Human Genome Center

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

Professor	Minoru Kanehisa, Ph.D.	教　授	理学博士	金　久　　實
Research Associate	Toshiaki Katayama, M.Sc.	助　手	理学修士	片　山　俊　明
Research Associate	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 genome sequences of over 300 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 several open source softwares 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." The KEGG DAS is weekly updated and freely available at http://das.hgc.jp/.

2. SOAP/WSDL interface for the KEGG system

Shuichi Kawashima, Toshiaki Katayama 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 networks in biological processes (PATHWAY database), 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. Recent key changes include the following: (1) Several new methods to retrieve the information concerning LIGAND and GLYCAN database have been added. (2) Methods to retrieve the information concerning KO (KEGG Orthology) have been re-organized. (2) bconv method have been added to exchange external database ID (e.g. NCBI Gene-ID) to the KEGG ID. (3) Methods to utilize "entry" element information in KGML (KEGG Markup Language) have been added. (4) C# programming language client was supported. The KEGG API is available at http://www.genome.jp/kegg/soap/.

3. High performance database 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 49,498,755 (Release 150). 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, OMIIM, 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.

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 the 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 publicly available and provides the community 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. SSS: a new 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 utilize 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/.

6. Integrative analysis of chemical and genomic information on the 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. In order to make full use of the potential of natural products as research tools as well as drug leads on the context of synthetic biology, it is of great importance to understand the biosynthetic strategies with the integrative computational analyses of chemical and genomic information. An increasing number of such information become available to allow us to extract the design principles of natural products coded on genomic information. Natural products are composed of a series of molecular building blocks, such as fatty acids and amino acids, which can be regarded as minimal units hierarchically organized into the biosynthetic circuits. We define molecular building blocks required for describing the chemical information of natural products. Each natural product is then 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 develop a computational method to extract distinctive network structures consisted of both chemical and genomic information, which will be useful for designing the biosynthetic circuits to help metabolic engineering of novel bioactive compounds.

7. Development of algorithms for biosynthetic process analysis

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

We are developing algorithms for identifying biosynthetic process of some given medicinal products by utilizing the database of identified building blocks in biosynthetic processes. Given a set of building block graphs, the problem is to find the most reasonable decomposition of a graph where each decomposed subgraph is the same or similar to some of the building block graphs. 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.

8. Development of algorithms for protein structure indexing

Tetsuo Shibuya

Protein structure analysis is one of the most important research issues in the post-genomic era, and faster and more accurate query 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. There are many efficient indexing structures for strings, but it has been considered very difficult to design such sophisticated data structures against 3-D structures like proteins. By using the data structure, we can search efficiently for all of their substructures whose RMSD (root mean square distance) or URMSD (unit-vector root mean square distances) to some given query 3-D structure are not larger than a given bound. Our data structure can be stored in O(n) space, where n is the sum of lengths of the set of proteins. We propose an efficient construction algorithm for it and a quasi-linear time search algorithm. Further algorithms for more flexible structure searching/function prediction/clustering/motif finding based on the data structure is also under development.

9. Construction of a knowledge base for tracing drug evolution

Michihiro Araki and Minoru Kanehisa

Current drugs are mostly derived by modifi-

cation of known drug structures or from lead structures to be optimized for targeting new molecules or obtaining improved efficacy. Recent computational approaches have been analyzing the chemical properties of drugs, lead compounds and seed compounds to provide different kinds of chemical rules for 'drug-likeness'. The chemical rules based on the chemical property distributions are very useful for filtering drug-like molecules out of empirically synthesized chemical libraries, but do not really explain how to modify known drugs or lead structures to new drug candidates. To explore a design rule in drug development, it is necessary to focus on the empirical modification processes to construct a knowledge base for tracing the chemical evolutions. We start collecting data on the drug evolutions from databases and literatures, and part of the data has already been implemented on the drug structure maps in the KEGG DRUG database.

Publications

Tamori A, Yamanishi Y, Kawashima S, Kanehisa M, Enomoto M, Tanaka H, Kubo S, Shiomi S, Nishiguchi S. Alteration of gene expression in human hepatocellular carcinoma with integrated hepatitis B virus DNA. Clin. Cancer Res. 11: 5821-5826, 2005

Yamada, T., Kawashima, S., Mamitsuka, H., Goto, S., Kanehisa, M. Comprehensive analysis and prediction of synthetic lethality using subcellular locations. Genome Inform Ser Workshop Genome Inform. 2005, 16(1): 150-158, 2005

Okuda S, Kawashima S, Goto S, Kanehisa M. Conservation of gene co-regulation between two prokaryotes: Bacillus subtilis and Escherichia coli. Genome Inform Ser Workshop Genome Inform. 2005, 16(1): 116-124, 2005

Honda, W., Kawashima, S., Kanehisa, M. Autoimmune diseases and peptide variations. Genome Inform. Ser. Workshop Genome Inform. 2005, 16(1): 272-280, 2005

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

Okuda, S., Katayama, T., Kawashima, S., Goto, S., Kanehisa, M. ODB: a database of operons accumulating known operons across multiple genomes. Nucleic Acids Res. 34: D358-362, 2006

Hashimoto, K., Goto, S., Kawano, S., Aoki-Kinoshita, K.F., Ueda, N., Hamajima, M., Kawasaki, T., Kanehisa, M. KEGG as a glycome informatics resource. Glycobiology, in press

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

Shibuya, T., Geometric Suffix Tree: A New Index Structure for Protein 3-D Structures, IPSJ SIG Notes SIGAL 105, to appear.

Sakai, H., Murakami, H., Aburatani, S., Shibuya, T., Horimoto, K., Kanehisa, M. Bayesian Approach for Sequence Pattern Search in Tissue Specific Alternative Splicing. The 9th World Multi-Conference on Systemics, Cybernetics and Informatics, Vol. VIII: 25-30, 2005

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)　長崎正朗
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 concern is 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. Estimating gene regulatory networks and protein-protein interactions of Saccharomyces cerevisiae from multiple genome-wide data

Naoki Nariai, Yoshinori Tamada, Seiya Imoto, Satoru Miyano

Biological processes in cells are properly performed by gene regulations, signal transductions and interactions between proteins. To understand such molecular networks, we proposed a statistical method to estimate gene regulatory networks and protein-protein interaction networks simultaneously from DNA microarray data, protein-protein interaction data and other genome-wide data. We unify Bayesian networks and Markov networks for estimating gene regulatory networks and protein-protein interaction networks according to the reliability of each biological information source. Through the simultaneous construction of gene regulatory networks and protein-protein interaction networks of *Saccharomyces cerevisiae* cell cycle, we predict the role of several genes whose functions are currently unknown. By using our probabilistic model, we can detect false positives of high through-put data, such as yeast two hybrid data. In a genome-wide experiment, we find possible gene regulatory relationships and protein-protein interactions between large protein complexes that underlie complex regulatory mechanisms of biological processes.

b. Utilizing evolutionary information and gene expression data for estimating gene networks with Bayesian network models

Yoshinori Tamada, Hideo Bannai[1], Seiya Imoto, Toshiaki Katayama, Minoru Kanehisa, Satoru Miyano: [1]Kyushu University

Since microarray gene expression data do not contain sufficient information for estimating accurate gene networks, other biological information has been considered to improve the estimated networks. Recent studies have revealed that highly conserved proteins that exhibit similar expression patterns in different organisms, have almost the same function in each organ-

ism. Such conserved proteins are also known to play similar roles in terms of the regulation of genes. Therefore, this evolutionary information can be used to refine regulatory relationships among genes, which are estimated from gene expression data. We proposed a statistical method for estimating gene networks from gene expression data by utilizing evolutionarily conserved relationships between genes. Our method simultaneously estimates two gene networks of two distinct organisms, with a Bayesian network model utilizing the evolutionary information so that gene expression data of one organism helps to estimate the gene network of the other. We show the effectiveness of the method through the analysis on *Saccharomyces cerevisiae* and *Homo sapiens* cell cycle gene expression data. Our method was successful in estimating gene networks that capture many known relationships as well as several unknown relationships which are likely to be novel.

c. Estimating gene networks from expression data and binding location data via Boolean networks

Osamu Hirose, Naoki Nariai, Yoshinori Tamada, Hideo Bannai[1], Seiya Imoto, Satoru Miyano

We proposed a computational method for estimating gene networks by the Boolean network model. The Boolean networks have some practical problems in analyzing DNA microarray gene expression data: One is the choice of threshold value for discretization of gene expression data, since expression data take continuous variables. The other problem is that it is often the case that the optimal gene network is not determined uniquely and it is difficult to choose the optimal one from the candidates by using expression data only. To solve these problems, we use the binding location data produced by Lee *et al*. (Science 298: 799-804, 2002) together with expression data and illustrate a strategy to decide the optimal threshold and gene network. To show the effectiveness of the proposed method, we analyze *Saccharomyces cerevisiae* cell cycle gene expression data as an application.

d. Error tolerant model for incorporating biological knowledge with expression data in estimating gene networks

Seiya Imoto, Tomoyuki Higuchi[2], Takao Goto, Satoru Miyano : [2]Instititute of Statistical Mathematics

We proposed a novel statistical method for estimating gene networks based on microarray gene expression data together with information from biological knowledge databases. Although a large amount of gene regulation information has already been stored in some biological databases, there are still errors and missing facts due to experimental problems and human errors. Therefore, we cannot blindly use them for understanding gene regulation and a robust procedure with a statistical model for using such database information is required. By using gene expression data, we provide a probabilistic framework of a joint learning model for repairing database information and for estimating a gene network based on dynamic Bayesian networks, simultaneously. To show the effectiveness of the proposed method, we analyze *Saccharomyces cerevisiae* cell-cycle gene expression data together with KEGG information.

