Human Genome Center

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

Professor	Minoru Kanehisa, Ph.D.	教　授(委嘱)	理学博士	金　久　　實
Research Associate	Toshiaki Katayama, M.Sc.	助　手	理学修士	片　山　俊　明
Research Associate	Shuichi Kawashima, M.Sc.	助　手	理学修士	川　島　秀　一
Lecturer	Tetsuo Shibuya, Ph.D.	講　師	理学博士	渋　谷　哲　朗
Research Associate	Michihiro Araki, Ph.D.	助　手	薬学博士	荒　木　通　啓

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

1. Comprehensive repository for community genome annotation

Toshiaki Katayama, Mari Watanabe and Minoru Kanehisa

KEGG DAS is an advanced genome database system providing DAS (Distributed Annotation System) service for all organisms in the GENOME and GENES databases in KEGG (Kyoto Encyclopedia of Genes and Genomes). Currently, KEGG DAS contains 6,943,951 annotations for genome sequences of 441 organisms. The KEGG DAS server provides gene annotations linked to the KEGG PATHWAY and LIGAND databases, as well as the SSDB database containing paralog, ortholog and motif information. In addition to the coding genes, information of non-coding RNAs predicted using Rfam database is also provided to fill the annotation of the intergenic regions of the genome. We have been developing the server based on open source software including BioRuby, BioPerl, BioDAS and GMOD/GBrowse to make the system consistent with the existing open standards. The contents of the KEGG DAS database can be accessed graphically in a web browser using GBrowse GUI (graphical user interface) and also programatically by the DAS protocol. The DAS, which is an XML over HTTP data retrieving protocol, enables the user to write various kinds of automated programs for analyzing genome sequences and annotations. For example, by combining KEGG DAS with KEGG API, a program to retrieve upstream sequences of a given set of genes which have similar expression patterns on the same pathway, can be written very easily. GBrowse, the graphical interface, enables user to browse, search, zoom and visualize a particular region of the genome. Moreover, users are also able to add their own annotations onto the GBrowse

view by providing another DAS server or by simply uploading their own data as a file. This functionality enables researchers to add various annotations on the genome and by sharing their annotations with the community they can continuously refine the genome annotation, so-called "community annotation." This year we have updated the GBrowse genome browser to the latest version, which has improved user interface. We are also responsible to Japanese localization of the browser. The KEGG DAS is weekly updated and freely available at http://das.hgc.jp/.

2. Automatic assignments of orthologs and paralogs in complete genomes

Toshiaki Katayama and Minoru Kanehisa

Accession of the number of sequenced genomes made it difficult to characterize orthologous relationship among organisms. We are developing a computational method for finding appropriate orthologous gene clusters automatically. It is based on a graph analysis of the KEGG SSDB database, containing sequence similarity relations among all the genes in the completely sequenced genomes. The nodes of the SSDB graph are genes and the edges are the Smith-Waterman sequence similarity scores computed by the SSEARCH program. The edges are not only weighted but also directed, indicating the best (top-scoring) hit when a gene in an organism is compared against all genes in another organism. Thus, a highly connected cluster of nodes containing a number of bidirectional best hits might be considered an ortholog cluster consisting of functionally identical genes. Such a cluster can be found by our heuristic method for finding quasi-cliques, but the SSDB graph is too large to perform quasi-clique finding at a time. Therefore, we introduce a hierarchy (evolutionary relationship) of organisms and treat the SSDB graph as a nested graph. However, the method still requires large computation time along with the number of organism increases, we are trying to refine the process to make it faster and accurate.

3. SOAP/WSDL interface for the KEGG system

Toshiaki Katayama, Shuichi Kawashima and Minoru Kanehisa

We have continued to develop KEGG API, a web service to facilitate usability of the KEGG system. KEGG is a suite of databases and associated software, integrating our current knowledge of molecular interaction/reaction pathways and other systemic functions (PATHWAY and BRITE databases), the information about the genomic space (GENES database), and information about the chemical space (LIGAND database). KEGG API provides valuable means to retrieve various kinds of information stored in the KEGG and has become an increasingly popular mode of access. Recently, we have introduced several new methods to search compounds, drugs and glycans by their structure, mass, composition and annotations. Additionally, the methods to colorize the PATHWAY diagram is enhanced to control the different elements sharing the same name can be distinguished by their ID. The KEGG API is available at http://www.genome.jp/kegg/soap/.

4. EGassembler: web server for large-scale clustering and assembling ESTs and genomic DNA fragments

Ali Masoudi-Nejad, Shuichi Kawashima, Koichiro Tonomura, Masanori Suzuki, Minoru Kanehisa

EST sequencing has proven to be an economically feasible alternative for gene discovery in species lacking a draft genome sequence. Ongoing large-scale EST sequencing projects feel the need for bioinformatics tools to facilitate uniform ESTs handling. This brings about a renewed importance to a universal tool for processing and functional annotation of large sets of ESTs in order to cover the complete transcriptome of an organism. EGassembler (http://egassembler.hgc.jp/) is a web server, which provides an automated as well as a user-customized analysis tool for cleaning, repeat masking, vector trimming, organelle masking, clustering and assembling of ESTs and genomic fragments. It is also designed to serve as a standalone web application for each one of those processes. The web server is freely available and provides the community with a unique all-in-one online application web service for large scale ESTs and genomic DNA clustering and assembling, especially for EST processing and annotation projects.

5. New version of MAGEST database

Shuichi Kawashima and Minoru Kanehisa

We have developed a new version of MAGEST database. Previously MAGEST was designed as a database for maternal gene expression information for an ascidian, *Halocynthia roretzi*. In the new version of MAGEST, it is ex-

tended to four embryonic developmental stages including the following additional stages, e.g. early cleavage, early gastrula and early neurula. Furthermore, we constructed gene clusters and assembled sequences of ESTs by EGassembler. These clusters enabled us to cross-refer the gene expression of the four different stages. The new web site is accessible at http://magest.hgc.jp/. Now we are comparing the MAGEST sequences with other species of ascidian, *Ciona* sp. Because the phylogenetic position of *H. roretzi* is a good outgroup for Enterogonia to which *Ciona* sp. belongs, we expect that the result from this comparative analysis lead us to understand the diversification of gene families among Urochodata.

6. SSS: a sequence similarity search service

Toshiaki Katayama, Kazuhiro Ohi, Minoru Kanehisa

There are various services in the world to find similar sequences from the database, such as the famous BLAST service provided at NCBI. However, the method to search and the database to be searched could not be added from outside. To provide our super computer resources at the Human Genome Center to the research community, we started to develop a new service for the sequence similarity search, SSS. In SSS, user can select the search algorithm from BLAST, FASTA, SSEARCH, TRANS and EXONERATE. This variety of options is unique among the public services. Then user is prompted to select appropriate database depending on the algorithm selected and the search is executed. On the backend, we implemented the search system on the Sun Grid Engine to provide efficient resources on distributed computers. As a result, we are able to provide time consuming services such as TRANS and EXONERATE in addition to the popular algorithms. The SSS service is freely available at http://sss.hgc.jp/.

7. High performance database entry retrieval system

Kazutomo Ushijima, Chiharu Kawagoe, Toshiaki Katayama, Shuichi Kawashima, Kenta Nakai, Minoru Kanehisa

Recently, the number of entries in biological databases is exponentially increasing year by year. For example, there were 10,106,023 entries in the GenBank database in the year 2000, which has now grown to 65,771,589 (Release 156 + daily updates). In order for such a vast amount of data to be searched at a high speed, we have developed a high performance database entry retrieval system, named HiGet. The HiGet system is constructed on the HiRDB, a commercial ORDBMS (Object-oriented Relational Database Management System) developed by Hitachi, Ltd. It is publicly accessible on the Web page at http://higet.hgc.jp/ or SOAP based web service at http://higet.hgc.jp/soap/. HiGet can execute full text search on various biological databases. In addition to the original plain format, the system contains data in the XML format in order to provide a field specific search facility. When a complicated search condition is issued to the system, the search processing is executed efficiently by combining several types of indices to reduce the number of records to be processed within the system. Current searchable databases are GenBank, UniProt, Prosite, OMIM, PDB and RefSeq. We are planning to include other valuable databases and also planning to develop an inter-database search interface and a complex search facility combining keyword search and sequence similarity search.

8. Development of algorithms for biosynthetic process analysis

Kohichi Suematsu, Tetsuo Shibuya, Michihiro Araki, and Minoru Kanehisa

We developed algorithms for identifying biosynthetic process of medicinal products by utilizing the database of sub-molecular building blocks in biosynthetic processes. The problem is to find the most reasonable decomposition of a graph into subgraphs which are annotated in the database. We have developed new efficient algorithms and tools based on the algorithms for the problem, though the problem is a very difficult NP-hard problem.

9. Geometric Suffix Tree: A New Data structure for Indexing Protein Structures

Tetsuo Shibuya

Protein structure analysis is one of the most important research issues in the post-genomic era, and faster and more accurate index data structures for such 3-D structures are highly desired for research on proteins. We proposed a new data structure for indexing protein 3-D structures. For strings, there are many efficient indexing structures such as suffix trees, but it has been considered very difficult to design such sophisticated data structures against 3-D structures like proteins. Our index structure is based on the suffix trees and is called the geometric suffix tree. By using the geometric suffix

tree for a set of protein structures, we can search for all of their substructures whose RMSDs (root mean square deviations) or URMSDs (unit-vector root mean square deviations) to a given query 3-D structure are not larger than a given bound. Though there are $O(N^2)$ substructures, our data structure requires only $O(N)$ space where N is the sum of lengths of the set of proteins. We propose an $O(N^2)$ construction algorithm for it, while a naive algorithm would require $O(N^3)$ time to construct it. Moreover we propose an efficient search algorithm. Experiments show that we can search for similar structures much faster than naive RMSD computation against all over the database. The experiments also show that the construction time of the geometric suffix tree is practically almost linear to the size of the database, when applied to a protein structure database.

10. Protein Function Prediction based on 3-D Structure Motifs

Hiroki Sakai and Tetsuo Shibuya

Protein functions are said to be determined by its 3-D structures, but not all functions have been known to be related to some 3-D structure motifs. The geometric suffix tree, a data structure for indexing 3-D protein structures, which is also developed by us, enables comprehensive enumeration of all the possible structural motifs among given set of proteins. We are developing a new algorithm based on the support vector machine that decides protein's function from the 3-D structure of a protein. This algorithm utilizes all the possible 3-D motifs found by using the geometric suffix tree.

11. Genotype Clustering based on Hidden Markov Models

Ritsuko Onuki and Tetsuo Shibuya

Analysis of single nucleotide polymorphisms (SNPs) is one of the most important research topics in the post-genomic era. Recently, the genotype data increases exponentially and development of tools for analyzing these data is highly desired. We are developing a new algorithm for clustering genotypes, by utilizing the hidden markov model (HMM) that infers haplotypes from genotpyes.

12. The Origin of Chemical Diversity in Biosynthetic Circuits of Medicinal Natural Products

Michihiro Araki, Tetsuo Shibuya, Kohichi Suematsu, and Minoru Kanehisa

Medicinal natural products have been the major sources of bioactive compounds with diverse pharmacological activities, and are enzymically synthesized as secondary metabolites for specific biological purposes. The diverse activities are obviously explained by extensive diversities in the chemical structures. Although several hypotheses have been proposed to understand the origin of the chemical diversity, no systematic analyses have been done so far. We thus define molecular building blocks required for describing the chemical information of natural products to elucidate how chemical diversities are designed in the biosynthetic circuits. Each natural product is expressed as a combination of molecular building blocks with corresponding enzymatic information as links between building blocks to be collected in a database. The knowledge database constructed from various resources enables us to identify the system structures of the biosynthetic circuits. We also extract distinctive network structures consisted of both chemical and genomic information, which will be useful for understanding the origin of chemical diversity in the biosynthetic circuits.

13. Extracting A Strategy for Drug Design by Tracing Drug Development

Daichi Shigemizu, Michihiro Araki and Minoru Kanehisa

Current drugs are mostly derived by modification of known drug structures or from lead structures to be optimized for targeting new molecules or obtaining improved efficacy. To explore a design rule in drug development, it is necessary to focus on the empirical modifications to construct a knowledge base for tracing the chemical evolutions. We have been collecting data on the drug evolutions from databases and literatures, which has already been implemented on the drug structure maps in the KEGG DRUG database. In order to trace the chemical development in the database, binary relations of chemical structures were defined between before and after the chemical changes. We subsequently compared chemical structures in the binary relations to define functional atoms and groups supposed to be important for drug development. We have been collecting a series of such transformation patterns in the process of drug development to analyze and provide a design strategy in drug development.

