may indicate the presence of here-to-fore unidentified or unpredicted imprinted genes on this chromosome. This approach would provide a novel way to identify DMRs or methylation imprints and hence novel imprinted genes in the human genome.

3. Toward a quantitiative analysis of gene expression networks in the budding yeast

Fumihito Miura, Miyuki Onda, Kazuhisa Ota[1], Takashi Ito[1], and Yoshiyuki Sakaki:[1]Cancer Research Institute, Kanazawa Univ.

For the systematic analysis of gene regulatory networks, we developed a PCR-based knock-in strategy to make a transcription factor constitutively active by replacing its activation domain with that of VP16. This strategy would facilitate the search of downstream target genes even in the absence of activating signals from upstream, which cannot be known a priori for the novel transcription factors revealed for the first time by genomic sequencing. The power of the strategy was demonstrated in the study of PDR1 regulating multidrug resistance in the budding yeast. Furthermore, we applied the strategy to the analysis of uncharacterized transcription factors YRR1 and YOR172W to reveal their unique participation in multidrug resistance. On the other hand, we are developing a unique adapter-tagged competitive PCR using genomic DNA as a standard for the absolutely quantitative description of the transcriptome. Integration of these strategies with chromatin immuno-precipitation technique would allow us to collect data on the “input” (*i.e.* stauts of binding of transcription factors to the promoter *in vivo*) and “output” (*i.e.* transcripts) of gene expression in a genome-wide scale. Such a dataset would help us understand the logic of gene expression network in a quantitative manner.

4. Screening for imprinted genes by Allelic Message Display

Yuriko Hagiwara, 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.

5. Mutation analysis of the Genes Encoding Dystrophin Associated Protein Complex in the Pathogenesis of Human Idiopathic Cardiomyopathy

Yuepeg Wang, Hisashi Hagiwara, Yuko Mitani, Yoshiyuki Sakaki

Idiopathic cardiomyopathy, one of the common heart diseases resulting in heart failure and sudden death, is divided into hypertrophic (HCM) and dilated (DCM) cardiomyopathy. The genetic abnormalities would account for about 50~70% of the diseases. Although both diseases are genetically heterogenerous, the most common abnormalities have been reported to involve mutations in the genes encoding dystrophin associated protein complex and sarcomere proteins. Dystrophin associated protein complex in the cardiac muscle consists of alpha- and beta-dystroglycans, alpha-, beta-, gamma- and delta-sarcoglycans, and dyntrophin. The complex is believed to play a key role in maintaining the normal architecture of the muscle sarcolemma by constituting a link between the subsarcolemmal cytoskeleton and the extracellular matrix. Previous studies have documented a disruption in the integrity of the complex in the skeletal and cardiac muscle in the Syrian hamster, a model animal of HCM and DCM. Recently, mutations in this complex have been reported to be the etiology of cardiomyopathies in the model animals. But in human, the mutation and the role in cardiomyopathies remain to be elucidated. Therefore, we aimed to explore whether these genes are mutated and are involved in the pathogenesis of HCM and/or DCM in Japanese patients. We have determined the genome structure of one of these genes. Now we are accumulating the DNA samples from patients and normal controls and are screening mutations in the genes in both encoding and promoter regions. By functional analysis of promoter activity and protein interactions, the role of positive mutations in the pathogenesis of human cardiomyopathies will be clarified.

6. Molecular mechanisms regulating mammalian

circadian clock

Hajime Tei, Akiko Hida, Rika Numano, Nobuya Koike, Shihoko Kojima, Yoko Sato, Satomi Shiozuka, 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

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

Levanon, D., Glusman, G., Bangsow, T., Ben-Asher, E., Male, D.A., Avidan, N., Bangsow, C., Hattori, M., Taylor, T.D., Taudien, S., Blechschmidt, K., Shimizu, N., Rosenthal, A., Sakaki, Y., Lancet, D. and Groner, Y.: Architecture and anatomy of the genomic locus encoding the human leukemia-associated transcription factor RUNX1/AML1. GENE 262: 23-33 (2001).

Uematsu, C., Nishida, J., Okano, K., Miura, F., Ito, T., Sakaki, Y. & Kambara, H.: Multiplex polymerase chain reaction (PCR) with color-tagged module-shuffling primers for comparing gene expression levels in various cells. Nucleic Acids Res. **29**, e84 (2001).

Miura, F., Yada, T., Nakaki, K., Sakaki, Y. & Ito, T.: Differential diaplay analysis of mutants for the transcription factor Pdr1p regulating multidrug resistance in the budding yeast. FEBS Lett. **505**: 103-108 (2001).

Shiose, A., Kuroda, J., Tsuruya, K., Hirai, M., Hirakata, H., Naito, S., Hattori, M., Sakaki, Y, and Sumimoto, H.: A Novel Superoxide-producing NAD(P)H Oxidase in Kidney. J. Biol. Chem. 276: 1417-1423 (2001).