2. Drug Target Gene Discovery with Gene Networks

a. Computational strategy for discovering druggable gene networks from genome-wide RNA expression profiles

Seiya Imoto, Yoshinori Tamada, Hiromitsu Araki[3], Kaori Yasuda[3], Cristin G. Print[4], Stephen D. Sharnock-Jones[5], Deborah Sanders[5], Christopher J. Savoie[3], Kousuke Tashiro[1], Satoru Kuhara[1], Satoru Miyano: [3]Gene Networks International, [4]University of Auckland, [5]Cambridge University

We proposed a computational strategy for discovering gene networks affected by a chemical compound. Two kinds of DNA microarray data are assumed to be used: One dataset is short time-course data that measure responses of genes following an experimental treatment. The other dataset is obtained by several hundred single gene knock-downs. These two datasets provide three kinds of information; (i) A gene network is estimated from time-course data by the dynamic Bayesian network model, (ii) Relationships between the knocked-down genes and their regulatees are estimated directly from knock-down microarrays and (iii) A gene network can be estimated by gene knock-down data alone using the Bayesian network model. We proposed a method that combines these three kinds of information to provide an accurate gene network that most strongly relates to the mode-of-action of the chemical compound in cells. This information plays an essential role in pharmacogenomics. We illustrate this method with an actual example where human endothelial cell gene networks were generated from a

novel time course of gene expression following treatment with the drug fenofibrate, and from 270 novel gene knock-downs. Finally, we succeeded in inferring the gene network related to PPAR-α, which is a known target of fenofibrate.

b. Identifying drug active pathways from gene networks estimated by gene expression data

Yoshinori Tamada, Seiya Imoto, Kousuke Tashiro[1], Satoru Kuhara[1], Satoru Miyano

We present a computational method for identifying genes and their regulatory pathways influenced by a drug, using microarray gene expression data collected by single gene disruptions and drug responses. The automatic identification of such genes and pathways in organisms' cells is an important problem for pharmacogenomics and the tailor-made medication. Our method estimates regulatory relationships between genes as a gene network from microarray data of gene disruptions with a Bayesian network model, then identifies the drug affected genes and their regulatory pathways on the estimated network with time course drug response microarray data. Compared to the existing method, our proposed method can identify not only the drug affected genes and the druggable genes, but also the drug responses of the pathways. For evaluating the proposed method, we conducted simulated examples based on artificial networks and expression data. Our method succeeded in identifying the pseudo drug affected genes and pathways with the high coverage greater than 80%. We also applied our method to *Saccharomyces cerevisiae* drug response mircorray data. In this real example, we identified the genes and the pathways that are potentially influenced by a drug. These computational experiments indicate that our method successfully identifies the drug-activated genes and pathways, and is capable of predicting undesirable side effects of the drug, identifying novel drug target genes, and understanding the unknown mechanisms of the drug.

3. Modeling and Simulation of Biological Pathwas

a. Automatic drawing of biological networks using cross cost and subcomponent data

Mitsuru Kato, Masao Nagasaki, Atsushi Doi, Satoru Miyano

Automatic graph drawing function for biopathways is indispensable for biopathway databases and software tools. We proposed a new grid-based algorithm for biopathway layout that considers (a) edge-edge crossing, (b) node-edge crossing, (c) distance measures between nodes, as its costs, and (d) subcelluar localization information from Gene Ontology, as its constraints. For this algorithm, we newly define cost functions, devise an efficient method for computing the costs (a)-(c) by employing a matrix representing the difference between two layouts, and take a steepest descent method for searching locally optimal solutions and multi-step layout method for finding better solutions. We implemented this algorithm on Cell Illustrator which is a biopathway modeling and simulation software. The algorithm is applied to a signal transduction pathway of apoptosis induced by fas ligand. We compare our layout with that of the grid-based algorithm by Li and Kurata (Bioinformatics 21 (9): 2036. 2042, 2005). The result shows that our algorithm reduces edge-edge crossings and node-edge crossings, and solves the "isolated island problem", that is, some groups of nodes are apart from other nodes in the layout. As a result, the biological understandability of the layout is fairly improved.

c. Simulation based validation of the p53 transcriptional activity with hybrid functional Petri net

Masao Nagasaki, Atsushi Doi, Hiroshi Matsuno[6], Satoru Miyano

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.

d. A new regulatory interactions suggested by simulations for circadian genetic control mechanism in mammals

Hiroshi Matsuno[6], Shin-Ichi T. Inouye[6], Yasuki Okitsu[6], Yasushi Fujii[6], Satoru Miyano: [6]Yamaguchi University

We employed hybrid functional Petri net to analyze the circadian genetic control mechanism, which consists of loops of clock genes and generates endogenous near 24 hour rhythms in mammals. Based on the available biological data, we constructed a model and, by using Cell Illustrator, we performed computer simulations for time courses of clock gene transcription and translation. Although the initial model successfully reproduced most of the circadian genetic control mechanisms, two discrepancies remained despite wide selection of the parameters. We found that addition of a hypothetical path into the initial model successfully simulated time courses and phase relations among clock genes. This also demonstrates usefulness of hybrid functional Petri net approach to biological system analysis.

e. Prediction of debacle points for robustness of biological pathways by using recurrent neural networks

Hironori Kitakaze[7], Hiroshi Matsuno[6], Nobuhiko Ikeda[6], Satoru Miyano: [7]Oshima College of Maritime Technology, [8]Tokuyama College of Technology

Living organisms have ingenious control mechanisms in which many molecular interactions work for keeping their normal activities against disturbances inside and outside of them. However, at the same time, the control mechanism has debacle points at which the stability can be broken easily. We proposed a new method which uses recurrent neural network for predicting debacle points in a hybrid functional Petri net model of a biological pathway. Evaluation on an apoptosis signaling pathway indicates that the rates of 96.5% of debacle points and 65.5% of non-debacle points can be predicted by the proposed method.

f. Petri net modeling of biological pathways

Masao Nagasaki, Atsushi Doi, Hiroshi Matsuno[6], Satoru Miyano

We have developed a software tool called Cell Illustrator (CI) for modeling and simulating biological pathways based on the concept of Petri net together with an XML format called Cell System Markup Language (CSML) describing biological pathways for simulation. This paper shows the concepts behind CI and CSML and presents our computational strategy with them for systems biology.