14. Comprehensive Analysis of Distinctive Polyketide and Nonribosomal Peptide

Structural Motifs Encoded in Microbial Genomes

Yohsuke Minowa, Michihiro Araki and Minoru Kanehisa

We developed a highly accurate method to predict polyketide (PK) and nonribosomal peptide (NRP) structures encoded in the microbial genome. PKs/NRPs are polymers of carbonyl or peptidyl chains synthesized by polyketide synthases (PKS) and nonribosomal peptide synthetases (NRPS). We analyzed domain sequences corresponding to specific substrates and physical interactions between PKSs/NRPSs in order to predict which substrates are selected and assembled into highly ordered chemical structures. The predicted PKs/NRPs were represented as the sequences of carbonyl/peptidyl units to extract the structural motifs effeciently. We applied our method to 4,529 PKSs/NRPSs and found 619 PKs/NRPs. We also collected 1,449 PKs/NRPs whose chemical structures have been experimentally determined. The structural sequences were compared using the Smith-Waterman algorithm, and clustered into 271 clusters. From the compound clusters, we extracted 33 structural motifs which are significantly related with their bioactivities. The integrative analysis of genomic and chemical information here will provide a strategy to predict the chemical structures, the biosynthetic pathways, and the biological activities of PKs/NRPs, which is useful for the rational design of novel PKs/NRPs.

Publications

Yoshizawa AC, Kawashima S, Okuda S, Fujita M, Itoh M, Moriya Y, Hattori M, Kanehisa M. Extracting sequence motifs and the phylogenetic features of SNARE-dependent membrane traffic. Traffic. 7(8): 1104-1118, 2006.

Masoudi-Nejad A, Tonomura K, Kawashima S, Moriya Y, Suzuki M, Itoh M, Kanehisa M, Endo T, Goto S. EGassembler: online bioinformatics service for large-scale processing, clustering and assembling ESTs and genomic DNA fragments. Nucleic Acids Res. 34: W459-62, 2006.

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., and Hirakawa, M. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. 34, D354-357 (2006).

Hashimoto, K., Goto, S., Kawano, S., Aoki-Kinoshita, K.F., Ueda, N., Hamajima, M., Kawasaki, T., and Kanehisa, M. KEGG as a glycome informatics resource. Glycobiology 16, 63R-70 R (2006).

Honda, W., Kawashima, S., Kanehisa, M. Metabolite antigens and pathway incompatibility. Genome Inform. Ser. Workshop Genome Inform. 2006, 17(1): 184-193, 2006

Shibuya, T., Geometric Suffix Tree: A New Index Structure for Protein 3-D Structures, Combinatorial Pattern Matching 2006 (CPM 2006), LNCS 4009, Barcelona, July 5-7, 2006, pp. 84-93.

Suematsu, K., Araki, M., and Shibuya, T. Graph tiling algorithm for molecular synthesis analysis, SIGAL 106, May 18, 2006, Ikaho, pp. 25-32.

Shibuya, T., Geometric Suffix Tree: A New Index Structure for Protein 3-D Structures, IPSJ SIG Notes SIGAL 105-3, March 18 2006, Atsugi, pp. 17-24.

Shibuya, T., Combinatorial Structural Pattern Matching, The Second Japan-Taiwan Bilateral Symposium on Bioinformatics, 2006.

Shibuya, T., Indexing Structures for Biomolecular Structures, The First Japan-Taiwan Bilateral Symposium on Bioinformatics, 2006.

バイオインフォマティクス事典，日本バイオインフォマティクス学会編，共立出版，2006.

荒木通啓，金久實「創薬科学のためのKEGG」化学工業　Vol. 57　№3，42–46，2006.

Human Genome Center

Laboratory of DNA Information Analysis
DNA情報解析分野

Professor	Satoru Miyano, Ph.D.		教　授	理学博士	宮　野　　悟
Assistant Professor	Seiya Imoto, Ph.D.		助　手	博士(数理学)	井　元　清　哉
Assistant Professor	Masao Nagasaki, Ph.D.		助　手	博士(理学)	長　﨑　正　朗
Project Assistant Professor	Ryo Yoshida, Ph.D.		特任助手	博士(理学)	吉　田　　亮

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 interests are focused on Computational Systems Biology and its related computational techniques. 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 Systems Biology

a. Modeling and estimation of dynamic EGFR pathway by data assimilation approach using time series proteomic data

Shinya Tasaki, Masao Nagasaki, Masaaki Oyama, Hiroko Hata, Kazuko Ueno, Ryo Yoshida, Tomoyuki Higuchi, Sumio Sugano, Satoru Miyano

Cell Illustrator is a model building tool based on the Hybrid Functional Petri net with extension (HFPNe). By using Cell Illustrator, we have succeeded in modeling biological pathways, e.g., metabolic pathways, gene regulatory networks, microRNA regulatory networks, cell signaling networks, and cell-cell interactions. The recent development of tandem mass spectrometry coupled with liquid chromatography (LC/MS/MS) technology has enabled researchers to quantify the dynamic profile of a wide range of proteins within the cell. The proteomic data obtained by using LC/MS/MS has been considerably useful for introducing dynamics to the HFPNe model. Here, we report the first introduction of the time-series proteomic data to our HFPNe model. We constructed an epidermal growth factor receptor signal transduction pathway model (EFGR model) by using the biological data available in the literature. Then, the kinetic parameters were determined in the data assimilation (DA) framework with some manual tuning so as to fit the proteomic data published by Blagoev et al. (Nat. Biotechnol., 22: 1139. 1145, 2004). This in silico model was further refined by adding or removing some regulation loops using biological background knowledge. The DA framework was used to select the most plausible model from among the refined models. By using the proteomic data, we semi-automatically constructed a well-tuned EGFR HFPNe model by using the Cell Illustrator coupled with the DA framework.

b. Genomic data assimilation for estimating hybrid functional Petri net from time-course gene expression data

Masao Nagasaki, Rui Yamaguchi, Ryo Yoshida, Seiya Imoto, Atsushi Doi, Yoshinori Tamada, Hiroshi Matsuno, Satoru Miyano, Tomoyuki Higuchi

We propose an automatic construction method of the hybrid functional Petri net as a simulation model of biological pathways. The problems we consider are how we choose the values of parameters and how we set the network structure. Usually, we tune these unknown factors empirically so that the simulation results are consistent with biological knowledge. Obviously, this approach has the limitation in the size of network of interest. To extend the capability of the simulation model, we propose the use of data assimilation approach that was originally established in the field of geophysical simulation science. We provide genomic data assimilation framework that establishes a link between our simulation model and observed data like microarray gene expression data by using a nonlinear state space model. A key idea of our genomic data assimilation is that the unknown parameters in simulation model are converted as the parameter of the state space model and the estimates are obtained as the maximum a posteriori estimators. In the parameter estimation process, the simulation model is used to generate the system model in the state space model. Such a formulation enables us to handle both the model construction and the parameter tuning within a framework of the Bayesian statistical inferences. In particular, the Bayesian approach provides us a way of controlling overfitting during the parameter estimations that is essential for constructing a reliable biological pathway. We demonstrate the effectiveness of our approach using synthetic data. As a result, parameter estimation using genomic data assimilation works very well and the network structure is suitably selected.

c. Cell fate simulation model of gustatory neurons with microRNAs double-negative feedback loop by hybrid functional Petri net with extension

Ayumu Saito, Masao Nagasaki, Atushi Doi, Kazuko Ueno, Satoru Miyano

Biological regulatory networks have been extensively researched. Recently, the microRNA regulation has been analyzed and its importance has increasingly emerged. We have applied the Hybrid Functional Petri net with extension (HFPNe) model and succeeded in creating model biological pathways, e.g. metabolic pathways, gene regulatory networks, cell signaling networks, and cell-cell interaction models with one of the HFPNe implementations Cell Illustrator. Thus, we have applied HFPNe to model regulatory networks that involve a new key regulator microRNA. As a test case, we selected the cell fate determination model of two gustatory neurons of Caenorhabditis elegans…ASE left (ASEL) and ASE right (ASER). These neurons are morphologically bilaterally symmetric but physically asymmetric in function. Johnston et al. have suggested that their cell fate is determined by the double-negative feedback loop involving the lsy-6 and mir-273 microRNAs. Our simulation model confirms their hypothesis. In addition, other well-known mutants that are related with the double-negative feedback loop are also well-modeled. The new upstream regulator of lsy-6 (lsy-2) that is mentioned in another paper is also integrated into this model for the mechanism of switching between ASEL and ASER without any contradictions. Therefore, the HFPNe-based modeling will be one of the promising modeling methods and simulation architectures that illustrate microRNA regulatory networks.

d. A combined pathway to simulate CDK-dependent phosphorylation and ARF-dependent stabilization for p53 transcriptional activity

Atsushi Doi, Masao Nagasaki, Kazuko Ueno, Hiroshi Matsuno, Satoru Miyano

The protein p53 is phosphorylated by a member of protein kinases such as CDK7, and stabilized by the protein ARF. The phosphorylation and stabilization of p53 is believed to enhance its transcriptional activity and act simultaneously. Biological pathways composed of experts knowledge obtained from the literature are including these activation mechanisms. However, the map of biological pathways does not reflect the combination effect of phosphorylation and stabilization. We have conducted some simulations of biological pathways with hybrid functional Petri net (HFPN) after careful reading of papers. In this paper, we constructed the HFPN based biological pathway of CDK-dependent phosphorylation pathway and combine with ARF-dependent pathway described previously, to observe the effect of the phosphorylation on the stabilization with simulation-based validation.

e. Analysis of gene networks for drug target discovery and validation

Seiya Imoto, Yoshinori Tamada, Christopher J. Savoie, Satoru Miyano

Understanding responses of cellular system for a dosing molecule is one of the most important problems in pharmacogenomics. In this

chapter, we describe computational methods for identifying and validating drug target genes based on the gene network estimated from microarray gene expression data. We use two types of microarray gene expression data: One is gene disruptant microarray data and the other is time-course drug response microarray data. For this purpose, the information of gene networks plays an essential role and is unattainable from clustering methods, which are the standard for gene expression analysis. The gene network is estimated from disruptant microarray data by the Bayesian network model and then the proposed method automatically identifies sets of genes or gene regulatory pathways affected by the drug. We use an actual example from analysis of Saccharomyces cerevisiae gene expression profile data to express a concrete strategy for the application of gene network information toward drug target discovery. Key words: Drug target, gene network, Bayesian network, Boolean network, microarray data, Bayes statistics.

f. Finding module-based gene networks in time-course gene expression data with state space models

Rui Yamaguchi, Ryo Yoshida, Seiya Imoto, Tomoyuki Higuchi, Satoru Miyano

We discuss the use of the state space models to analyze time-course microarray gene expression data. Typical features of time-course microarray data are short time-course and high-dimensional observational vector. By these aspects, conventional statistical models such as the multivariate autoregressive models lead to unsuitable results due to the overfitting. The state space models have the potential to overcome this problem by the dimension reduction process. This paper provides (1) a survey of the state space models, (2) a review of some existing researches using the state space models for time-course microarray data together with a biological meaning of the model, (3) a solution for the lack of parameter identifiability, and (4) identification of biological system that is a central topic in bioinformatics. Finally, we show the usefulness of the state space models for time-course microarray data by the analysis of Saccharomyces cerevisiae cell cycle gene expression data.

g. Structural modeling and analysis of signaling pathways based on Petri nets

Chen Li, Shunichi Suzuki, Qi-Wei Ge, Mitsuru Nakata, Hiroshi Matsuno, Satoru Miyano

The purpose of this study is to discuss how to model and analyze signaling pathways by using Petri net. Firstly, we propose a modeling method based on Petri net by paying attention to the molecular interactions and mechanisms. Then, we introduce a new notion "activation transduction component" in order to describe an enzymic activation process of reactions in signaling pathways and shows its correspondence to a so-called elementary T-invariant in the Petri net models. Further, we design an algorithm to effectively find basic enzymic activation processes by obtaining a series of elementary T-invariants in the Petri net models. Finally, we demonstrate how our method is practically used in modeling and analyzing signaling pathway mediated by thrombopoietin as an example.