Wang, Y., Chen, J., Wang, Y., Taylor, C.W., Hirata, Y., Hagiwara, H., Mikoshiba, K., Toyo-Oka, T., Omata, M. and Sakaki, Y.: Crucial Role of Type 1, but Not Type 3, Inositol 1,4,5-Trisphosphate (IP(3)) Receptors in IP(3)-Induced Ca(2+) Release, Capacitative Ca(2+) Entry, and Proliferation of A7r5 Vascular Smooth Muscle Cells. Circ. Res. 88: 202-209 (2001).

Yutaka Suzuki, Hirotoshi Taira, Tatsuhiko Tsunoda, Junko Mizushima-Sugano, Jun Sese, Hiroko Hata,Toshio Ota,Takao Isogai, Toshihiro Tanaka, Shinichi Morishita, Kousaku Okubo, Yoshiyuki Sakaki, Yusuke Nakamura, Akira Suyama and Sumio Sugano: Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. EMBO reports vol.2, No.5: 388-393(2001).

Dong-Kug Choi, Yutaka Suzuki, Shinichiro Yoshimura, Takushi Togashi, Munetomo Hida, Todd D. Taylor, Yuepeng Wang, Sumio Sugano, Masahira Hattori, Yoshiyuki Sakaki: Molecular cloning and characterization of a gene expressed in mouse developing tongue, *mDscr5* gene, a homolog of human DSCR5 (Down syndrome Critical Region gene 5). Mammalian Genome 12: 347-351(2001).

Yutaka Suzuki, Tatsuhiko Tsunoda, Jun Sese, Hirotoshi Taira, Junko Mizushima-Sugano, Hiroko Hata, Tosio Ota, Takao Isogai, Toshihiro Tanaka, Yusuke Nakamura, Akira Suyama, Yoshiyuki Sakaki, Shinichi Morishita, Kousaku Okubo, and Sumio Sugano: Identification and Characterization of the Potential Promoter Regions of 1031 Kinds of Human Genes. Genome Research Vol.11, No.5: 677-684 (2001).

Ayako Yamamoto, Takashi Suzuki, Yoshiyuki Sakaki: Isolation of hNap1BP which interacts with human Nap1(NCKAP1) whose expression is down-regulated in Alzheimer's disease. Gene 271: 159-169 (2001).

Yaeko Ichikawa, Jun Goto, Masahiro Hattori, Atsushi Toyoda, Kazuo Ishii, Seon-Yong Jeong, Hideji Hashida, Naoki Masuda, Katsuhisa Ogata, Fumio Kasai ,Momoki Hirai, Patoricia Maciel, Guy A. Rouleau, Yoshiyuki Sakaki, Ichiro kanazawa: The genomic structure and expression of MJD, the Macado-Joseph disease gene. J Hum Genet46:413-422 (2001)

Tomoharu Osada, Gen Watanabe, Yoshiyuki Sakaki, and Takashi Takeuchi: Puromycin-Sensitive Aminopeptidase Is Essential for the Maternal Recognition of Pregnancy in Mice. Molecular Endocrinology 15(6): 882-893 (2001).

Tomoharu Osada, Gen Watanabe, Shunzo Kondo, Masashi Toyoda, Yoshiyuki Sakaki and Takashi Takeuchi: Male Reproductive Defects Caused by Puromycin-Sensitive Aminopeptidase Deficiency in Mice: Molecular Endocrinology 15(6): 960-971 (2001).

Yuzo Nakagawa-Yagi, Dong-Kug Choi, Nobuo Ogane, Shin-ichi Shimada, Motohide Seya, Takashi Momoi, Takashi Ito, Yoshiyuki Sakaki: Discovery of a novel compound: insight into mechanisms for acrylamide-induced axonopathy and colchicine-induced apoptoic neuronal cell death. Brain Research 909: 8-19 (2001).

Yoshihisa Watanabe, Asao Fujiyama, Yuta Ichiba, Masahira Hattori, Tetsushi Yada, Yoshiyuki Sakaki and Toshimichi Ikemura: Chromosome-wide assessment of replication timing for human chromosomes 11q and 21q: disease-related genes in timing-switch regions. Human Molecular Genetics Vol.11, No.1: 13-21(2002).

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

Kubota, H., Ota, K., Sakaki, Y. & Ito, T. Budding yeast GCN1 binds the GI domain to activate the eIF2α kinase GCN2. J. Biol. Chem. 276: 17591-17596 (2001).

Miura, F., Yada, T., Nakaki, K., Sakaki, Y. & Ito, T. Differential diaplay analysis of mutants for the transcription factor Pdr1p regulating multidrug resistance in the budding yeast. FEBS Lett. 505: 103-108 (2001).

Stokkan, K-A., Yamazaki, S., Tei, H., Sakaki, Y. and Menaker, M. Entrainment of the circadian clock in the liver by feeling. Science 291: 490-493 (2001).