4. Algorithmic and Statistical Methods for Bioinformatics

a. Prediction of transcriptional terminators in *Bacillus subtilis* and related species

Michiel J.L. de Hoon, Yuko Makita, Kenta Nakai, Satoru Miyano

In prokaryotes, genes belonging to the same operon are transcribed in a single mRNA molecule. Transcription starts as the RNA polymerase binds to the promoter and continues until it reaches a transcriptional terminator. Some terminators rely on the presence of the Rho protein, whereas others function independently of Rho. Such Rho-independent terminators consist of an inverted repeat followed by a stretch of thymine residues, allowing us to predict their presence directly from the DNA sequence. Unlike in *Escherichia coli*, the Rho protein is dispensable in *Bacillus subtilis*, suggesting a limited role for Rho-dependent termination in this organism and possibly in other Firmicutes. We analyzed 463 experimentally known terminating sequences in B. subtilis and found a decision rule to distinguish Rho-independent transcriptional terminators from non-terminating sequences. The decision rule allowed us to find the boundaries of operons in *B. subtilis* with a sensitivity and specificity of about 94%. Using the same decision rule, we found an average sensitivity of 94% for 57 bacteria belonging to the *Firmicutes phylum*, and a considerably lower sensitivity for other bacteria. Our analysis shows that Rho-independent termination is dominant for *Firmicutes* in general, and that the properties of the transcriptional terminators are conserved. Terminator prediction can be used to reliably predict the operon structure in these organisms, even in the absence of experimentally known operons.

b. ArrayCluster: an analytic tool for clustering, data visualization and module finder on gene expression profiles

Ryo Yoshida, Tomoyuki Higuchi[2], Seiya Imoto, Satoru Miyano

One significant challenge of gene expression profiling based on microarray technology is to

find unknown subtypes of several diseases at the molecular levels. This task can be addressed by grouping gene expression patterns of the collected samples in the basis of a large number of genes. Application of commonly used clustering methods to such a dataset, however, are likely to fail due to the overlearning, because the number of samples to be grouped is much smaller than the data dimension which is equal to the number of genes involved in the profiling. To overcome such difficulty, we developed a novel model-based clustering method, refereed to as the mixed factors analysis. The ArrayCluster is a freely available software to perform the mixed factors analysis. It provides us some analytic tools for clustering DNA microarray experiments, data visualization and an automatic detector of module transcriptional genes that are relevant to the calibrated molecular subtypes and so on.

c. Statistical model selection method to analyze combinatorial effects of SNPs and environmental factors for binary disease

Reiichiro Nakamichi, Seiya Imoto, Satoru Miyano

We proposed a model selection method to estimate the relation of multiple SNPs, environmental factors and the binary disease trait. We applied the combination of logistic regression and genetic algorithm for this study. The logistic regression model can capture the continuous effects of environments without categorization, which causes the loss of the information. To construct an accurate prediction rule for binary trait, we adopted Akaike's information criterion (AIC) to find the most effective set of SNPs and environments. That is, the set of SNPs and environments that gives the smallest AIC is chosen as the optimal set. Since the number of combinations of SNPs and environments is usually huge, we proposed the use of the genetic algorithm for choosing the optimal SNPs and environments in the sense of AIC. We show the effectiveness of the proposed method through the analysis of the case/control populations of diabetes, Alzheimer's disease and obesity patients. We succeeded in finding an efficient set to predict types of diabetes and some SNPs which have strong interactions to age while it is not significant as a single locus.

d. A weighted profile based method for protein-RNA interacting residue prediction

Euna Jeong, Satoru Miyano

The prediction of putative RNA-interacting residues in proteins is an important problem in a field of molecular recognition. We suggest a weighted profile based method for predicting RNA-interacting residues, which utilizes the trained neural network. Most neural networks have a learning rule which allows the network to adjust its connection weights in order to correctly classify the training data. We focus on the network weights that are dependent on the training data set and give evidence of which inputs were more influential in the network. A large set of the network weights trained on sequence profiles is analyzed and qualified. We explore the feasibility of utilizing the qualified information to improve the prediction performance for protein-RNA interaction. Our proposed method shows a considerable improvement, which has been applied to the profiles of the PSI -BLAST alignment. Results for predictions using alternative representations of profile are included for comparison.

Publications

Ando, T., Konishi, S., Imoto, S. Nonlinear regression modeling via regularized radial basis function networks. J. Statistical Planning and Inference. In press.

De Hoon, Michiel J.L., Makita, Y., Nakai, K., Miyano, S. Prediction of transcriptional terminators in *Bacillus subtilis* and related species. PLoS Computational Biology. 1(3): e25, 2005.

Doi, A., Nagasaki, M., Matsuno, H., Miyano, S. Simulation based validation of the p53 transcriptional activity with hybrid unctional Petri net. In Silico Biology. In press.

Heinrich, T., DeLisi, C., Kanehisa, M., Miyano, S. Genome Informatics 16(1) (Universal Academy Press). 2005.

Heinrich, R., Mamitsuka, H., Kanehisa, M., Miyano, S., Takagi, T. Genome Informatics 16 (2) (Universal Academy Press). 2005.

Hirose, O., Nariai, N., Tamada, Y., Bannai, H., Imoto, S., Miyano, S. Estimating gene networks from expression data and binding location data via Boolean networks. Proc. First International Workshop on Data Mining and Bioinformatics (DMBIO2005). Lecture Notes in Computer Science. 3482: 349-356, 2005.

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.

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 discovering druggable gene networks from genome-wide RNA expression profiles. Pacific Symposium on Biocomputing. 11, In press.

Imoto, S., Matsuno, H., Miyano, S. Gene networks: estimation, modeling and simulation. in R. Eils and A. Kriete (Eds.), Computational Systems Biology, Academic Press, 205-228, 2005.

Imoto, S., Tamada, Y., Savoie, C.J., Miyano, S., Analysis of gene networks for drug target discovery and validation. in J. Walker and M. Sioud (Eds.), Target Discovery and Validation (a volume of "Methods in Molecular Biology" series), Humana Press, USA. In press.

Jeong, E., Miyano, S., A weighted profile based method for protein-RNA interacting residue prediction. Transactions on Computational Systems Biology. In press.

Kato, M., Nagasaki, M., Doi, A., Miyano, S. Automatic drawing of biological networks using cross cost and subcomponent data. Genome Informatics 16(2): 22-31, 2005.

Kitakaze, H., Matsuno, H., Ikeda, N., Miyano, S. Prediction of debacle points for robustness of biological pathways by using recurrent neural networks. Genome Informatics. 16(1): 192-202, 2005.

Li, C., Suzuki, S., Ge, Q.-W., Nakata, M., Matsuno, H., Miyano, S. On modeling and analyzing signaling pathways with inhibitory interactions based on Petri net. Proc. The 2005 Internatinal Joint Conference of InCoB, AASBi and KSBI (BIOINFO 2005), 348-353, 2005.

Makita, Y., De Hoon, M.J., Ogasawara, N., Miyano, S., Nakai, K. Bayesian joint prediction of associated transcription factors in Bacillus subtilis. Pacific Symposium on Biocomputing. 10: 507-518, 2005.

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. Bioinformatics and Computational Biology, In press.

Miyano, S., Mesirov, J.P., Kasif, S., Istrail, S., Pevzner, P.A., Waterman, M.S.: Proc. 9th Annual International Conference on Research in Computational Molecular Biology (RECOMB 2005), Lecture Notes in Bioinformatics (Springer). Vol. 3500, 2005.

Nagasaki, M., Doi, A., Matsuno, H., Miyano, S. Petri net modeling of biological pathways. Proc. Algebraic Biology 2005 (Universal Academy Press). 19-31, 2005.

Nagasaki, M., Doi, A., Matsuno, H., Miyano, S. Computational modeling of biological processes with Petri net based architecture. In "Bioinformatics Technologies" (Y.P. Chen, ed) (Springer Press). 179-243, 2005.

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, In press.

Nariai, N., Tamada, Y., Imoto, S., Miyano, S. Estimating gene regulatory networks and protein-protein interactions of Saccharomyces cerevisiae from multiple genome-wide data. Bioinformatics. 21: ii206-ii212, 2005.

Ohtsubo, S., Iida, A., Nitta, K., Tanaka, T., Yamada, R., Ohnishi, Y., Maeda, S., Tsunoda, T., Takei, T., Obara, W., Akiyama, F., Ito, K., Honda, K., Uchida, K., Tsuchiya, K., Yumura, W., Ujiie, T., Nagane, Y., Miyano, S., Suzuki, Y., Narita, I., Gejyo, F., Fujioka, T., Nihei, H., Nakamura, Y. Association of a single-nucleotide polymorphism in the immunoglobulin mu-binding protein 2 gene with immunoglobulin A nephropathy. J. Hum. Genet. 50(1): 30-35, 2005.

Ott, S., Hansen, A., Kim, S.-Y., and Miyano, S. Superiority of network motifs over optimal networks and an application to the revelation of gene network evolution. Bioinformatics. 21 (2): 227-238, 2005.

Tamada, Y., Bannai, H., Imoto, S., Katayama, T., Kanehisa, M., Miyano, S. Utilizing evolutionary information and gene expression data for estimating gene networks with Bayesian network models. J. Bioinformatics and Computational Biology. 3(6): 1295-1313, 2005.

Tamada, Y., Imoto, S., Tashiro, K., Kuhara, S., Miyano, S. Identifying drug active pathways from gene networks estimated by gene expression data. Genome Informatics. 16(1): 182-191, 2005.

Yoshida, R., Higuchi, T., Imoto, S., Miyano, S. ArrayCluster: an analytic tool for clustering, data visualization and module finder on gene expression profiles. Bioinformatics. In press.

Yoshida, R., Higuchi, T., Imoto, S. Estimating time-dependent gene networks from time series DNA microarray data by dynamic linear model with Markov switching. Proc. IEEE 4th Computational Systems Bioinformatics. 289-298, 2005.