2. Statistical and Computational Knowledge Discovery

a. A statistical framework for genome-wide discovery of biomarker splice variations with GeneChip Human Exon 1.0 ST arrays

Ryo Yoshida, Kazuyuki Numata, Seiya Imoto, Masao Nagasaki, Atsushi Doi, Kazuko Ueno, Satoru Miyano

Alternative splicing is an important regulatory mechanism that generates multiple mRNA transcripts which are transcribed into functionally diverse proteins. According to the current studies, aberrant transcripts due to splicing mutations are known to cause for 15% of genetic diseases. Therefore understanding regulatory mechanism of alternative splicing is essential for identifying potential biomarkers for several types of human diseases. Most recently, advent of GeneChip® Human Exon 1.0 ST Array enables us to measure genome-wide expression profiles of over one million exons. With this new microarray platform, analysis of functional gene expressions could be extended to detect not only differentially expressed genes, but also a set of specific-splicing events that are differentially observed between one or more experimental conditions, e.g. tumor or normal control cells. In this study, we address the statistical problems to identify differentially observed splicing variations from exon expression profiles. The proposed method is organized according to the following process: (1) Data preprocessing for removing systematic biases from the probe intensities. (2) Whole transcript analysis with the analysis of variance (ANOVA) to identify a set of loci that cause the alternative splicing-related to a certain disease. We test the proposed statistical approach on exon expression profiles of colorectal carcinoma. The applicability is veri-

fied and discussed in relation to the existing biological knowledge. This paper intends to highlight the potential role of statistical analysis of all exon microarray data. Our work is an important first step toward development of more advanced statistical technology. Supplementary information and materials are available from http://bonsai.ims.u-tokyou.ac.jp/~yoshidar/IBSB2006_ExonArray.htm

b. Machine learning prediction of amino acid patterns in protein N-myristoylation

Ryo Okada, Chigusa Miyakawa, Manabu Sugii, Hiroshi Matsuno, Satoru Miyano

Protein N-myristoylation is the lipid modification in which the 14-th carbon saturated fatty acid binds covalently to N-terminal of virus-based and eukaryotic protein. In this study, we suggest an approach to predict the pattern of N-myristoylation signal using the machine learning system BONSAI. BONSAI finds rules in combination of an alphabet indexings and decision trees. Computational experiments with BONSAI classified amino acid residues depending on effect for N-myristoylation and found rules of the alphabet indexing. In addition, BONSAI suggested new requirements for the position of an amino acid in N-myristoylation singal.

c. Contribution of comparative fish studies to general endocrinology: structure and function of some osmoregulatory hormones

Yoshio Takei, Akatsuki Kawakoshi, Takehiro Tsukada, Shinya Yuge, Maho Ogoshi, Koji Inoue, Susumu Hyodo, Hideo Bannai, Satoru Miyano

Fish endocrinologists are commonly motivated to pursue their research driven by their own interests in these aquatic animals. However, the data obtained in fish studies not only satisfy their own interests but often contribute more generally to the studies of other vertebrates, including mammals. The life of fishes is characterized by the aquatic habitat, which demands many physiological adjustments distinct from the terrestrial life. Among them, body fluid regulation is of particular importance as the body fluids are exposed to media of varying salinities only across the thin respiratory epithelia of the gills. Endocrine systems play pivotal roles in the homeostatic control of body fluid balance. Judging from the habitat-dependent control mechanisms, some osmoregulatory hormones of fish should have undergone functional and molecular evolution during the ecological transition to the terrestrial life. In fact, water-regulating hormones such as vasopressin are essential for survival on the land, whereas ion-regulating hormones such as natriuretic peptides, guanylins and adrenomedullins are diversified and exhibit more critical functions in aquatic species. In this short review, we introduce some examples illustrating how comparative fish studies contribute to general endocrinology by taking advantage of such differences between fishes and tetrapods. In a functional context, fish studies often afford a deeper understanding of the essential actions of a hormone across vertebrate taxa. Using the natriuretic peptide family as an example, we suggest that more functional studies on fishes will bring similar rewards of understanding. At the molecular level, recent establishment of genome databases in fishes and mammals brings clues to the evolutionary history of hormone molecules via a comparative genomic approach. Because of the functional and molecular diversification of ion-regulating hormones in fishes, this approach sometimes leads to the discovery of new hormones in tetrapods as exemplified by adrenomedullin 2.

Publications

1. DeLisi, C., Kanehisa, M., Heinrich, R., Miyano, S. (Eds.) Genome Informatics. 17 (1), 2006.
2. Doi, A., Nagasaki, M., Matsuno, H., Miyano, S. Simulation based validation of the p 53 transcriptional activity with hybrid unctional Petri net. In Silico Biology. 6 (1-2): 1-13, 2006.
3. Doi, A., Nagasaki, M., Ueno, K., Matsuno, H., Miyano, S.A combined pathway to simulate CDK-dependent phosphorylation and ARF-dependent stabilization for p53 transcriptional activity. Genome Informatics. 17 (1): 112-123, 2006.
4. Imoto, S., Miyano, S. Bayesian network approach to estimate gene networks. in A. Mittal, A. Kassim and T. Tan (Eds.), Bayesian Network Technologies: Applications and Graphical Models, Idea Group Publishers, USA. In press.
5. Imoto, S., Tamada, Y., Araki, H., Yasuda, K., Print, C.G., Charnock-Jones, S.D., Sanders, D., Savoie, C.J., Tashiro, K., Kuhara, S., Miyano, S. Computational strategy for dis-

covering druggable gene networks from genome-wide RNA expression profiles. Pacific Symposium on Biocomputing. 11: 559-571, 2006.
6. Imoto, S., Tamada, Y., Savoie, C.J., Miyano, S., Analysis of gene networks for drug target discovery and validation. Methods in Molecular Biology. 360: 33-56, 2006.
7. Imoto, S., Higuchi, T., Goto, T., Miyano, S. Error tolerant model for incorporating biological knowledge with expression data in estimating gene networks. Statistical Methodology. 3 (1): 1-16, 2006.
8. Jeong, E., Miyano, S.A weighted profile based method for protein-RNA interacting residue prediction. Transactions on Computational Systems Biology. 123-139, 2006.
9. Li, C., Suzuki, S., Ge, Q.-W., Nakata, M., Matsuno, H., Miyano, S. Structural modeling and analysis of signaling pathways based on petri nets. J. Bioinf. Comput. Biol. 4 (5): 1119-1140, 2006.
10. Matsuno, H., Inouye, S.-T., Okitsu, Y., Fujii, Y., Miyano, S.A new regulatory interactions suggested by simulations for circadian genetic control mechanism in mammals. J. Bioinf. Comput. Biol. 4 (1): 139-154, 2006.
11. Matsuno, H., Li, C., Miyano, S. Petri net based description for systematic understanding of biological pathways. IEICE Trans. Fundamentals. E89-A (11): 3166-3174, 2006.
12. Miyano, S. (Ed.) Special RECOMB 2005 Issue. J. Comp. Biol. 13 (2), 2006.
13. Nagasaki, M., Yamaguchi, R., Yoshida, R., Imoto, S., Doi, A., Tamada, Y., Matsuno, H., Miyano, S., Higuchi, T. Genomic data assimilation for estimating hybrid functional Petri net from time-course gene expression data. Genome Informatics. 17 (1). 46-61, 2006.
14. Nakamichi, R., Imoto, S., Miyano, S. Statistical model selection method to analyze combinatorial effects of SNPs and environmental factors for binary disease. International Journal on Artificial Intelligence Tools. 15 (5), 711-724, 2006.
15. Okada, R., Sugii, M., Matsuno, H., Miyano, S. Machine learning prediction of amino acid patterns in protein N-myristoylation. Lecture Notes in Bioinformatics. 4146: 4-14, 2006.
16. Saito, A., Nagasaki, M., Doi, A., Ueno, K., Miyano, S. Cell fate simulation model of gustatory neurons with microRNAs double-negative feedback loops by hybrid functional Petri net with extension. Genome Informatics 17 (1): 100-111, 2006.
17. Sakakibara, Y., Smith, T.F., Kanehisa, M., Miyano, S., Takagi, T. (Eds.) Genome Informatics. 17 (2), 2006.
18. Takei, Y., Kawakoshi, A., Tsukada, T., Yuge, S., Ogoshi, M., Inoue, K., Hyodo, S., Bannai, H., Miyano, S. Contribution of comparative fish studies to general endocrinology: structure and function of some osmoregulatory hormones. J. Experimental Zoology. Part A, Comparative Experimental Biology. 305 (9): 787-798, 2006.
19. Tamada, Y., Imoto, S., Miyano, S. Estimating gene networks from gene expression data utilizing biological information. Proc. Inst. Statist. Math. 54 (2): 333-356, 2006.
20. Tasaki, S., Nagasaki, M., Oyama, M., Hata, H., Ueno, K., Yoshida, R., Higuchi, T., Sugano, S., Miyano, S. Modeling and estimation of dynamic EGFR pathway by data assimilation approach using time series proteomic data. Genome Informatics. 17 (2): 226-228, 2006.
21. Washio, T., Higuchi, T., Imoto, S., Tamada, Y., Sato, K., Motoda, H. Graph mining and its application to statistical modeling. Proc. Inst. Statist. Math. 54 (2): 315-331, 2006.
22. Yamaguchi, R., Yoshida, R., Imoto, S., Higuchi, T., Miyano, S. Finding module-based gene networks in time-course gene expression data with state space models. IEEE Signal Processing Magazine. In press.
23. Yoshida, R., Higuchi, T., Imoto, S., Miyano, S. ArrayCluster: an analytic tool for clustering, data visualization and module finder on gene expression profiles. Bioinformatics. 22 (12): 1538-1539, 2006.
24. Yoshida, R., Numata, K., Imoto, S., Nagasaki, M., Doi, A., Ueno, K., Miyano, S.A statistical framework for genome-wide discovery of biomarker splice variations with GeneChip Human Exon 1.0 ST arrays. Genome Informatics. 17 (1): 88-99, 2006.
25. 宮野　悟，江口至洋，金久　實，高木利久，中井謙太(編)．バイオインフォマティクス事典．共立出版．2006．

Human Genome Center

Laboratory of Molecular Medicine
Laboratory of Genome Technology
Division of Advanced Clinical Proteomics

ゲノムシークエンス解析分野
シークエンス技術開発分野
先端臨床プロテオミクス共同研究ユニット

Professor	Yusuke Nakamura, M.D., Ph.D.	教　授	医学博士	中　村　祐　輔
Assistant Professor	Hidewaki Nakagawa, M.D., Ph.D.	助　手	医学博士	中　川　英　刀
Assistant Professor	Koichi Matsuda, M.D., Ph.D.	助　手	医学博士	松　田　浩　一
Assistant Professor	Hitoshi Zembutsu, M.D., Ph.D.	助　手	医学博士	前　佛　　均
Associate Professor	Toyomasa Katagiri, Ph.D.	助教授	医学博士	片　桐　豊　雅
Assistant Professor	Ryuji Hamamoto, Ph.D.	助　手	理学博士	浜　本　隆　二
Project Associate Professor	Yataro Daigo, M.D., Ph.D.	特任助教授	医学博士	醍　醐　弥太郎

The major goal of the Human Genome Project is to identify genes of medical importance, and to develop new diagnostic and therapeutic tools. We have been attempting to isolate genes involving in carcinogenesis and also those causing or predisposing to various diseases as well as those related to drug efficacies and adverse reactions. By means of technologies developed through the genome project including a high-resolution SNP map, a large-scale DNA sequencing, and the cDNA microarray method, we have isolated a number of biologically and/or medically important genes.

1. Genes playing significant roles in human cancers

Toyomasa Katagiri, Yataro Daigo, Hidewaki Nakagawa, Koichi Matsuda, Hitoshi Zembutsu, Ryo Takata, Mitsugu Kanehira, Koji Takahashi, Chiyuki Furukawa, Atsushi Takano, Nobuhisa Ishikawa, Tatsuya Kato, Satoshi Hayama, Chie Suzuki, Akira Togashi, Kazuhito Morioka, Tomohide Kidokoro, Chizu Tanikawa, Asahi Hishida, Miki Akiyama, Sachiko Dobashi, Meng-Lay Lin, Jae-Hyun Park, Tomomi Ueki, Yosuke Harada, Koichiro Inaki, Chikako Fukukawa, Koji Ueda, Minh Hue Nguyen, Yuria Mano, Masaya Taniwaki, Ryohei Nishino, Daizaburo Hirata, Takumi Yamabuki, Nagato Sato, Kenji Tamura, Masayo Hosokawa, Su-Youn Chung, Motohide Uemura, Akio Takehara, Arata Shimo, Eiji Hirota, and Yusuke Nakamura

(1) Chemosensitivity

To predict the efficacy of the M-VAC neoadjuvant chemotherapy for invasive bladder cancers, we previously established the method to calcu-

late the prediction score on the basis of expression profiles of 14 predictive genes. This scoring system had clearly distinguished the responder group from the non-responder group. To further validate the clinical significance of the system, we applied it to 22 additional cases of bladder cancer patients and found that the scoring system correctly predicted clinical response for 19 of the 22 test cases. The group of patients with positive predictive scores had significantly longer survival than that with negative scores. When we compared our results with the previous report describing the prognosis of the patients with cystectomy alone, the results imply that patients with positive scores are likely to have benefit by having M-VAC neoadjuvant chemotherapy, but that the chemotherapy would shorten lives of patients with negative scores. We are confident that our prediction system to M-VAC therapy should provide opportunities for achieving better prognosis and improving quality of life of patients. Taken together, our data suggest that the goal of "personalized medicine," prescribing the appropriate treatment regimen for each patient, may be achievable by selecting specific sets of genes for their predictive values.

Philadelphia-chromosome-positive acute lymphoblastic leukemia (Ph+ALL) revealed one of the poorest prognoses among leukemias due to its high incidence of relapse. Although more than 96% of patients with Ph-positive ALL achieved complete remission (CR) by the imatinib-combined chemotherapy in a phase II study conducted by the Japan adult leukemia study group (JALSG), 25% of them experienced relapse. To establish a prediction system for a risk of relapse after CR, we analyzed gene expression profiles of 23 bone marrow samples from patients with Ph+ALL using cDNA microarray consisting of 27,648 cDNA sequences. Using the 19 randomly-selected test cases, we identified 16 genes that were expressed significantly differently between "non-relapse" (8 patients) and "relapse" (11 patients) groups; from the list of 16 genes, we selected the 6 "predictive" genes that showed significant differences and constructed a numerical prediction scoring system by which the non-relapse group (with positive scores) was clearly separated from the relapse group (with negative scores). Scoring of 4 cases that were reserved from the original 23 cases predicted correctly the responses to imatinib-combined chemotherapy. In addition, three cases, that were resistant to the imatinib-combined therapy and failed to induce remission, also revealed the negative scores. Because real-time reverse transcription-PCR data were highly concordant with the cDNA microarray data for those 6 genes, we developed a quantitative reverse transcription-PCR based prediction system that could be feasible for routine clinical use. Our results suggest that possibility of the relapse after complete remission by the imatinib -combined chemotherapy can be predicted by expression patterns in this set of genes, leading to achievement of "personalized therapy" for treatment of this disease.

(2) Lung cancer

We found co-transactivation of CDCA1 (cell division associated 1) and KNTC2 (kinetocore associated 2), members of the evolutionarily-conserved centromere protein complex, in non-small cell lung carcinomas (NSCLCs). Immunohistochemical analysis using lung-cancer tissue microarray confirmed high levels of CDCA1 and KNTC2 proteins in the great majority of lung cancers of various histological types. Their elevated expressions were associated with poorer prognosis of NSCLC patients. Knockdown of either CDCA1 or KNTC2 expression with siRNA significantly suppressed growth of NSCLC cells. Furthermore, inhibition of their binding by a cell-permeable peptide carrying the CDCA1-derived 19 amino-acid peptide (11R-$CDCA1_{398-416}$) that correspond to the binding domain to KNTC2 effectively suppressed growth of NSCLC cells. As our data imply that human CDCA1 and KNTC2 appear to fall in the category of cancer-testis antigens and their simultaneous up-regulation is a frequent and important feature of cell growth/survival of lung-cancer, selective suppression of CDCA1 or KNTC2 activity, and/or inhibition of the CDCA1-KNTC2 complex formation could be a promising therapeutic target for treatment of lung cancer.

We also identified abundant expression of neuromedin U (NMU) in the great majority of lung cancers. Immunohistochemical analysis demonstrated significant association of NMU expression with poorer prognosis of NSCLC patients. Treatment of NSCLC cells with siRNA against NMU suppressed its expression and inhibited growth of the cells; on the other hand, induction of exogenous expression of NMU conferred growth-promoting activity and enhanced the cell mobility *in vitro*. We found that two G protein-coupled receptors, growth hormone secretagogue receptor 1b (GHSR1b) and neurotensin receptor 1 (NTSR1), were also over-expressed in NSCLC cells and a hetero-dimer complex of these receptors functioned as an NMU receptor. The NMU-receptor interaction subsequently induced generation of a second messenger, cAMP, to activate its downstream genes including transcription factors and cell cy-

cle regulators. Treatment of NSCLC cells with siRNAs for GHSR or NTSR1 suppressed expression of those genes and the growth of NSCLC cells. These data strongly implied that targeting the NMU signaling pathway would be a promising therapeutic strategy for treatment of lung cancers.

In addition, we found over-expression of a *MAPJD* (Myc-associated protein with JmjC domain) gene in the great majority of NSCLC cases. Induction of exogenous expression of MAPJD into NIH3T3 cells conferred growth-promoting activity. Concordantly, *in vitro* suppression of MAPJD expression with siRNA effectively suppressed growth of NSCLC cells in which MAPJD was over-expressed. We found four candidate MAPJD-target genes, *SBNO1, TGFBRAP1, RIOK1*, and *RASGEF1A*, which were the most significantly induced by exogenous MAPJD expression. Through interaction with MYC protein, MAPJD transactivates a set of genes including kinases and cell signal transducers that are possibly related to proliferation of lung cancer cells. As our data imply that MAPJD is a novel member of the MYC transcriptional complex and its activation is a common feature of lung-cancer, selective suppression of this pathway could be a promising therapeutic target for treatment of lung cancers.

To identify molecules that might serve as biomarkers or targets for development of novel molecular therapies, we have been screening genes encoding transmembrane/secretory proteins that are up-regulated in lung cancers, using cDNA microarrays coupled with purification of tumor cells by laser microdissection. A gene encoding seizure-related 6 homolog (mouse)-like 2 (SEZ6L2) protein, was chosen as a candidate for such molecule. Semi-quantitative RT-PCR and western-blot analyses documented increased expression of SEZ6L2 in the majority of primary lung cancers and lung-cancer cell lines examined. SEZ6L2 protein was proven to be present on the surface of lung-cancer cells by flow cytometrical analysis using anti-SEZ6L2 antibody. Immunohistochemical staining for tumor tissue microarray consisting of 440 archived lung-cancer specimens detected positive SEZ6L2 staining in 327 (78%) of 420 non-small cell lung cancers (NSCLCs) and 13 (65%) of 20 small-cell lung cancers (SCLCs) examined. Moreover, NSCLC patients whose tumors revealed a higher level of SEZ6L2 expression suffered shorter tumor-specific survival compared to those with no SEZ6L2 expression. These results indicate that SEZ6L2 should be a useful prognostic marker of lung cancers.

Furthermore, to characterize the molecular mechanisms involved in carcinogenesis and progression of small-cell lung cancer (SCLC) and identify molecules to be applicable as novel diagnostic markers and/or for development of molecular-targeted drugs, we applied cDNA microarray profile analysis coupled with purification of cancer cells by laser-microbeam microdissection (LMM). Expression profiles of 32,256 genes in 15 SCLCs identified 252 genes that were commonly up-regulated and 851 transcripts that were down-regulated in SCLC cells compared with non-cancerous lung tissue cells. An unsupervised clustering algorithm applied to the expression data easily distinguished SCLC from the other major histological type of non-small cell lung cancer (NSCLC) and identified 475 genes that may represent distinct molecular features of each of the two histological types. In particular, SCLC was characterized by altered expression of genes related to neuroendocrine cell differentiation and/or growth such as *ASCL 1, NRCAM*, and *INSM1*. We also identified 68 genes that were abundantly expressed both in advanced SCLCs and advanced adenocarcinomas (ADCs), both of which had been obtained from patients with extensive chemotherapy treatment. Some of them are known to be transcription factors and/or gene expression regulators such as *TAF5L, TFCP2L4, PHF20, LMO4, TCF20, RFX2*, and *DKFZp547I048* as well as those encoding nucleotide-binding proteins such as *C9orf76, EHD3*, and *GIMAP4*. Our data provide valuable information for better understanding of lung carcinogenesis and chemoresistance.

(3) Pancreatic cancer

Pancreatic ductal adenocarcinoma (PDAC) shows the worst mortality among the common malignancies, with a 5-year survival rate of only 4%, and the majority of PDAC patients are diagnosed at an advanced stage, in which no effective therapy is available at present. Although the proportion of curable cases is still not so high, surgical resection of early-staged PDACs is the only way to cure the disease. Hence, establishment of a screening strategy to detect early-staged PDACs by novel serological markers is urgently required, and development of novel molecular therapies for PDAC treatment is also eagerly expected. We here report over-expression of REG4, a new member of the REG family, in PDAC cells on the basis of the genome-wide cDNA microarray analysis as well as RT-PCR and immunohistochemical analysis. We also detected significant elevation of REG4 in serum of some parts of patients with early-staged PDACs by our ELISA system, indicating the possibility of REG4 as a new serological marker of PDACs. Furthermore, we found that

knockdown of the endogenous REG4 expression in PDAC cell lines with siRNA caused decrease of cell viability. Concordantly, addition of recombinant REG4 to the culture medium enhanced growth of PDAC cell line in a dose-dependent manner. A monoclonal antibody against REG4 neutralized its growth-promoting effects and attenuated significantly the growth of PDAC cells. These findings implicate that REG4 is a promising tumor marker to screen early-staged PDAC and also that neutralization of REG4 by the antibody may offer us novel potential tools for treatment of PDACs.

Among dozens of up-regulated genes in PDAC cells, we also focused on one tyrosine kinase receptor, Eph receptor A4 (EphA4), as a molecular target for PDAC therapy. Immunohistochemical analysis validated EphA4 overexpression in approximately a half of PDAC tissues. To investigate its biological function in PDAC cells, we knocked down *EphA4* expression by siRNA, which drastically attenuated PDAC cell viability. Concordantly to the siRNA experiment, PDAC-derivative cells that were designed to constitutively express exogenous EphA4 showed more rapid growth rate than cells transfected with mock vector, suggesting the growth-promoting effect of EphA4 on PDAC cells. Furthermore, the expression analysis for ephrin ligand family members indicated coexistence of ephrinA3 ligand in PDAC cells with EphA4 receptor, and knockdown of *ephrinA3* by siRNA also attenuated PDAC cell viability as well as *EphA4*. These results implicate that the EphA4-ephrinA3 pathway is likely to be a promising molecular target for pancreatic cancer therapy.

(4) Prostate cancer

Through genome-wide cDNA microarray analysis coupled with microdissection of prostate cancer cells, we identified *MICAL2-PV (Molecule Interacting with CasL-2 Prostate Cancer Variants)*, novel splicing variants of *MICAL2*, showing overexpression in PC cells. Northern blot analysis demonstrated that *MICAL2-PVs* were cancer-testis specific transcripts. Immunohistochemical analysis using an antibody generated specific to MICAL2-PV revealed that MICAL2-PV was expressed in the cytoplasm of cancer cells with various staining patterns and intensities, while it was not or hardly detectable in adjacent normal prostate epithelium or prostatic intraepithelial neoplasia (PIN). Interestingly, immunohistochemical analysis of 105 PC specimens on the tissue microarray indicated that MICAL2-PV expression status was significantly correlated with Gleason scores ($P<0.0001$) or tumor classification ($p<0.0001$). Furthermore, the expression levels of MICAL2-PVs were also concordant to those of c-Met, a marker of tumor aggressiveness, with statistical significance ($p = 0.0018$). To investigate its biological function in PC cells *in vitro*, we knocked down endogenous *MICAL2-PVs* in PC cells by siRNA, which resulted in the significant reduction of PC cell viability. Our findings suggest that MICAL2-PVs is likely to be involved in tumor progression or aggressiveness of PC, and could be a candidate as a novel molecular marker and/or a target for treatment of advanced PCs.

(5) Breast Cancer

We previously reported that up-regulation of SMYD3, a histone H3 lysine-4 specific methyltransferase, plays a key role in the proliferation of colorectal carcinoma (CRC) and hepatocellular carcinoma (HCC). In this study, we reveal that SMYD3 expression is also elevated in a great majority of breast cancer tissues. Similarly to CRC and HCC, silencing of SMYD3 by siRNA to this gene resulted in the inhibited growth of breast cancer cells, suggesting that the increased SMYD3 expression is also essential for the proliferation of the breast cancer cells. Moreover, we show here that SMYD3 could promote breast carcinogenesis by directly regulating the expression of the proto-oncogene *WNT10B*. These data imply that augmented SMYD3 plays a crucial role in breast carcinogenesis, and that inhibition of SMYD3 should be a novel therapeutic strategy for treatment of breast cancer.

Cancer therapy directing at specific molecular targets in signaling pathways of cancer cells such as Tamoxifen, aromatase inhibitors and trastuzumab has been proven its usefulness for treatment of advanced breast cancers. However, increases of the risk of endometrial cancer by long-term tamoxifen administration as well as those of bone fracture due to osteoporosis in postmenopausal women with aromatase inhibitor prescription are recognized as their side effects. Due to the emergence of these side effects and also drug resistance, it is necessary to search novel targets for molecularly-orientated drugs on the basis of well-characterized mechanisms of action. Using the accurate genome-wide expression profiles of breast cancers, we found *maternal embryonic leucine-zipper kinase (MELK)* that was significantly overexpressed in the great majority of breast cancer cells. To assess a possible role of *MELK* in mammary carcinogenesis, we knocked down the expression of endogenous *MELK* in breast cancer cell-lines by means of the mammalian vector-based RNA interference (RNAi) technique. Furthermore, we

identified a long isoform of Bcl-G (Bcl-G_L), a pro -apoptotic member of Bcl-2 family, as a possible substrate (s) for the MELK kinase by pull-down assay with wild-type-and kinase-dead-MELK recombinant proteins. Finally, we performed TUNEL assay and FACS analysis to measure the proportions of sub-G1 population to investigate MELK is involved in apoptosis cascade through the Bcl-G_L-related pathway. The multiple human tissues-and cancer cell lines-northern blot analyses demonstrated that *MELK* was overexpressed at a significantly high level in a great majority of breast cancers and cancer cell lines, but not expressed in normal vital organs (heart, liver, lung, hidney). Suppression of *MELK* expression with small-interfering RNA significantly inhibited growth of human breast cancer cells. We also found that MELK protein physically interacted with Bcl-G_L protein, a pro-apoptotic member of the Bcl-2 family through its N-terminal region. Subsequent immunocomplex kinase assay showed Bcl-G_L was specifically phosphorylated by MELK *in vitro*. TUNEL assay and FACS analysis revealed that overexpression of WT-MELK suppressed Bcl-G_L-induced apoptosis, while that of D150A-MELK did not. Our findings suggest that kinase activity of MELK is likely to be involved in mammary carcinogenesis through inhibition of pro-apoptotic function of Bcl-G_L. The kinase activity of MELK should be a promising target for development of molecular-targeting therapy for patients with breast cancers.