Yoshida, R., Imoto, S., Higuchi, T. A penalized likelihood estimation on transcriptional module-based clustering, Proc. First International Workshop on Data Mining and Bioinformatics (DMBIO 2005). Lecture Note in Computer Science. 3482: 389-401, 2005.

長崎正朗，土井淳，宮野悟．ダイナミックパス

ウェイモデリング言語Cell System Markup Language (CSML). タンパク質核酸酵素. 50 (16 Suppl.) : 2269–2274.
藤井靖, 松野浩嗣, 宮野悟, 井上愼一. ハイブリッド関数ペトリネットによる哺乳類の時計遺伝子機構のモデル化とシミュレーション. 時間生物学. 11(1) : 8–16, 2005.

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
Associate Professor	Toyomasa Katagiri, Ph.D.	助教授	医学博士	片桐豊雅
Associate Professor	Yataro Daigo, M.D., Ph.D.	特任助教授	医学博士	醍醐弥太郎
Assistant Professor	Hidewaki Nakagawa, M.D., Ph.D.	助手	医学博士	中川英刀
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
Assistant Professor	Ryuji Hamamoto, Ph.D.	助手	理学博士	浜本隆二

The major goal of the Human Genome Project is to identify genes predisposing to diseases, and to develop new diagnostic and therapeutic tools. We have been attempting to isolate genes involving in carcinogenesis and also those causing or predisposing to other diseases such as IgA nephropathy, and Crohn's disease. 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 associated with common diseases

a. IgA nephropathy

Shigeru Ohtsubo, Aritoshi Iida[2], Kosaku Nitta[1], Toshihiro Tanaka[3], Ryo Yamada[4], Yozo Ohnishi[3], Shiro Maeda[5], Tatsuhiko Tsunoda[6], Takashi Takei[1], Wataru Obara[7], Fumihiro Akiyama[8], Kyoko Ito[1], Kazuho Honda[1], Keiko Uchida[1], Ken Tsuchiya[1], Wako Yumura[1], Takashi Ujiie[9], Yutaka Nagane[10], Satoru Miyano, Yasushi Suzuki[7], Ichiei Narita[8], Fumitake Gejyo[8], Tomoaki Fujioka[9], Hiroshi Nihei[1] and Yusuke Nakamura: [1]Department of Medicine, Kidney Center, Tokyo Women's Medical University, Tokyo, Japan, [2]Laboratory for Genotyping, [3]Laboratory for Cardiovascular Diseases, [4]Laboratory for Rheumatic Diseases, [5] Laboratory for Diabetic Nephropathy, and [6] Laboratory for Medical Informatics, SNP Research Center, The Institute of Physical and Chemical Research (RIKEN), Tokyo, Japan, [7] Department of Urology, Iwate Medical University, Iwate, Japan, [8]Division of Clinical Nephrology and Rheumatology, Niigata University Graduate School of Medical and Dental Sci-

ences, Niigata, Japan, [9]Department of Urology, Iwate Prefectural Ofunato Hospital, Iwate, Japan, [10]Department of Urology, Sanai Hospital, Iwate, Japan

Immunogobulin A (IgA) nephropathy is the most common form of primary glomerulonephritis worldwide. The pathogenesis of IgA nephropathy is unknown, but it is certain that some genetic factors are involved in susceptibility to the disease. Employing a large-scale, case-control association study using gene-based single-nucleotide polymorphism (SNP) markers, we previously reported three candidate genes. We report here an additional significant association between IgA nephropathy and a SNP located in the gene encoding immunoglobulin μ-binding protein 2 (IGHMBP2) at chromosome 11 q13.2-q13.4. The association ($\chi^2=17.1$, $p=0.00003$; odds ratio of 1.85 with 95% confidence interval of 1.39-2.50 in a dominant association model) was found using DNA from 465 affected individuals and 634 controls. The SNP (G34448A) caused an amino-acid substitution from glutamine to lysine (E928K). As the gene product is involved in immunoglobulin-class switching and patients with the A allele revealed higher serum levels of IgA ($p=0.048$), the amino-acid change might influence a class-switch to increase serum IgA levels, resulting in a higher risk of IgA nephropathy.

So far, we have identified five candidate genes that may be related to susceptibility to IgA nephropathy. On the basis of that information, we propose the potential mechanisms of IgA nephropathy. The onset of IgA nephropathy could be associated with antigens such as viruses, fungus, bacteria or food that are processed and presented to T cells. HLA-DR, which regulates immune responses against protein antigens, is of great importance in the selection and activation of CD4-positive T cells; we identified the gene encoding HLA-DR earlier as a candidate susceptibility gene (Akiyama et al. 2002). HLA-DR molecules with the V724L substitution might account for individual differences in immune responses of T cells, which activate antibody-producing B cells. For its part, as noted above, the 928K variant of IGHMBP2 might influence a class-switch leading to increased serum IgA levels.

The third of our candidates, PIGR, is an integral membrane secretory component localized on the basolateral surface of secretory epithelial cells, where it is thought to mediate the transepithelial transport of polymeric IgA. We showed earlier that a genetic variation in the promoter region of the PIGR gene caused an A 580V substitution associated with IgA nephropathy, and suggested that the V allele might affect binding of polymeric IgA to PIGR and cause deposition of mesangial IgA. IgA deposits in the kidney can trigger production of a variety of cytokines and growth factors by renal cells and by circulating inflammatory cells, leading to the characteristic histopathological features of mesangial-cell proliferation and depositions of immunoglobulin and complement in mesangial regions.

SELL and *SELE* genes encode cell-cell adhesion molecules involved in the leukocyte-endothelial cell interaction required for extravasation at sites of tissue injury. SELE is expressed predominantly in cytokine-activated endothelium, and SELL is present in circulating leukocytes. We reported that Y468H in the *SELE* gene, as well as P238S-SELL and a SNP in the promoter region of SELL, were strongly associated with IgA, and suggested that these substitutions could affect the quality and/or quantity of gene products and possibly play a significant role in inflammatory changes leading to renal fibrosis and ultimately renal failure.

Although functional studies must be undertaken to determine how these genetic variations, now including E928K-IGHMBP2, can affect the onset and development of IgA nephropathy, the results of our genetic studies have suggested several potential mechanisms for investigation.

b. Crohn's disease

Keiko Yamazaki, Masakazu Takazoe[1], Torao Tanaka[1], Toshiki Ichimori[1], Susumu Saito[2], Aritoshi Iida[2], Yoshihiro Onouchi[2], Akira Hata[2] and Yusuke Nakamura: [1]Department of Medicine, Division of Gastroenterology, Social Insurance Chuo General Hospital, Tokyo, Japan, [2]SNP Research Center, the Institute of Physical and Chemical Research (RIKEN), Kanagawa, Japan

The inflammatory bowel diseases (IBD), Crohn's disease (CD) and ulcerative colitis (UC), are chronic inflammatory disorders of the digestive tract. The pathogenesis of IBD is complicated, and it is widely accepted that immunologic, environmental and genetic components contribute to its etiology. In order to identify genetic susceptibility factors in CD, we performed a genome-wide association study in Japanese patients and controls using nearly 80,000 gene-based single nucleotide polymorphism (SNP) markers, and investigated the haplotype structure of the candidate locus in Japanese and European patients. We identified highly significant associations ($p=1.71\times10^{-14}$ with odds ratio of 2.17) of SNPs and haplotypes within the

TNFSF15 (the gene encoding tumor necrosis factor superfamily, member 15) genes in Japanese CD patients. The association was confirmed in the study of two European IBD cohorts. Iinterestingly, a core *TNFSF15* haplotype showing the association with increased risk to the disease was common in the two ethnic groups. Our results suggest that the genetic variations in the *TNFSF15* gene contribute to the susceptibility to IBD in the Japanese and European populationsns.

2. Genes playing significant roles in human cancer

a. Genes that are inducible by p53

Park Woong Ryeon, Chizu Tanikawa, Koichi Matsuda and Yusuke Nakamura

The p53 tumor suppressor gene is more frequently mutated in human cancers than any other cancer-associated genes yet identified; p53 mutations are found in more than half of all cancers examined. The wild-type active form of its product exerts its tumor-suppressing functions either by regulating cell-cycle arrest and DNA repair, or by inducing apoptosis, depending on the specific transcriptional targets that are activated. Although the selection of transcriptional targets seems to depend on the level of cellular stress and to differ by cell type, p53 protein binds to DNA in a sequence-specific manner to activate transcription of genes encoding, for example, $p21^{WAF1}$, p53R2, MDM2, p53 DINP1, p53AIP1, Bax, and GADD45. Modification of the p53 molecule is considered to be important in the process of selecting transcriptional targets, but the mechanism for protein modification is still not well understood. Phosphorylation of p53 at Ser-15 and Ser-20 has been shown to be involved in activating p53. Although the roles of these modifications are not fully characterized, ATM and CHK2 protein are candidates for kinases responsible for phosphorylation of the Ser-15 or Ser-20 residues of p53, respectively (8, 9). After having a low level of DNA damage, p53 is phosphorylated at residues of Ser-15 and Ser-20, and promotes binding of p53 to promoters of genes involved in the G1 arrest and DNA repair. However, if DNA damage is severe, Ser-46 of p53 is phosphorylated and the modified p 53 leads to induction of apoptosis-related genes, such as p53AIP1. Although dozens of p53-target genes involved in p53-dependent tumor suppression, i.e. growth arrest, DNA repair, and apoptosis, have been reported to date, the genetic mechanisms responsible for p53-dependent cell-survival after exposure to various genotoxic stresses remains to be elucidated.

We reported this year isolation of a novel p53 -target gene, designated p53-inducible cell-survival factor (p53CSV). p53CSV contains a p53 -binding site within its second exon and the reduction of expression by small interfering RNA (siRNA) enhanced apoptosis, whereas overexpression protected cells from apoptosis caused by DNA damage. p53CSV is induced significantly when cells have a low level of genotoxic stresses, but not when DNA damage is severe. p 53CSV can modulate apoptotic pathways through interaction with Hsp70 that probably inhibits activity of Apaf-1. Our results imply that under specific conditions of stress, p53 regulates transcription of p53CSV and that p53 CSV is one of important players in the p53-mediated cell survival.

b. Colon, Liver, and Gastric cancers

Ryuji Hamamoto, Masataka Tsuge, Pittella Fabio, Natini Jinawath, Kazutaka Obama, Yoichi Furukawa, and Yusuke Nakamura

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.

We also found a significant association of a genetic polymorphism (VNTR of a "CCGCC" unit) with an increased risk of colorectal cancer ($\chi^2=17.86$, $p=3.8\times10^{-5}$, odds ratio$=2.20$), hepatoma ($\chi^2=25.39$, $p=4.7\times10^{-7}$, odds ratio$=2.74$) and also breast cancer ($\chi^2=38.91$, $p=3.4\times10^{-9}$, odds ratio$=3.79$), but not with that of gastric cancer. This polymorphic region was proven to be a binding-site of the transcriptional factor E2 F-1. The reporter assay exhibited that the reporter plasmid containing three-tandem repeats of the binding motif (corresponding to the risk allele) showed significantly higher reporter activity than those containing two-tandem repeats (the low-risk allele). These data suggest that the

SMYD3 polymorphism enhancing its promoter activity is a common susceptible factor for human cancer.

Among the genes that were up-regulated in tumors, we selected a gene encoding peptidylprolyl isomerase like 1 (*PPIL1*), a cyclophilin-related protein, because it showed a growth-promoting effect on NIH3T3 and HEK293 cells. Moreover, transfection of short-interfering RNA specific to *PPIL1* into SNUC4 and SNUC5 cells effectively reduced expression of the gene and suppressed growth of those colon-cancer cells. In addition, we documented interaction between PPIL1 protein and stathmin. Since stathmin is up-regulated in various types of malignancy, interaction between these two proteins may play an important role in cell proliferation. The findings reported here may offer new insight into colonic carcinogenesis and should contribute to development of new molecular strategies for treatment of human colorectal tumors.

C. cDNA microarray analysis of cancers

Toyomasa Katagiri, Yataro Daigo, Hidewaki Nakagawa, Yoichi Furukawa, Takefumi Kikuchi, Soji Kakiuchi, Toru Nakamura, Koichi Okada, Satoshi Nagayama, Shingo Ashida, Toshihiko Nishidate, Chie Suzuki, Nobuhisa Ishikawa, Ryo Takata, Tatsuya Kato, Akira Togashi, Satoshi Hayama, Megumi Iizumi, Keisuke Taniuchi, and Yusuke Nakamura

(1) Chemosensitivity

Neoadjuvant chemotherapy for invasive bladder cancer, involving a regimen of methotrexate, vinblastin, doxorubicin, and cisplatin (M-VAC), can improve the resectability of larger neoplasms for some patients and offer a better prognosis. However, some suffer severe adverse drug reactions without any effect, and no method yet exists for predicting the response of an individual patient to chemotherapy. Our purpose in this study is to establish a method for predicting response to the M-VAC therapy. We analyzed gene-expression profiles of biopsy materials from 27 invasive bladder cancers using a cDNA microarray consisting of 27,648 genes, after populations of cancer cells had been purified by laser-microbeam microdissection. We identified dozens of genes that were expressed differently between nine "responder" and nine "non-responder" tumors; from that list we selected the 14 "predictive" genes that showed the most significant differences and devised a numerical prediction-scoring system that clearly separated the responder group from the non-responder group. This system accurately predicted the drug responses of eight of nine test cases that were reserved from the original 27 cases. As real-time RT-PCR data were highly concordant with the cDNA microarray data for those 14 genes, we developed a quantitative RT-PCR based-prediction system that could be feasible for routine clinical use. Our results suggest that the sensitivity of an invasive bladder cancer to the M-VAC neoadjuvant chemotherapy can be predicted by expression patterns in this set of genes, a step toward achievement of "personalized therapy" for treatment of this disease.

Serum levels of amphiregulin (AREG) and transforming growth factor-alpha (TGFA) that were previously identified to be expressed at high levels in non-small cell lung cancer (NSCLC) with poor response to gefitinib, were examined by ELISA using blood samples taken from 50 patients with advanced NSCLCs. Of 14 cases that revealed above the cut-off line for AREG in serum, twelve responded poorly (PD) to gefitinib, whereas 18 of the 36 cases showing below the cut-off revealed partial response (PR) or stable condition (SD) ($P=0.026$). Thirteen of 15 patients who were positive for TGFA responded poorly to gefitinib, while 18 of the 35 patients with negative TGFA levels turned out to be relatively good responders ($P=0.014$). Of 22 patients with positive values for either or both marker, 19 were poor responders. On the other hand, among 28 patients negative for both markers, 17 were classified into the PR or SD groups ($P=0.001$). Gefitinib-treated NSCLC patients whose serum AREG or TGFA was positive showed a poorer tumor-specific survival ($P=0.037$ and 0.002 respectively, by univariate analysis), compared with those whose serum AREG or TGFA concentrations were negative. Multivariate analysis showed an independent association between positivity for TGFA and shorter survival times among NSCLC patients treated with gefitinib ($P=0.034$). AREG or TGFA positivity in NSCLC tissues was significantly higher in male, non-adenocarcinomas, and smokers. Our data suggest that the status of AREG and TGFA in serum can be an important predictor of the resistance to gefitinib among patients with advanced NSCLC.

(2) Lung cancer

We have been investigating genes involved in pulmonary carcinogenesis by examining genome-wide gene-expression profiles of non-small cell lung cancers (NSCLCs), to identify molecules that might serve as diagnostic markers or targets for development of new molecular therapies. Distant metastasis is one of crucial parameters that determine types of treatment and prognosis of patients. Numbers of previous reports discovered important factors involved in multi-

ple steps of metastasis, the precise mechanisms of metastasis still remain to be clarified. To identify genes associated with this complicated biological feature of cancer, we analyzed expression profiles of 16 metastatic brain tumors derived from primary lung adenocarcinoma (ADC) using cDNA microarray representing 23,040 genes. We applied bioinformatical algorithm to compare the expression data of these 16 brain metastatic loci with those of 37 primary NSCLCs including 22 ADCs, and found that metastatic tumor cells has very different characteristics of gene expression patterns from primary ones. 244 genes that showed significantly different expression levels between the two groups included plasma membrane bounding proteins, cellular antigens, and cytoskeletal proteins that might play important roles in altering cell-cell communication, attachment, and cell motility, and enhance the metastatic ability of cancer cells. Our results provide valuable information for development of predictive markers as well as novel therapeutic target molecules for metastatic brain tumor of ADC of the lung.