We also focused on one gene that encodes PBK/TOPK including a kinase domain. Northern blot analyses using mRNAs of normal human organs, breast cancer tissues, and cancer cell-lines indicated this molecule to be a novel cancer/testis antigen. Reduction of PBK/TOPK expression by siRNA resulted in significant suppression of cell growth probably due to dysfunction in the cytokinetic process. Immunocytochemical analysis with anti-PBK/TOPK antibody implicated a critical role of PBK/TOPK in an early step of mitosis. PBK/TOPK could phosphorylate histone H3 at Ser10 *in viro* and *in vivo*, and medicated its growth-promoting effect through histone H3 modification. Since PBK/TOPK is the cancer/testis antigen and its kinase function is likely to be related to its oncogenic activity, we suggest PBK/TOPK to be a promising molecular target for breast cancer therapy.

Among the up-regulated genes, we further focused on functional significance of *PRC1* (protein regulator of cytokinesis 1) in development of breast cancer. Western blot analysis using breast cancer cell-lines revealed a significant increase of endogenous PRC1 level in G_2/M phase. Treatment of breast cancer cells with small-interfering RNAs (siRNAs) against PRC1 effectively suppressed PRC1 expression and inhibited the growth of breast cancer cells, T47D and HBC5. Furthermore, we found interaction of PRC1 and KIF2C/MCAK (Kinesin family member 2C/Mitotic centromere-associated kinesin) by co-immunoprecipitation and immunoblotting using COS-7 cells in which these molecules were introduced exogenously. These findings suggest a possible involvement of the PRC1 -KIF2C/MCAK complex in breast tumorigenesis and that this complex should be a promising target for development of novel treatment for breast cancer.

(6) Colon cancer

Through a genome-wide cDNA microarray analysis, we identified a number of genes whose expression was up-regulated frequently in colorectal cancer. Among them, we here report a gene termed *FAM84A* that was expressed in none of 22 normal tissues examined except the testis. Although immunocytochemical staining revealed localization of FAM84A protein in the sub-cellular membrane region, the staining was observed limitedly in the region lacking the attachment with neighboring cells. In addition, we found that exogenous FAM84A expression increased cell motility in NIH3T3 cells, and that phosphorylation of serine38 of FAM84A was associated with morphology of cells. Our results indicate a possibility that up-regulation of FAM 84A plays a critical role in progression of colon cancer.

(7) Renal cancer

In order to clarify the molecular mechanism involved in renal carcinogenesis, and identify molecular targets for diagnosis and treatment, we analyzed genome-wide gene expression profiles of 15 surgical specimens of clear cell renal cell carcinoma (RCC), compared to normal renal cortex, using a combination of laser microbeam microdissection (LMM) with a cDNA microarray representing 27,648 genes. We identified 257 genes that were commonly up-regulated and 721 genes that were down-regulated in RCCs. Interestingly, none of top 24 up-regulated genes that showed most significant differences in informative RCC-cases were included in previously reports describing expression profiles of ccRCC using RNAs isolated from bulk tissues. These findings suggest that it is important to purify as much as possible the populations of cancerous and normal epithelial cells obtained from surgical specimens. Among significantly-transactivated genes, we in this study focused

on Semaphorin 5B (*SEMA5B*) and knocked-down its expression in RCC cells by small-interfering RNA (siRNA). Effective down-regulation of its expression levels in RCC cells significantly attenuated RCC cell viability. In conclusion, our data should be helpful for a better understanding of the tumorigenesis of RCC and should contribute to the development of diagnostic tumor marker and molecular-targeting therapy for patients with RCC.

(8) Bladder cancer

To disclose molecular mechanism of bladder cancer, the second most common genitourinary tumor, we had previously performed genome-wide expression profile analysis of 26 bladder cancers by means of cDNA microarray representing 27,648 genes. Among genes that were significantly up-regulated in the majority of bladder cancers, we here report identification of *MPHOSPH1 (M-phase phosphoprotein 1)* as a candidate molecule for drug development for bladder cancer. Northern blot analyses using mRNAs of normal human organs and cancer cell-lines indicated this molecule to be a novel cancer/testis antigen. Introduction of *MPHOSPH1* into NIH3T3 cells significantly enhanced cell growth at *in vitro* and *in vivo* conditions. We subsequently found an interaction between MPHOSPH1 and PRC1 (Protein Regulator of Cytokinesis 1), which was also up-regulated in bladder cancer cells. Immunocytochemical analysis revealed co-localization of endogenous MPHOSPH1 and PRC1 proteins in bladder cancer cells. Interestingly, knockdown of either of MPHOSPH1 or PRC1 expression with specific siRNAs caused significant increase of multi-nuclear cells and subsequent cell death of bladder cancer cells. Our results imply that the MPHOSPH1/PRC1 complex is likely to play a crucial role in bladder carcinogenesis and that inhibition of the MPHOSPH1/PRC1 expression or their interaction should be novel therapeutic targets for bladder cancers.

(9) Cholangiocarcinoma

Intrahepatic cholangiocarcinoma (ICC) is the second most common primary cancer in the liver, and its incidence is highest in northeastern part of Thailand. ICCs in this region are known to be associated with infection with liver flukes, particularly Opisthorchis viverrini (OV), as well as nitrosamines from food. To clarify molecular mechanisms of ICC associated with or without liver flukes, we analyzed gene-expression profiles of fluke-associated ICCs from 20 Thai patients, and compared their profiles with those of 20 Japanese ICCs that were not associated with fluke by means of laser-microbeam-microdissection and a cDNA microarray containing 27,648 genes. We identified 77 commonly upregulated genes and 325 commonly downregulated genes in the two ICC groups. Unsupervised hierarchical cluster analysis separated the 40 ICCs into two major branches almost completely according to the fluke status. The putative signature of liver fluke-associated ICC exhibited elevated expression of genes involved in xenobiotic metabolism (*UGT2B11, UGT1A10, CHST4, SULT1C1*), while that of non-liver fluke-associated ICC represented enhanced expression of genes related to growth factor signaling (*TGFBI, PGF, IGFBP1, IGFBP3*). Additional random permutation tests identified a total of 52 genes whose expression levels were significantly different between the two groups. We also identified 17 genes associated with macroscopic type of fluke-associated ICCs. These data may not only contribute to clarification of common and fluke-specific mechanisms underlying ICC, but also serve as a starting point for the identification of novel diagnostic markers and/or therapeutic targets for the disease.

(10) Biomarker for esophageal and lung cancer

Gene-expression profile analysis of lung and esophageal carcinomas revealed that Dikkopf-1 (DKK1) was highly transactivated in the great majority of lung cancers and esophageal squamous-cell carcinomas (ESCCs). Immunohistochemical staining using tumor tissue microarrays consisting of 279 archived non-small cell lung cancers (NSCLCs) and 280 ESCC specimens demonstrated that a high level of DKK1 expression was associated with poor prognosis of patients with NSCLC as well as ESCC, and multivariate analysis confirmed its independent prognostic value for NSCLC. In addition, we identified that exogenous expression of DKK1 increased the migratory activity of mammalian cells, suggesting that DKK1 may play a significant role in progression of human cancer. We established an ELISA system to measure serum levels of DKK1 and found that serum DKK1 levels were significantly higher in lung and esophageal cancer patients than in healthy controls. The proportion of the DKK1-positive cases was 126 (70.0%) of 180 NSCLC, 59 (69.4%) of 85 SCLC, and 51 (63.0%) of 81 ESCC patients, while only 10 (4.8%) of 207 healthy volunteers were falsely diagnosed as positive. A combined ELISA assays for both DKK1 and CEA increased sensitivity, and classified 82.2% of the NSCLC patients as positive while only 7.7% of healthy

volunteers were falsely diagnosed to be positive. The use of both DKK1 and ProGRP increased sensitivity to detect SCLCs up to 89.4%, while false positive rate in healthy donors were only 6.3%. Our data imply that DKK1 should be useful as a novel diagnostic/prognostic biomarker in clinic and probably as a therapeutic target for lung and esophageal cancer.

2. Common diseases

(1) cerebral infarction

Authors: Michiaki Kubo[1,2,4], Jun Hata[1,2,4], Toshiharu Ninomiya[1,2], Koichi Matsuda[4], Koji Yonemoto[1], Toshiaki Nakano[2,3], Tomonaga Matsushita[2,4], Keiko Yamanaka-Yamazaki[4], Yozo Ohnishi[5], Susumu Saito[5], Takanori Kitazono[2], Setsuro Ibayashi[2], Katsuo Sueishi[3], Mitsuo Iida[2], Yusuke Nakamura[4], and Yutaka Kiyohara[1].: [1]Department of Environmental Medicine, [2]Department of Medicine and Clinical Science, [3]Pathophysiological and Experimental Pathology, Graduate School of Medical Sciences, Kyushu University, Fukuoka 812-8582, Japan. [4]Laboratory of Molecular Medicine, Human Genome Center, Institute of Medical Science, University of Tokyo, Tokyo 108-8639, Japan. [5]Laboratory for Genotyping, SNP Research Center, the Institute of Physical and Chemical Research (RIKEN), Yokohama 230-0045, Japan.

Cerebral infarction is the most common type of stroke and often causes long-term disability. To investigate the genetic contribution to cerebral infarction, we conducted a case-control study using 52,608 gene-based tag-SNPs selected from JSNP database. Here we reported that one non-synonymous SNP in a member of protein kinase C (PKC) family, *PRKCH*, was significantly associated with lacunar infarction in two independent Japanese samples ($p = 3.2 \times 10^{-7}$, crude odds ratio of 1.39). This SNP was likely to affect the PKC activity. Furthermore, a 14-year follow-up cohort study in Hisayama (Fukuoka, Japan) supported involvement of this SNP for the development of cerebral infarction ($p = 0.03$, age-and sex-adjusted hazard ratio of 2.83). We also found that PKCη was mainly expressed in vascular endothelial cells and foamy macrophages in human atherosclerotic lesions, and its expression was enhanced as the lesion type progressed. Our results support a role for *PRKCH* in the pathogenesis of cerebral infarction.

Publications

1. R. Hamamoto, F.P. Silva, M. Tsuge, T. Nishidate, T. Katagiri, Y. Nakamura and Y. Furukawa: Enhanced SMYD3 expression is essential for the growth of breast cancer cells. Cancer Science, 97: 113-118, 2006
2. K. Obama, T. Kato, S. Hasegawa, S. Satoh, Y. Nakamura, and Y. Furukawa: Overexpression of peptidyl-prolyl isomerase-like 1 is associated with the growth of colon cancer cells. Clinical Cancer Research, 12: 70-76, 2006
3. A. Iida, S. Saito, A. Sekine, A. Takahashi, N. Kamatani, and Y. Nakamura: Japanese single nucleotide polymorphism database for 267 possible drug-related genes. Cancer Science, 97: 16-24, 2006
4. T. Kikuchi, Y. Daigo, N. Ishikawa, T. Katagiri, T. Tsunoda, S. Yoshida, and Y. Nakamura: Expression profiles of metastatic brain tumor from lung adenocarcinomas on cDNA microarray. International Journal of Oncology, 28: 799-805, 2006
5. T. Mushiroda, Y. Ohnishi, S. Saito, A. Takahashi, Y. Kikuchi, S. Saito, H. Shimomura, Y. Wanibuchi, T. Suzuki, N. Kamatani, and Y. Nakamura: Association of VKORC1 and CYP2C9 polymorphisms with warfarin dose requirements in Japanese patients. Journal of Human Genetics, 51: 249-253, 2006
6. T. Suda, T. Tsunoda, N. Uchida, T. Watanabe, S. Hasegawa, S. Satoh, S. Ohgi, Y. Furukawa, Y. Nakamura, and H. Tahara: Identification of secernin 1 as a novel immunotherapy target for gastric cancer using the expression profiles of cDNA microarray. Cancer Science, 97: 411-419, 2006
7. K. Yoshiura, A. Kinoshita, T. Ishida, A. Ninokata, T. Ishikawa, T. Kaname, M. Bannai, K. Tokunaga, S. Sonoda, R. Komaki, M. Ihara, V. Saenko, A.G. Kaimovich, I. Sekine, K. Komatsu, H. Takahashi, M. Nakashima, N. Sosonkina, C.K. Mapendano, M. Ghadami, M. Nomura, D.-S. Linag, D.-K. Kim, A. Garidkhuu, N. Natsume, T. Ohta, H. Tomita, K. Hirayama, M. Ishibashi, A. Takahashi, N. Saito, S. Saitou, Y. Nakamura, and N, Niikawa: A SNP in the ABCC11 gene is the determinant of human earwax type. Nature Genetics, 38: 324-330, 2006
8. T. Yamabuki, Y. Daigo, T. Kato, S. Hayama, T. Tsunoda, M. Miyamoto, T. Ito, M. Fujita, M. Hosokawa, S. Kondo, and Y. Nakamura: Genome-wide gene expression profile analysis of esophageal squamous cell carcinomas.