We found that human ANLN, a homologue of anillin, an actin-binding protein in *Drosophila*, was transactivated in lung-cancer cells and appeared to play a significant role in pulmonary carcinogenesis. Induction of small interfering RNAs (siRNAs) against *ANLN* in NSCLC cells suppressed its expression and resulted in growth suppression; moreover, siRNA treatment yielded cells with larger morphology and multiple nuclei, which subsequently died. On the other hand, induction of exogenous expression of ANLN enhanced the migrating ability of mammalian cells by interacting with RHOA, a small GTPase, and inducing actin stress fibers. Interestingly, inhibition of PI3K/AKT activity in NSCLC cells decreased the stability of ANLN and caused reduction of the nuclear ANLN level. Immunohistochemical staining of nuclear ANLN (n-ANLN) on lung-cancer tissue microarrays was associated with poor survival of NSCLC patients, indicating that this molecule might serve as a prognostic indicator. Our data imply that up-regulation of ANLN is a common feature of the carcinogenetic process in lung tissue, and suggest that selective suppression of ANLN could be a promising approach for developing a new strategy to treat lung cancers.

We reported evidence that a member of the armadillo protein family, plakophilin 3 (PKP3), is a potential molecular target for treatment of lung cancers and might also serve as a prognostic indicator. We documented elevated expression of *PKP3* in the great majority of NSCLC samples examined. Treatment of NSCLC cells with small interfering RNAs (siRNAs) of *PKP3* suppressed growth of the cancer cells; on the other hand, induction of exogenous expression of PKP3 conferred growth-promoting activity on COS-7 cells and enhanced their mobility *in vitro*. To investigate its function, we searched for PKP 3-interacting proteins and identified dynamin 1-like (DNM1L), which was also activated in NSCLC. In addition, a high level of *PKP3* expression was associated with poor survival as well as disease stage and node status for patients with lung adenocarcinoma (ADC), suggesting an important role of the protein in development and progression of this disease. As our data imply that up-regulation of PKP3 is a frequent and important feature of lung carcinogenesis, we suggest that targeting the PKP3 molecule might hold promise for development of a new therapeutic and diagnostic strategy for clinical management of lung cancers.

An increased level of dihydrouridine in transfer RNA^{Phe} was found in human malignant tissues nearly three decades ago, but its biological significance in carcinogenesis has remained unclear. Through analysis of genome-wide gene-expression profiles among non-small cell lung carcinomas (NSCLCs), we identified overexpression of a novel human gene, termed *hDUS2*, encoding a protein that shared structural features with tRNA-dihydrouridine synthases (DUS). The deduced 493-amino-acid sequence showed 39% homology to the Dus2 enzyme (dihydrouridine synthase 2) of *Saccharomyces cerevisiae*, and contained a conserved double-strand RNA-binding motif (DSRM). We found that hDUS2 protein had tRNA-dihydrouridine synthase activity and that it physically interacted with EPRS, a glutamyl-prolyl tRNA synthetase, and was likely to enhance translational efficiencies. An siRNA against *hDUS2* transfected into NSCLC cells suppressed expression of the gene, reduced the amount of dihydrouridine in tRNA molecules, and suppressed growth. Immunohistochemical analysis demonstrated significant association between higher levels of hDUS2 in tumors and poorer prognosis of lung-cancer patients. Our data imply that up-regulation of *hDUS2* is a relatively common feature of pulmonary carcinogenesis, and that selective suppression of hDUS2 enzyme activity and/or inhibition of formation of the hDUS2-tRNA synthetase complex could be a promising therapeutic strategy for treatment of many lung cancers.

(3) Pancreatic cancer

Through functional analysis of genes that were transactivated in pancreatic ductal adenocarcinomas (PDACs), we identified *RAB6KIFL* as a good candidate for development of drugs to

treat PDACs at the molecular level. Knockdown of endogenous *RAB6KIFL* expression in PDAC cell lines by siRNA drastically attenuated growth of those cells, suggesting an essential role for the gene product in maintaining viability of PDAC cells. RAB6KIFL belongs to the kinesin superfamily of motor proteins, which have critical functions in trafficking of molecules and organelles. Proteomics analyses using a polyclonal anti-RAB6KIFL antibody identified one of the cargoes transported by RAB6KIFL as discs large homolog 5 (DLG5), a scaffolding protein that may link the vinexin-β-catenin complex at sites of cell-cell contact. Like *RAB6KIFL*, *DLG 5* was up-regulated in PDACs, and knockdown of endogenous DLG5 by siRNA significantly suppressed the growth of PDAC cells as well. Decreased levels of endogenous RAB6KIFL in PDAC cells altered the sub-cellular localization of DLG5 from cytoplasmic membranes to cytoplasm. Our results imply that collaboration of RAB6KIFL and DLG5 is likely to be involved in pancreatic carcinogenesis. These molecules should be promising targets for development of new therapeutic strategies for ductal adenocarcinomas of the pancreas.

P-cadherin/CDH3 belongs to the family of classical cadherins that are engaged in various cellular activities including motility, invasion, and signaling of tumor cells, in addition to cell adhesion. However, the biological roles of P-cadherin itself are not fully characterized. Based on information derived from a previous genome -wide cDNA microarray analysis of microdissected PDACs, we focused on *P-cadherin* as one of the genes most strongly over-expressed in the great majority of PDACs. To investigate the consequences of over-expression of P-cadherin in terms of pancreatic carcinogenesis and tumor progression, we used a P-cadherin-deficient PDAC cell line, Panc1, to construct a cell line (Panc1-CDH3) that stably over-expressed P-cadherin. Induction of P-cadherin in Panc1-CDH 3 increased the motility of the cancer cells, but a blocking antibody against P-cadherin suppressed the motility *in vitro*. Over-expression of P-cadherin was strongly associated with cytoplasmic accumulation of one of the catenins, p 120ctn, and cadherin-switching in PDAC cells. Moreover, P-cadherin-dependent activation of cell motility was associated with activation of Rho GTPases, Rac1 and Cdc42, through accumulation of p120ctn in cytoplasm and cadherin-switching. These findings suggest that over-expression of P-cadherin is likely to be related to the biological aggressiveness of PDACs; blocking of P-cadherin activity or its associated signaling could be a novel therapeutic approach for treatment of aggressive pancreatic cancers.

(4) Prostate cancer

Through genome-wide cDNA microarray analysis coupled with microdissection of prostate cancer cells, we identified a novel gene, *PCOTH* (*Prostate Collagen Triple Helix*), showing over-expression in prostate cancer cells and its precursor cells, PINs (prostatic intraepithelial neoplasia). Immunohistochemical analysis using polyclonal anti-PCOTH antibody confirmed elevated expression of PCOTH, a 100-amino-acid protein containing collagen triple-helix repeats, in tumor cells. Knocking-down of its expression using siRNA resulted in drastic attenuation of prostate cancer cell growth. Concordantly, LNCaP-derivative cells that were designed to express PCOTH highly and stably demonstrated the higher growth rate than LNCaP cells transfected with mock vector. Using 2D-DIGE analysis as well as subsequent western blotting and in-gel kinase assay, we found that phosphorylation level of oncoprotein TAF-Iβ/SET was significantly elevated in LNCaP cells transfected with PCOTH than control LNCaP cells. Furthermore, knockdown of endogenous TAF-Iβ expression using siRNA also attenuated viability of prostate cancer cells as well. These findings suggest that PCOTH is involved in growth and survival of prostate cancer cells thorough, in parts, the TAF-Iβ pathway, and that this molecule should be a promising target for development of new therapeutic strategies for prostate cancers.

(5) Renal cancer

To identify molecules to serve as diagnostic markers for renal cell carcinoma (RCC) and as targets for novel therapeutic drugs, we investigated genome-wide expression profiles of RCCs using a cDNA microarray. We subsequently confirmed that hypoxia-inducible protein-2 (*HIG 2*) was expressed exclusively in RCCs and fetal kidney. Induction of *HIG2* cDNA into COS7 cells led to secretion of the gene product into culture media and resulted in enhancement of cell growth. Small interfering RNA (siRNA) effectively inhibited expression of *HIG2* in human RCC cells that endogenously expressed high levels of the protein, and significantly suppressed cell growth. Moreover, addition of polyclonal anti-HIG2 antibody into culture media induced apoptosis in RCC-derived cell lines. By binding to an extracellular domain of frizzled homolog 10 (FZD10), HIG2 protein enhanced oncogenic Wnt-signaling and its own transcription, suggesting that this product is likely to function as an autocrine growth factor. ELISA analysis of clinical samples identified secretion of HIG2 protein into the plasma of RCC patients even at an early stage of tumor development, whereas it

was detected at significantly lower levels in healthy volunteers or patients with chronic glomerulonephritis. The combined evidence suggests that this molecule represents a promising candidate for development of molecular-targeting therapy and could serve as a prominent diagnostic tumor-marker for patients with renal carcinomas.