International Journal of Oncology, 28: 1375-1384, 2006
9. A. Iida, H. Kizawa, Y. Nakamura, and S. Ikegawa: High-resolution SNP map of ASPN, a susceptibility gene for osteoarthritis. Journal of Human Genetics, 51: 151-154, 2006
10. F.P. Silva, R. Hamamoto, Y. Furukawa, and Y. Nakamura: TIPUH1 encodes a novel KRAB zinc-finger protein highly expressed in human hepatocellular carcinomas. Oncogene, 25: 5063-5070, 2006
11. K. Nakashima, T. Hirota, Y. Suzuki, A. Matsuda, M. Akahoshi, M. Shimizu, A. Jodo, S. Doi, K. Fujita, M. Ebisawa, S. Yoshihara,T. Enomoto, T. Shirakawa, F. Kishi, Y. Nakamura, and M. Tamari: Association of the RIP 2 gene with childhood atopic asthma. Alleology International, 55: 77-83, 2006
12. K. Nakashima, T. Hirota, K. Obara, M. Shimizu, A. Jodo, M. Kameda, S. Doi, K. Fujita, T. Shirakawa, T. Enomoto, F. Kishi, S. Yoshihara, K. Matsumoto, H. Saito, Y. Suzuki, Y. Nakamura, and M. Tamari: An association study of asthma and related phenotypes with polymorphisms in negative regulator molecules of the TLR signaling pathway. Journal of Human Genetics, 51: 284-291, 2006
13. S. Ashida, M. Furihata, T. Katagiri, K. Tamura, Y. Anazawa, H. Yoshioka, T. Miki, T. Fujioka, T. Shuin, Y. Nakamura, and H. Nakagawa: Expression of novel molecules, MICAL2-PV (MICAL2 prostate cancer variants), increases with high gleason score and prostate cancer progression. Clinical Cancer Research, 12: 2767-2773, 2006
14. N. Ishikawa, Y. Daigo, A. Takano, M. Taniwaki, T. Kato, S. Tanaka, W. Yasui, Y. Takeshima, K. Inai, H. Nishimura, E. Tsuchiya, N. Kohno, and Y. Nakamura: Characterization of SEZ6L2 cell-surface protein as a novel prognostic marker for lung cancer. Cancer Science, 97737-745, 2006
15. T. Kobayashi, T. Masaki, M. Sugiyama, Y. Atomi, Y. Furukawa, and Y. Nakamura: A gene encoding a family with sequence similarity 84, member A (FAM84A) enhanced migration of human colon cancer cells. International Journal of Oncology, 29: 341-347, 2006
16. M. Taniwaki, Y. Daigo, N. Ishikawa, A. Takano, T. Tsunoda, W. Yasui, K. Inai, N. Kohno, and Y. Nakamura: Gene expression profiles of small-cell lung cancers: molecular signatures of lung cancer. International Journal of Oncology, 29: 567-575, 2006
17. E. Hirota, L. Yan, T. Tsunoda, S. Ashida, M. Fujime, T. Shuin, T. Miki, Y. Nakamura and T. Katagiri: Genome-wide gene expression profiles of clear cell renal cell carcinoma; International Journal of Oncology, 29: 799-827, 2006
18. J.-H. Park, M.-L. Lin, T. Nishidate, Y. Nakamura, and T. Katagiri: PDZ-binding kinase/T-LAK cell-originated protein kinase, a putative cancer/testis antigen with an oncogenic activity in breast cancer. Cancer Research, 66: 9186-9195, 2006
19. K. Ozaki, H. Sato, A. Iida, H. Mizuno, T. Nakamura, Y. Miyamoto, A. Takahashi, T. Tsunoda, S. Ikegawa, N. Kamatani, M. Hori, Y. Nakamura, and T. Tanaka: A functional SNP in PSMA6 confers risk of myocardial infarction in the Japanese population. Nature Genetics, 38: 921-925, 2006
20. N. Jinawath, Y. Chamgramol, Y. Furukawa, K. Obama, T. Tsunoda, B. Sripa, C. Pairojkul, and Y. Nakamura: Comparison of gene -expression profiles between opisthorchis viverrini and non-opisthorchis viverrini associated human intrahepatic cholangiocarcinoma. Hepatology, 44: 1025-1038, 2006
21. Y. Nakamura, M. Futamura, H. Kamino, K. Yoshida, Y. Nakamura, and H. Arakawa: Identification of p53-46F as a super p53 with an enhanced ability to induce p53-dependent apoptosis. Cancer Science, 97: 633-641, 2006
22. A. Takehara, H. Eguchi, H. Ohigashi, O. Ishikawa, T. Kasugai, M. Hosokawa, T. Katagiri, Y. Nakamura, and H. Nakagawa: Novel tumor marker REG4 detected in serum of patients with resectable pancreatic cancer and feasibility for antibody therapy targeting REG4. Cancer Science, 97: 1191-1197, 2006
23. M. Iiizumi, H. Nakagawa, M. Hosokawa, A. Takehara, C. Su-Youn, T. Nakamura, T. Katagiri, H. Eguchi, H. Ohigashi, O. Ishikawa, Y. Nakamura, and H. Nakagawa: EphA4 receptor, overexpressed in pancreatic ductal adenocarcinoma, promotes cancer cell growth. Cancer Science, 97: 1211-1216, 2006
24. K. Takahashi, C. Furukawa, A. Takano, N. Ishikawa, T. Kato, S. Hayama, C. Suzuki, W. Yasui, K. Inai, S. Sone, T. Ito, H. Nishimura, E. Tsuchiya, Y. Nakamura and Y. Daigo: The neuromedin u-growth hormone secretagogue receptor 1b/neurotensin receptor 1 oncogenic signaling pathway as a therapeutic target for lung cancer. Cancer Research, 66: 9408-9419, 2006
25. S. Hayama, Y. Daigo, T. Kato, N. Ishikawa, T. Yamabuki, M. Miyamoto, T. Ito, E. Tsuchiya, S. Kondo, and Y. Nakamura: Activation of CDCA1-KNTC2, members of centromere protein complex, involved in pul-

monary carcinogenesis. Cancer Research, 66: 10339-10348, 2006
26. K. Ura, K. Obama, S. Satoh, Y. Sakai, Y. Nakamura, and Y. Furukawa: Enhanced RASGEF1A expression is involved in the growth and migration of intrahepatic cholangiocarcinoma. Clinical Cancer Research, 12: 6611-6616, 2006
27. N. Ishii, K. Ozaki, H. Sato, H. Mizuno, S. Saito, A. Takahashi, Y. Miyamoto, S. Ikegawa, N. Kamatani, M. Hori, S. Saito, Y. Nakamura and T. Tanaka: Identification of a novel non-coding RNA, MIAT, that confers risk of myocardial Infarction. Journal of Human Genetics, 51: 1087-1099, 2006
28. S. Mahasirimongkol, W. Chantratita, S. Promso, E. Pasomsab, N. Jinawath, W. Jongjaroenprasert, V. Lulitanond, P. Krittayapoositpot, S. Tongsima, P. Sawanpanyalert, N. Kamatani, Y. Nakamura, T. Sura: Similarity of the allele frequency and linkage disequilibrium pattern of single nucleotide polymorphisms in drug-related gene loci between Thai and northern East Asian populations: implications for tagging SNP selection in Thais. Journal of Human Genetics, 51: 896 -904, 2006
29. Y. Nakazaki, H. Hase, H. Inoue, Y. Beppu, M. Xin, G. Sakaguchi, R. Kurita, S. Asano, Y. Nakamura, and K. Tani: Serial analysis of gene expression in progressing and regressing mouse tumor implicates the involvement of RANTES andTARC in antitumor immune responses. Molecular Therapy, 14: 599-606, 2006
30. M. Kato, A. Sekine, Y. Ohnishi, T.A. Johnson, T. Tanaka, Y. Nakamura and T. Tsunoda: Linkage disequilibrium of evolutionarily conserved regions in the human genome. BMC Genomics, 7: 326, 2006

Human Genome Center

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

Professor	Kenta Nakai, Ph.D.	教　授　理学博士	中　井　謙　太
Associate Professor	Kengo Kinoshita, Ph.D.	助教授　理学博士	木　下　賢　吾

The mission of our laboratory is to conduct computational ("in silico") studies on the functional aspects of genome information. Roughly speaking, genome information represents what kind of proteins/RNAs are synthesized on what conditions. Thus, our study includes the structural analysis of molecular function of each gene product as well as the analysis of its regulatory information, which will lead us to the understanding of its cellular role represented by the networks of inter-gene interaction.

1. Conservation of regulation systems in firmicutes

Nicolas Sierro and Kenta Nakai

Due to the probable co-regulation by a common transcription factor of genes showing a similar expression profile, investigation of their promoter regions is an important step towards the understanding of global cell regulation networks. By coupling the current knowledge about experimentally proven transcriptional regulation with the currently available raw genetic data, a better understanding of the similarities and differences between various species could be obtained. Most of the bacterial transcriptional regulation data have however been obtained in two specific organisms, *Escherichia coli* and *Bacillus subtilis*, and comparative genomics is therefore necessary to evaluate to which extent the acquired knowledge is applicable to other bacterial species. Therefore, the annotated proteins of 66 complete firmicutes genomes, including that of *B. subtilis*, were compared to each other in order to build clusters of homologous proteins and their upstream intergenic regions. Two approaches were then used to analyze these upstream intergenic regions: the mapping of known *B. subtilis* transcription factor binding sites on every genome using the position specific weight matrices provided by DBTBS, and the analysis of the upstream intergenic region conservation for each cluster of homologous genes.

To provide a comprehensive yet easy to understand and interpret representation of the known motif conservation pattern, new tools were developed that can generate a graphic for each cluster indicating on one side in which stains homologous proteins are found, and on the other side whether or not these proteins possess a certain transcription factor binding site in their upstream intergenic region. With this method, the existence of different regulation systems for the CtsR and HrcA heat shock response regulons within the firmicutes could for instance be shown. Our data suggest for example that the mollicutes, which are characterized by a very small genome size and lack CtsR, have placed the regulation of genes typically regulated by CtsR under the control of HrcA, or that in the Staphylococcus strains the HrcA regulons is not only regulated by HrcA itself, but also by CtsR.

The analysis of the upstream intergenic region conservation for each cluster of homologous

genes is also of interest because it is expected that regions involved in gene regulation are more conserved than regions with no particular function. To investigate this conservation, each cluster was therefore divided in subclusters based both on the bacterial genus and the size of the intergenic regions and aligned by ClustalW. The aligned sequences were used to calculate the sequence conservation and position specific weight matrices generated for the conserved regions. By comparing these matrices to each other, motifs corresponding to both known transcription factor binding sites and unknown conserved regions could be obtained; the analysis of the latter group being currently underway.

By carrying out comparative analysis of a large number of related genomes, concentrating particularly on the conservation of the promoter regions of homologous genes, and using the experimental data available in literature, significant differences in the regulatory networks of not only a single strain, but of whole genus could be highlighted. This approach will therefore not only allow a refinement of the current understanding of bacterial regulation networks, but also provide new input for experimental research by helping in the design of experiments and the interpretation of their results.

2. Construction of a promoter model for tissue-specific expression in *Ciona intestinalis*

Alex Vandenbon, Takehiro Kusakabe[1], and Kenta Nakai: [1]Graduate School of Life Science, University of Hyogo

Transcriptional regulation of gene expression in eukaryotes is controlled by transcription factors binding *cis*-regulatory elements in regulatory sequences. Genes containing binding sites for the same set of transcription factors in their non-coding regions can be expected to be co-regulated at the transcriptional level. However, using merely the presence of these binding sites to predict new target genes for tissue-specific expression might result in high numbers of false positives, as the computational prediction of these binding sites has been shown to suffer from very bad specificity. To increase accuracy of the prediction of new target genes we constructed a promoter model that does not only take into account the presence of *cis*-regulatory elements, but also their order in the regulatory sequence and the distances between pairs of elements. We trained this model on 5 sets of tissue-specifically expressed genes of *C. intestinalis* and used it to predict new target genes in a genome-wide set of promoter sequences. Blast results for high-scoring genes indicate that this promoter model might be useful for predicting promising candidates for wet-lab experiments.