Publications

1 S. Ohtsubo, A. Iida, K. Nitta, T. Tanaka, R. Yamada, Y. Ohnishi, S. Maeda, T. Tsunoda, T. Takei, W. Obara, F. Akiyama, K. Ito, K. Honda, K. Uchida, K. Tsuchiya, W. Yumura, T. Ujiie, Y. Nagane, S. Miyano, Y. Suzuki, I. Narita, F. Gejyo, T. Fujioka, H. Nihei and Y. Nakamura: Association of a single-nucleotide polymorphism in the immunoglobulin μ-binding protein 2 gene with immunoglobulin A nephropathy. Journal of Human Genetics, 50: 30-35, 2005

2 F.P. Silva, R. Hamamoto, Y. Nakamura, and Y. Furukawa: WDRPUH, a novel WD-repeat containing protein, is highly expressed in human hepatocellular carcinoma and involved in cell proliferation. Neoplasia, 7: 348-355, 2005

3 A. Iida, K. Ozaki, T. Tanaka, and Y. Nakamura: Fine-scale SNP map of an 11-kb genomic region at 22q13.1 containing the galectin-1 gene. Journal of Human Genetics, 50: 42-45, 2005

4 W.-R. Park and Y. Nakamura: p53CSV, a novel p53-inducible gene involved in the p53 -dependent cell-survival pathway. Cancer Research 65: 1197-1206, 2005

5 K. Taniuchi, H. Nakagawa, T. Nakamura, H. Eguchi, H. Ohigashi, O. Ishikawa, T. Katagiri, and Y. Nakamura: Over-expressed P-cadherin/CDH3 Promotes Motility of Pancreatic Cancer Cells by Interacting with p120 ctn and Activating Rho-Family GTPases. Cancer Research 65: 3092-3099, 2005

6 R. Takata, T. Katagiri, M. Kanehira, T. Tsunoda, T. Shuin, T. Miki, M. Namiki, K. Kohri, Y. Matsushita, T. Fujioka and Y. Nakamura: Predicting response to M-VAC neoadjuvant chemotherapy for bladder cancers through genome-wide gene expression profiling. Clinical Cancer Research 11: 2625-2636, 2005

7 H. Kizawa, I. Kou, A. Iida, A. Sudo, Y. Miyamoto, A. Fukuda, A. Mabuchi, A. Kotani, A. Kawakami, S. Yamamoto, N. Uchida, K. Nakamura, K. Notoya, Y. Nakamura and S. Ikegawa: An aspartic acid repeat polymorphism in asporin negatively affects chondrogenesis and increases susceptibility to osteoarthritis. Nature Genetics 37: 138-144, 2005

8 M. Nishiu, Y. Tomita, S. Nakatsuka, T. Takakuwa, N. Iizuka, Y. Hoshida, J. Ikeda, K. Iuchi, R. Yanagawa, Y. Nakamura and K. Aozasa: Distinct Pattern of gene expression in pyothorax-associated lymphoma (PAL), a lymphoma developing inlong-standing inflammation. Cancer Science, 95: 828-834, 2005

9 K. Taniuchi, H. Nakagawa, T. Nakamura, H. Eguchi, H. Ohigashi, O. Ishikawa, T. Katagiri, and Y. Nakamura: Down-regulation of RAB6KIFL/KIF20A, a Kinesin Involved with Membrane Trafficking of Discs Large Homolog 5, Can Attenuate Growth of Pancreatic Cancer Cells. Cancer Research, 65: 105 -112, 2005

10 A. Iida, Y. Nakamura: Identification of 156 novel SNPs in 29 genes encoding G-protein coupled receptors. Journal of Human Genetics 50: 182-191, 2005

11 Y. Anazawa, H. Nakagawa, M. Furihara, S. Ashida, K. Tamura, H. Yoshioka, T. Shuin, T. Fujioka, T. Katagiri, and Y. Nakamura: PCOTH, a Novel Gene Over-expressed in Prostate Cancers, Promotes Prostate Cancer Cell Growth Through Phosphorylation of Oncoprotein TAF-Ib/SET. Cancer Research, 65: 4578-4586, 2005

12 A. Togashi, T. Katagiri, S. Ashida, T. Fujioka, O. Maruyama, Y. Wakumoto, Y. Sakamoto, M. Fujime, Y. Kawachi, T. Shuin and Yusuke Nakamura: Hypoxia-inducible protein 2 (HIG2), a novel diagnostic marker for renal cell carcinoma (RCC) and potential target for molecular therapy. Cancer Research, 65: 4817 -4826, 2005

13 S. Nagayama, C. Fukukawa, T. Katagiri, T. Okamoto, T. Aoyama, N. Oyaizu, M. Imamura, J. Toguchida and Y. Nakamura: Therapeutic potential of antibodies against FZD10, a cell-surface protein, for synovial sarcomas. Oncogene, 24: 6201-6212, 2005

14 H. Mototani, A. Manuchi, S. Saito, M. Fujioka, A. Iida, Y. Takatori, A. Kotani, T. Kubo, K. Nakamura, A. Sekine, Y. Murakami, T. Tsunoda, K. Notoya, Y. Nakamura and S. Ikegawa: A functional single nucleotide polymorphism in the core promoter region of CALM1 is associated with hip osteoarthritis in Japanese. Human Molecular Genetics, 14: 1009-1017, 2005

15 K. Obama, K. Ura, M. Li, T. Katagiri, T. Tsunoda, A. Nomura, S. Satoh, Y. Nakamura, and Y. Furukawa: Genome-wide

analysis of gene expression in human intrahepatic cholangiocarcinomas. Hepatology, 41: 1339-1348, 2005
16 K. Obama, K. Ura, S. Satoh, Y. Nakamura, and Y. Furukawa: Up-regulation of PSF2, a member of multiprotein complex GINS, involved in cholangiocarcinogenesis. Oncology Report, 14: 701-706, 2005
17 T. Watanabe, T. Suda, T. Tsunoda, N. Uchida, K. Ura, T. Kato, S. Hasegawa, S. Satoh, S. Ohgi, H. Tahara, Y. Furukawa and Y. Nakamura: Identification of Immunoglobulin superfamily 11 (IGSF11) as a novel target for cancer immunotherapy of gastro-intestinal cancers. Cancer Science, 96: 498-506, 2005
18 S. Maeda, S. Tsukada, A. Kanazawa, A. Sekine, T. Tsunoda, D. Koya, H. Maegawa, A. Kashiwagi, T. Babazono, M. Matsuda, Y. Tanaka, T. Fujioka, H. Hirose, T. Eguchi, Y. Ohno, C. Groves, A. Hattersley, G. Hitman, M. Walker, K. Kaku, Y. Iwamoto, R. Kawamori, R. Kikkawa, N. Kamatani, M. McCarthy, and Y. Nakamura: Genetic Variations in the gene encoding TFAP2B are associated with type 2 diabetes. Journal of Human Genetics, 50: 283-292, 2005
19 S. Shimazaki, Y. Kawamura, A. Kanazawa, A. Sekine, S. Saito, T. Tsunoda, D. Koya,T. Babazono, Y. Tanaka, M. Matsuda, K. Kawai, T. Iiizumi, M. Imanishi, T. Shinosaki, T. Yanagimoto, M. Ikeda, S. Omachi, A. Kashiwagi, K. Kaku, Y. Iwamoto, R. Kawamori, R. Kikkawa, M. Nakajima, Y. Nakamura, and S. Maeda: Genetic variations in the gene encoding engulfment and cell motility 1 (ELMO1) are associated with susceptibility to diabetic nephropathy. Diabetes, 54: 1171-1178, 2005
20 M. Akahoshi, K. Obara, T. Hirota, A. Matsuda, K. Hasegawa, N. Takahashi, M. Shimizu, K. Nakashima, S. Doi, H. Fujiwara, A. Miyatake, K. Fujita, N. Higashi, M. Taniguchi, T. Enomoto, X.Q. Mao, K. Nakashima, C.N. Adra, Y. Nakamura, M. Tamari, and T. Shirakawa: A functional promoter polymorphism in the TBX21 gene is associated with aspirin-induced asthma. Human Genetics, 117: 16-26, 2005
21 T. Kato, Y. Daigo, S. Hayama, N. Ishikawa, T. Yamabuki, T. Ito, M. Miyamoto, S. Kondo and Y. Nakamura: A novel human tRNA-dihydrouridine synthase involved in pulmonary carcinogenesis. Cancer Research, 65: 5638-5646, 2005
22 K. Asamura, S. Abe, Y. Imamura, A. Aszodi, N. Suzuki, S. Hashimoto, Y. Takumi, T. Hayashi, R. Fassler, Y. Nakamura, and S. Usami: Type IX collagen is crucial for normal hearing. Neuroscience, 132: 493-500, 2005
23 R. Kawaida, R. Yamada, K. Kobayashi, S. Tokuhiro, A. Suzuki, Y Kochi, X. Chang, A. Sekine, T. Tsunoda, T. Sawada, H. Furukawa, Y. Nakamura, and K. Yamamoto: CUL1, a component of E3 ubiquitin ligase, alters lymphocyte signal transduction with possible effect on rheumatoid arthritis. Genes Immun, 6: 194-202, 2005
24 S. Seki, Y. Kawaguchi, K. Chiba, Y. Mikami, H. Kizawa, T. Oya, F. Mio, M. Mori, Y. Miyamoto, I. Masuda, T. Tsunoda, M. Kamata, T. Kubo, Y. Toyama, T. Kimura, Y. Nakamura and S. Ikegawa: A functional SNP in cartilage intermediate layer protein (CILP) is associated with susceptibility to lumbar disc disease. Nature Genetics, 37: 607-612, 2005
25 Y. Kochi, R. Yamada, A. Suzuki, J.B. Harley, S. Shirasawa, T. Sawada, S.C. Bae, S. Tokuhiro, X. Chang, A. Sekine, A. Takahashi, T. Tsunoda, Y. Ohnishi, K.M. Kaufman, C.P. Kang, C. Kang, S. Otsubo, W. Yumura, A. Mimori, T. Koike, Y. Nakamura, T. Sasazuki and K. Yamamoto: A functional variant in FCRL3, encoding Fc receptor-like 3, is associated with rheumatoid arthritis and several autoimmunities. Nature Genetics, 37: 478-485, 2005
26 C. Furukawa, Y. Daigo, N. Ishikawa, T. Kato, T. Ito, E. Tsuchiya, S. Sone, and Y. Nakamura: PKP3 oncogene as prognostic marker and therapeutic target for lung cancer. Cancer Research, 65: 7102-7110, 2005
27 M. Tsuge, R. Hamamoto, F.P. Silva, Y. Ohnishi, K. Chayama, Y. Furukawa, and Y. Nakamura: VNTR Polymorphism of E2F-1 binding element in the 5′ flanking region of SMYD3 is a risk factor for human cancers. Nature Genetics, 37: 1104-1107, 2005
28 M. Sakai, T. Shimokawa, T. Kobayashi, S. Matsushima, Y. Yamada, Y. Nakamura, and Y. Furukawa: Elevated expression of C10orf3 (Chromosome 10 open reading frame 3) is involved in the growth of human colon tumor. Oncogene, in press, 2005
29 M. Takahashi, K. Obama, Y. Nakamura, and Y. Furukawa: Identification of SP5 as a downstream gene of the beta-catenin/Tcf pathway and its enhanced expression in human colon cancer. International Journal of Oncology, 27: 1483-1487, 2005
30 N. Ishikawa, Y. Daigo, A. Takano, M. Taniwaki, T. Kato, S. Hayama, H. Murakami, Y. Takeshima, K. Inai, H. Nishimura, E. Tsuchiya, N. Kohno, and Y. Nakamura: Increases of amphiregulin and transforming growth factor-alpha in serum as predictors of poor response to Gefitinib among patients with advanced non-small cell lung cancers. Cancer Research, 65: 9176-9184, 2005
31 C. Suzuki, Y. Daigo, N. Ishikawa, T. Kato, S.