3. Analysis of trans-splicing in *Ciona intestinalis*

Li Shuang, Riu Yamashita, Takehiro Kusakabe[1] and Kenta Nakai

Trans-splicing, in which the original 5′-ends of some pre-mRNAs are discarded and replaced by the 5′-region of a certain SL RNA, has been reported in *Ciona intestinalis*. We analyzed this phenomenon using its EST and genome sequences. A conserved head sequence, which is called Spliced-Leader (SL), ATTCTATTTGAATA AG, and could not be mapped to the genome sequence, has been found in 419 of 2,077 5′-ESTs. Depending on the existence of this Spliced -Leader sequence, we classified the 5′-ESTs into 2 groups: SL+ and SL−. Using the UniGene database, the gene names of each SL+ and SL− groups were obtained. The fact that there were only small overlaps between these groups suggests that trans-splicing occurs on a specific gene set. By regarding the number of EST clones for each gene as its expression strength, we found that the expression level of SL+ genes is significantly high while that of SL− genes varies greatly. In addition, by regarding the number of libraries where a gene's EST(s) are observed as a degree of anti-tissue/developmental stage specificity, we found that the gene expression between SL+ and SL− groups shows a significant difference at the developmental stage of juvenile. We further used the Gene Ontology information of human homologous genes to annotate corresponding Ciona genes. As a result, we found that a significant number of SL− genes seem to be related to ribosomal functions while mitochondrion-related genes are found more frequently in SL+ genes.

4. Updating the database of transcriptional start sites (DBTSS)

Riu Yamashita, Yutaka Suzuki[2], Sumio Sugano[2], and Kenta Nakai: [2]Graduate School of Frontier Sciences

DBTSS (http://dbtss.hgc.jp) was constructed in 2002 based on experimentally-determined 5′-end clones. During this year, we have added several features. First, the number of clones corresponding to human genes has significantly increased, from 190,964 to 1,359,000. Second, we defined putative promoter groups by clustering TSSs within a 500 base range because the con-

tent of our database is now large enough to analyze the existence of multiple promoters for each gene. If a gene has several putative TSS clusters, we regard them as the evidence of alternative promoters. We found 8,308 human genes and 4,276 mouse genes which have alternative promoters. Third, DBTSS now supports a function that enables detailed comparison between any pair of TSSs present in it. Finally, we have added TSS information of zebrafish (15,189 TSSs: 32,263 clones), malaria (6,908 TSSs: 10,236 clones), and schyzon (14,029 TSSs: 22,923 clones).

5. Comprehensive detection of human terminal oligo-pyrimidine (TOP) genes and analysis of their characteristics

Riu Yamashita, Yutaka Suzuki[2], Nono Tomita-Takeuchi[2], Hiroyuki Wakaguri[2], Takuya Ueda[2], Sumio Sugano[2], and Kenta Nakai

It is known that several genes have a terminal oligo-pyrimidine sequence at the 5′-end of their mRNA, and are hence called TOP genes. These genes are also known to be transcriptionally or translationally regulated. But it is not known how many of these genes are present in the human genome. We performed a detection of TOP gene candidates using accurate TSS information provided by DBTSS. By using a position specific weight matrix constructed from 48 known TOP genes, we could detect 1,645 candidate TOP genes from the 13,717 human genes in our database. These 1,645 genes were broadly expressed compared with the rest of genes ($p<e-200$). 239 of these 1,645 genes satisfied the same criteria for TOP genes also in their mouse homologs. We experimentally validated 83 of the 239 candidates, and found 41 (49%) of them are translationally regulated. We also suggest that this translational regulation is affected by the length of mRNAs.

6. Comparative studies of alternative promoters of human and mouse genes

Katsuki Tsuritani[3], Takuma Irie[2], Riu Yamashita, Yuta Sakakibara[2], Hiroyuki Wakaguri[2], Akinori Kanai[2], Junko Mizushima-Sugano[2], Sumio Sugano[2], Kenta Nakai, and Yutaka Suzuki[2]: [3]Taisho Pharmaceutical, Co. Ltd.

It gradually becomes clear that a large population of human genes are regulated by more than one alternative promoters (APs). Here we report large-scale comparative studies of putative alternative promoters (PAPs) between human and mouse counterpart genes. We classified the 17,245 putative promoter regions (PPRs) in 5,764 PAP-containing human genes into three categories: "conserved", "marginal" and "non-conserved" according to the results of sequence comparison. The conserved PPRs are major and rich in CpG-island. In contrast, the non-conserved PPRs are minor and rich in repetitive elements. The latter also are similar to intergenic region in the distribution of CG content, and they produce more transcripts that encode small or no proteins than the conserved ones. Systematic luciferase assays of these PPRs revealed that both classes of PPRs did have promoter activity, but that their strength ranges were significantly different. Furthermore, we demonstrated that these characteristic features of the non-conserved PPRs are shared with the PPRs of previously discovered putative non-protein coding transcripts. Taken together, our data suggest that there are two distinct classes of promoters in humans, with the latter class of promoters emerging frequently during evolution.

7. Relationships between protein conservation and promoter conservation revealed by comparative sequence analysis between human and mouse

Hirokazu Chiba, Riu Yamashita, Kengo Kinoshita, and Kenta Nakai

Comparative sequence analysis is a powerful tool to extract functional or evolutionary information from genomes of organisms. With the use of complete genome sequences, there have been many studies comparing protein sequences or promoter sequences to provide insights into genomics. However, the relationships between protein conservation and promoter conservation are poorly understood. In this study, we compared not only protein sequences but also promoter sequences for 6,901 human and mouse orthologous genes to address this issue. First, we carried out the comparison of promoter sequences, and examined the relationship between gene function and promoter conservation. New functional categories with significant promoter conservation levels were identified in addition to the ones already reported previously. Next, the relationship between protein conservation and promoter conservation was examined. The correlation between them was weak, suggesting that protein sequences and promoter sequences are under different kinds of evolutionary pressures. Specifically, the 'ribosome' category shows significantly low promoter conservation, in spite of high protein conservation; while inversely, the 'extracellular matrix' category shows significantly high promoter conservation, in

spite of low protein conservation.

8. MLV integration sites analysis using DBTSS

Yoshiaki Tanaka, Riu Yamashita, and Kenta Nakai

Historically, it was thought that the integration sites of retroviruses were at random. Recently some researchers compared integration sites with the NCBI reference sequences (RefSeq) database, and reported Murine Leukemia Virus (MLV) is inserted preferentially close to transcription start sites (TSSs). However, these results rely on the TSS information in RefSeq, and it is known that many RefSeq genes do not have an accurate 5′ end. To obtain precise correlation between MLV integration sites and TSSs, we compared MLV integration sites with the Database of Transcription Start Sites (DBTSS), which covers precise TSSs. From this result, we could obtain a clear peak in TSS±2kb, which was narrower than previous reports. The integration frequency in TSS±5kb had a higher percentage than previous reports. Now we are analyzing the features of MLV integration sites using this result.

9. Computational analysis of microRNA recognition sites

Keishin Nishida, Riu Yamashita, Kengo Kinoshita, and Kenta Nakai

microRNAs (miRNAs) are ~22-nucleotide-long RNAs responsible for posttranscriptional regulation of genes by pairing with mRNA. It is known that the miRNA 5′-terminal region recognition of the mRNA 3′-untranslated region leads to translation inhibition or mRNA cleavage. This miRNA region is called "seed". However, although recent research increases the seed importance, concrete seed position and length are not defined. Our research detects the detail of seed position and length from an experimentally supported miRNA-target dataset. Those data indicate seed tendencies for translation downregulation and mRNA cleavage are different. We apply this analysis to the public microarray dataset. Two datasets indicate clear seeds. Many datasets do not have clear seeds, indicating a dependency on the quality of the microarray.

10. ATTED-II: a database of co-expressed genes and *cis*-elements for identifying co-regulated gene groups in *Arabidopsis*

Takeshi Obayashi, Kengo Kinoshita, Kenta Nakai, Masayuki Shibaoka[4], Shinpei Hayashi[4], Motoshi Saeki[4], Daisuke Shibata[5], Kazuki Saito[6], and Hiroyuki Ohta[4]: [4]Tokyo Institute of Technology, [5]Kazusa DNA Research Institute, [6] Chiba University

Publicly available database of co-expressed gene sets would be a valuable tool for a wide variety of experimental designs. We constructed an *Arabidopsis thaliana* trans-factor and cis-element prediction database (ATTED-II) that provides co-regulated gene relationships based on co-expressed genes deduced from microarray data and the predicted cis-elements. ATTED-II (http://atted.hgc.jp) includes the following features: (i) lists and networks of co-expressed genes calculated from 58 publicly available experimental series (1,388 GeneChip data) in *A. thaliana*; (ii) prediction of cis-regulatory elements to predict co-regulated genes amongst the co-expressed genes; and (iii) visual representation of expression patterns for individual genes. ATTED-II can thus help researchers to clarify the function and regulation of particular genes and gene networks.

11. COXPRES: co-expressed gene database for mouse and human

Takeshi Obayashi and Kengo Kinoshita

The number of publicly available gene expression data is abundant in mouse and human, and is ten or twenty times larger than that for Arabidopsis. However there is no database of co-expressed genes such as ATTED-II, although the information is very valuable to predict gene function. We are thus constructing a new database named COXPRES (co-expression) for co-expressed genes in mouse and human from such publicly available gene expression data. The information of gene co-expression is calculated from thousands of oligonucleotide microarray (GeneChip) data and then represented as gene lists and gene networks. This information of gene co-expression will widely promote experimental researches on mouse and human.

12. Assessing gene similarity with their co-expression and its use in prediction of gene function

Takeshi Obayashi, Atsushi Takabayashi[7], Noriko Ishikawa[7], Fumihiko Sato[7], and Kengo

Kinoshita: [7]Kyoto University

The information of gene co-expression is valuable to predict gene functions. The performance of gene prediction depends on the quality of the gene co-expression data and the method of prediction. The quality of gene co-expression data depends on not only the quality and quantity of the original data but also on the normalizing method and the method used to calculate correlation coefficient between gene expression patterns. To compare these methods, we developed a new system to automatically predict gene function from its co-expression data. Using this system, we assessed the methods to construct gene co-expression data. In addition to computational prediction, the methods are compared by large-scale gene disruption.

13. Analyses of homo-oligomer interfaces of proteins from the complementarity of molecular surface, electrostatic potential and hydrophobicity

Yuko Tsuchiya, Kengo Kinoshita, and Haruki Nakamura[8]: [8]Osaka University

To extract the structural interacting patterns between proteins, we developed a method to estimate the complementarities for hydrophobicity, electrostatic potential on the molecular surfaces and shape of the surfaces in protein-protein interfaces. We have found that the homo -oligomer interfaces can be classified into five groups according to the structure of the interface; cyclic-oligomer, twisted-dimer, dimer-parallel, dimer-perpendicular and dimer-circular. As the results of correlation analyses between the shape classification and the property complementary, new characteristic trends as the possible necessary conditions of protein-protein interactions are emerging.

14. PreBI: prediction of biological interfaces of proteins in crystals

Yuko Tsuchiya, Kengo Kinoshita, Nobutoshi Ito[9], and Haruki Nakamura[8]: [9]Tokyo Medical and Dental University

PreBI is a WWW server for predicting biological interfaces in protein crystal structures according to the complementarities of the electrostatic potential, hydrophobicity and shape of the interfaces, along with the area of the interfaces. The results can be checked through our interactive viewer. (http://pre-s.protein.osaka-u.ac.jp/~prebi/)

15. Probabilistic alignment detects remote homology in a pair of protein sequences without homologous sequence information

Ryotaro Koike[10], Kengo Kinoshita, and Akinori Kidera[10]: [10]Yokohama City University

Dynamic programming algorithm and its heuristics are the fundamental methods for similarity searches of biological sequences. Including additional information, such as homologous sequences in the profile comparison, has refined the detection power of sequence comparison. We described a new approach, probabilistic alignment (PA), which gives improved detection power using only a pair of amino acid sequences. Receiver operating characteristic (ROC) analysis showed that the PA method is far superior to BLAST, and that its sensitivity and selectivity approach to those of PSI-BLAST. Particularly for orphan proteins, PA exhibits much better performance than PSI-BLAST.

16. Prediction of disordered region and tertiary structure of proteins from its amino acid sequence

Takashi Ishida and Kengo Kinoshita

Identification of protein intrinsically disordered or unstructured regions is useful for protein structure determination and protein tertiary structure prediction. We developed a method to predict protein-disordered regions from its amino acid sequence. We achieved high prediction accuracy by combining the prediction from local sequence information and the prediction from global sequence alignment. At the same time, prediction of protein tertiary structure from amino acid sequence without the information of structure of homologues is still a big problem. We developed a new potential function based on contact number prediction to evaluate the matching between an amino acid sequence and a tertiary structure, and applied this potential to protein tertiary structure prediction.

17. Int-surf: a database for interacting sites for protein-protein interaction

Miho Higurashi, Takashi Ishida, and Kengo Kinoshita

Int-surf is a database of interacting sites of proteins. For all PDB entries, interacting sites with other proteins, small molecules and DNA are calculated from the atomic coordinates in PDB. The information of the interacting sites of

homologous proteins is mapped onto each protein structure, so users can see possible interaction sites of the considering protein, when its complex structure is not available, if the complex structures of close homologues are available. Using jV version 3, an interactive 3D viewer program, Int-surf allow users to observe the interactions with interactive visualization.

18. Molecular dynamics study of lipid membranes containing cholesterol

Naoya Fujita, Takashi Ishida, and Kengo Kinoshita

Cholesterol, as a component of raft structures which are important in signal transduction, protein transport, etc., is a fundamental molecule in most eukaryotic cell membranes. A Molecular dynamics simulation was applied to dipalmitoylphosphatidylcholine (DPPC) with several concentrations of cholesterol. Comparing a pure phospholipid membrane and a cholesterol-containing one suggests that sterol modifies the membrane to a higher order phase. These mixed bilayers could be useful as surrounding systems for eukaryotic membrane proteins.

Publications

Horton, P., Park, K.-J., Obayashi, T., and Nakai, K. Protein subcellular localization prediction with WoLF PSORT, In Proc. 4th Asia-Pacific BIOINFORMATICS Conference (APBC2006) Edited by Jiang, T. et al. (Imperial College Press), pp. 39-48, 2006.

Kimura, K., Watanabe, A., Suzuki, Y., Ota, T., Nishikawa, T., Yamashita, R., Yamamoto, J., Sekine, M., Tsuritani, K., Ishii, S., Sugiyama, T., Saito, K., Isono, Y., Irie, R., Kushida, N., Yoneyama, T., Otsuka, R., Kanda, K., Yokoi, T., Kondo, H., Wagatsuma, M., Murakawa, K., Ishida, S., Ishibashi, T., Takahashi-Fujii, A., Tanase, T., Nagai, K., Kikuchi, H., Nakai, K., Isogai, T., and Sugano, S. Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes, Genome Res., 16: 55-65, 2006.

Sierro, N., Kusakabe, T., Park, K.-J., Yamashita, R., Kinoshita, K., and Nakai, K. DBTGR: a database of tunicate promoters and their regulatory elements, Nucl. Acids Res., 34: D552-D 555, 2006.

Yamashita, R., Suzuki, Y., Wakaguri, H., Tsuritani, K., Nakai, K., and Sugano, S. DBTSS: Database of Human Transcription Start Sites, Progress Report 2006, Nucl. Acids Res., 34: D 86-D89, 2006.

Cheong, J., Yamada, Y., Yamashita, R., Irie, T., Kanai, A., Wakaguri, H., Nakai, K., Ito, T., Saito, I., Sugano, S., and Suzuki, Y. Diverse DNA methylation status at alternative promoters of human genes in various normal tissues, DNA Res., 13(14): 155-167, 2006.

Tsuchiya, Y., Kinoshita, K., and Nakamura, H. Analysis of homo-oligomer interfaces of proteins from the complementarity of molecular surface, electrostatic potential and hydrophobicity, Protein Eng Des Sel., 19: 421-429, 2006.

Tsuchiya, Y., Kinoshita, K., Ito, N., and Nakamura, N., PreBI: Prediction of biological interfaces of proteins in crystals, Nucl Acid Res, 34: W320-324, 2006.

Kinoshita, K., Kusunoki, M., and Miyai, K. Analysis of the three-dimensional structure of the "CXGXC" motif in the CMGCC and CAGYC regions of alpha-and beta-subunits of human chorionic gonadotropin: Importance of Glycine residue in the motif, *Endocrine J.,* **53**, 51-38, 2006.

Ishida, T., Nakamura, S., and Shimizu, K. Potential for assessing quality of protein structure based on contact number prediction. *Proteins* 64: 940-947, 2006.

中井謙太，生物学の未来を考える，蛋白質核酸酵素　51(12)：1704−1707, 2006.

中井謙太，第８章：バイオインフォマティクス～ホモロジー解析からシステム生物学まで～，遺伝子工学集中マスター，山本雅・仙波憲太郎編集，羊土社，pp. 102−111, 2006.

木下賢吾，eF-siteを利用した機能部位の予測，バイオテクノロジージャーナル, 6：444−447, 2006.

Human Genome Center

Laboratory of Biostatistics (Biostatistics Training Unit)

バイオスタティスティクス人材養成ユニット

Project Lecturer	Rui Yamaguchi, Ph.D.	特任講師　博士(理学)	山　口　　類
Project Research Associate	Atsushi Doi, Ph.D.	特任助手　博士(理学)	土　井　　淳

The main projects of our laboratory are to discover new biological meaning at molecular level from vast array of biological data by various statistical approaches, and to train the researchers for the right use of statistical techniques. The subjects under investigation are as follows: Extraction of useful information from time-course gene expression data by using time-series models, development of parameter estimation methods for bio-simulation models with statistical approaches, and building bio-pathway models with wide variety of background knowledge.

1. Extraction of Useful Information from Time Course Gene Expression Data with Statistical Time Series Models

a. State-space Approach with the Maximum Likelihood Principle to Identify the System Generating Time-course Gene Expression Data of Yeast

Rui Yamaguchi and Tomoyuki Higuchi[1]: [1]Institute of Statistical Mathematics

We use linear Gaussian state-space models to analyze time-course gene expression data. They are modeled to be generated from hidden state variable in a system. To identify the system, we estimate parameters of the model by EM algorithm and determine dimension of the state variable by BIC. We apply this method to a published cell-cycle gene expression data of yeast (*Saccharomyces cerevisiae*). The determined dimension of the state variable are compared with those reported in other papers, which were obtained by other parameter estimation methods and criteria.

b. Finding Module-Based Gene Networks with State-Space Models

Rui Yamaguchi, Ryo Yoshida[2], Seiya Imoto[2], Tomoyuki Higuchi[2], Satoru Miyano[2]: [2]Human Genome Center, Laboratory of DNA Information Analysis

We discuss the use of the state space models to analyze time-course microarray gene expression data. Typical features of time-course microarray data are short time-course and high-dimensional observational vector. By these aspects, conventional statistical models such as the multivariate autoregressive models lead to unsuitable results due to the overfitting. The state space models have the potential to overcome this problem by the dimension reduction process. This study provides (1) a survey of the state space models, (2) a review of some existing researches using the state space models for time-course microarray data together with a biological meaning of the model, (3) a solution for the lack of parameter identifiability, and (4) identification of biological system that is a central topic

in bioinformatics. Finally, we show the usefulness of the state space models for time-course microarray data by the analysis of *Saccharomyces cerevisiae* cell cycle gene expression data. As a result, we estimated the number of the gene modules, and a gene-module network.

2. Development of Parameter Estimation Methods for Bio-Simulation Models with Statistical Approaches

a. Genomic Data Assimilation for Estimating Hybrid Functional Petri Net from Time-Course Gene Expression Data

Masao Nagasaki[2], Rui Yamaguchi, Ryo Yoshida[2], Seiya Imoto[2], Atsushi Doi, Yoshinori Tamada[1], Hiroshi Matsuno[3], Satoru Miyano[2], and Tomoyuki Higuchi[1]: [3]Fuculty of Science, Yamaguchi University

For simulation models of biological pathways, in most of cases, parameters to govern them are tuned by experts empirically at the present time. In this study we take a novel approach for this area, in order to estimate these parameters and to select the best model from several candidate ones by using observed data. The approach is called the data assimilation (DA) of which the concept is to incorporate information from observed data into a simulation model. We can expect to obtaine more plausible results from the simulation model by this approach. From the point of view of the statistical modeling, DA can be realized by solving an inverse problem to estimate unknown state and parameters of a simulation model by using observed data. To formulate the problem, we use a statistical time series model which is called a nonlinear state space model (SSM). Using an SSM, we can employ effective statistical methods to estimate parameters, i.e. a particle filter, which is based on Monte Carlo simulation. In order to examine the applicability of the approach for biological simulation models, the methods was applied to a model of circadian rhythm represented by a hybrid functional Petri net (HFPN) with a synthesized data.

3. Building Bio-Pathway Models with Background Knowledge

a. Simulation-Based Validation of the p53 Transcriptional Activity with Hybrid Functional Petri Net

Atsushi Doi, Masao Nagasaki[2], Hiroshi Matsuno[3], and Satoru Miyano[2]

MDM2 and p19ARF are essential proteins in cancer pathways forming a complex with protein p53 to control the transcriptional activity of protein p53. It is confirmed that protein p53 loses its transcriptional activity by forming the functional dimer with protein MDM2. However, it is still unclear that protein p53 keeps its transcriptional activity when it forms the trimer with proteins MDM2 and p19ARF. We have observed mutual behaviors among genes p53, MDM2, p19ARF and their products on a computational model with hybrid functional Petri net (HFPN) which is constructed based on information described in the literature. The simulation results suggested that protein p53 should have the transcriptional activity in the forms of the trimer of proteins p53, MDM2, and p19ARF. This paper also discusses the advantages of HFPN based modeling method in terms of pathway description for simulations.

b. A Combined Pathway to Simulate CDK-Dependent Phosphorylation and ARF-Dependent Stabilization for p53 Transcriptional Activity

Atsushi Doi, Masao Nagasaki[2], Hiroshi Matsuno[3], and Satoru Miyano[2]

The protein p53 is phosphorylated by a member of protein kinases such as CDK7, and stabilized by the protein ARF. The phosphorylation and stabilization of p53 is believed to enhance its transcriptional activity and act simultaneously. Biological pathways composed of experts knowledge obtained from the literature are including these activation mechanisms. However, the map of biological pathways does not reflect the combination effect of phosphorylation and stabilization. We have conducted some simulations of biological pathways with hybrid functional Petri net (HFPN) after careful reading of papers. In this paper, we constructed the HFPN based biological pathway of CDK-dependent phosphorylation pathway and combine with ARF-dependent pathway described previously, to observe the effect of the phosphorylation on the stabilization with simulation-based validation.

Publications

Doi, A., Nagasaki, M., Matsuno, H., and Miyano, S., Simulation based validation of the p53 transcriptional activity with hybrid functional Petri net, In Silico Biol., 6(0001), 2006. <http://www.bioinfo.de/isb/2006/06/0001>.

Doi, A., Nagasaki, M., Matsuno, H., and Miyano, S., A combined pathway to simulate CDK-dependent phosphorylation and ARF-dependent stabilization for p53 transcriptional activity, Genome inform., 17(1): 112-123, 2006.

Nagasaki, M., Yamaguchi, R., Yoshida, R., Imoto, S., Doi, A., Tamada, Y., Matsuno, H., Miyano, S., and Higuchi, T., Genomic Data Assimilation for Estimating Hybrid Functional Petri Net from Time-Course Gene Expression Data, Genome Informatics (IBSB2006), 17(1), 46-61, 2006.

Yamaguchi, R., and Higuchi, T., State-space Approach with the Maximum Likelihood Principle to Identify the System Generating Time-Course Gene Expression Date of Yeast, International Journal of Data Mining and Bioinformatics, 1(1): 77-87, 2006.

Human Genome Center

Department of Public Policy
公共政策研究分野

Associate Professor Kaori Muto, Ph.D.

助教授 武 藤 香 織

Department of Public Policy has launched since September 2007 as a new and the first social scientific section on medical sciences featuring genomics. We value conducting empirical studies in many fields of social science for future policy making. We work for three major missions; public policy studies on translational research, its application to healthcare and its impact on social security, practical advices and survey for research projects to build public trust, and "minority-centered" scientific communication.

1. Public policy studies on translational research, its application to healthcare and its impact on social security

Exploring the human genome give us lots of critical keys to approach human health and disease and to develop new diagnostic tools and preventive methods. However, ethical, legal and social aspects have been featured since the launch of the Human Genome Project. Various ethical concerns have been raised concerning these scientific quest as well as expectation for change. Ethical discussions, however, tend to go before reviewing empirical evidences and facts. Our major mission is to conduct empirical analysis by interdisciplinary approach for future public discussions. We collaborate with experts of several fields of social sciences; e.g. sociology, anthropology, psychology, disability studies, economics, finance, policy science, law, marketing, STS (science, technology and society) and so on. One of the urgent issues should be an empirical analysis of social and financial aspects of personalized medicine. We're examining social and financial validity and draw regulatory framework of clinical introduction of drug sensitivity genetic testing, by estimating saving costs and impact on healthcare insurance system. Also we're planning to conduct surveys towards research participants and candidates to evaluate the consent process of ongoing genomic studies to suggest future change or correction of ethical guidelines for genomic research.

2. Practical advices and survey for research projects to build public trust

Medical scientists sometimes feel anxiety about protection of human participants and ethical issues. We help them to picture the ethically-valid roadmap towards their goals and suggest regulatory scheme for social responsibility of science, which includes creating consent forms and brochures with high readability, considering procedures of informing participants and obtaining consents, considering ethical balance of the whole protocol before applying ethical review boards, and public disclosure throughout the research process.

3. "Minority-centered" scientific communication

Japanese government emphasizes the importance of promoting communication between researchers, engineers and society in Science and Technology Basic Plan (2006). In a word "society", we have to imagine various people with

diverse values. We will organize a range of public engagement projects and science participation events and particularly we focus on "minority-centered" scientific communication, empowered by disability studies and gay/lesbian studies. We feature perspectives from people with chronic disorders, the disabled/challenged, LGBT/queers and ethnic minorities. These people are not exactly "minority" in our society, but their claims towards science are not easy to be heard. We conduct surveys and suggest policy agenda how science and policy works throughout these practices.

Collaborating with modern artists, we think future of science and society. One of our related projects is "Delivery by Male Project" (2004-) with Hiroko Okada, which features ethical and social issues of male pregnancy using assisted reproductive technologies.