Hayama, T. Ito, E. Tsuchiya, and Y. Nakamura: ANLN plays a critical role in human lung carcinogenesis through activation of RHOA and by involvement in PI3K/AKT pathway. Cancer Research, in press, 2005
32 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, in press, 2005
33 K. Asamura, S. Abe, H. Fukuoka, Y. Nakamura, S. Usami: Mutation analysis of COL9A 3, a gene highly expressed in the cochlea, in hearing loss patients. Auris Nasus Larynx, 32: 113-117, 2005
34 Kanazawa, Y. Kawamura, A. Sekine, A. Iida, T. Tsunoda, A. Kashiwagi, Y. Tanaka, T. Babazono, M. Matsuda, K. Kawai, T. Iiizumi, T. Fujioka, M. Imanishi, K. Kaku, Y. Iwamoto, R. Kawamori, R. Kikkawa, Y. Nakamura, and S. Maeda: Single nucleotide polymorphosms in the gene encoding Kruppel-like factor 7 are associated with type 2 diabetes. Diabetologia, 48: 1315-1322, 2005
35 C. Furukawa, Y. Nakamura and T. Katagiri: Molecular target therapy of synovial sarcoma. Future Oncology, in press, 2005
36 T. Watanabe, M. Suzuki, Y. Yamasaki, S. Okuno, H. Hishigaki, T. Ono, K. Oga, A. Mizoguchi-Miyakita, A. Tsuji, N. Kanemoto, S. Wakitani, T. Takagi, Y. Nakamura, and A. Tanigami: Mutated G-protein-coupled receptor GPR10 is responsible for the hyperphagia/dyslipidaemia/obesity locus of Dmol in the OLETD rat. Clinical and Experimental Pharmacology and Physiology, 32: 355-366, 2005
37 K. Yamazaki, D. McGovern, J. Ragoussis, M. Paolucci, H. Butler, D. Jewell, L. Cardon, M. Takazoe, T. Tanaka, T. Ichimori, S. Saito, A. Sekine, A. Iida, A. Takahashi, T. Tsunoda, M. Lathrop and Y. Nakamura: Single nucleotide polymorphisms in TNFSF15 confer susceptibility to Crohn's disease. Human Molecular Genetics, 14: 3299-3506, 2005
38 The International HapMap Consortium: A haplotype map of the human genome. Nature, 4371299-1320, 2005
39 T. Ishibe, T. Nakayama, T. Okamoto, T. Aoyama, K. Nishijo, K. Shibata, Y. Shima, S. Nagayama, T. Katagiri, Y. Nakamura, T. Nakamura, and J. Toguchida: Disruption of fibroblast growth factor signal pathway inhibits the growth of synovial sarcomas: potential application of signal inhibitor to molecular target therapy. Clinical Cancer Research, 11: 2702-2712, 2005
40 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, in press, 2005
41 A. Iida, S. Saito, A. Sekine, A. Takahashi, N. Kamatani, and Y. Nakamura: Japanese SNP database for 267 possible drug-related genes. Cancer Science, in press, 2005
42 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, in press, 2005
43 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 CYP 2C9 polymorphisms with warfarin dose requirements in Japanese patients. Journal of Human Genetics, in press, 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. DBTGR: A Database for Comparative Analysis of Tunicates Promoter Sequences

Nicolas Sierro, Takehiro Kusakabe[1], Keun-Joon Park, Riu Yamashita, Kengo Kinoshita and Kenta Nakai: [1]Graduate School of Life Science, University of Hyogo

The two ascidians *Ciona intestinalis* and *Ciona savignyi* belong to the tunicate subphylum, which is particularly interesting because it shares many developmental and physiological characteristics, as well as basic gene repertoires, with the vertebrates. The rapid development of a fertilized ascidian egg into a transparent tadpole-like larva having a body plan similar to its vertebrate counterpart and the availability of a well-established cell-lineage has made of these organisms a favored tool to elucidate the genetic regulatory systems underlying the developmental and physiological processes of vertebrates. In order to understand the regulation of these numerous genes identified after the recent release of the *C. intestinalis* and *C. savignyi* genomes, a database was created containing information on regulation of tunicate genes collected from literature, as well as predicted binding sited for the identified transcription factors. The information contained in the DataBase of Tunicate Gene Regulation (DBTGR, http://dbtgr.hgc.jp) originates from two sources. On one hand, gene regulation, transcription factor and their binding sites obtained by searching published literature. And on the other hand *C. intestinalis* and *C. savignyi* promoter sequences extracted from the most current genome releases, either by using the information provided in the literature and with the genome releases, or by sequence alignment and homology searches. Additionally, the recognition sites of the reported transcriptions factors were used to identify new potential binding sites within the promoters. More importantly, DBTGR provides an alignment between corresponding *C. intestinalis* and *C. savignyi* gene promoter sequences facilitating the identification of actual regulatory elements and of regions conserved in both promoters.

2. Motif Analysis of Tissue-specific Promoters in *Ciona intestinalis*

Keun-Joon Park, Nicolas Sierro, Takehiro Kusakabe[1], Riu Yamashita, Kengo Kinoshita and Kenta Nakai

DBTGR is a database of tunicate promoters and their regulatory elements. While constructing the DBTGR database, we also investigated promoters leading to tissue-specific expression in *Ciona* (tunicates). We have found muscle-specific motifs in the *Ciona intestinalis* dataset constructed from DBTGR and obtained with PSSM (position-specific scoring matrix) for each motif. After finding muscle-specific motifs in the restricted DBTGR *Ciona intestinalis* dataset, we searched for the same motifs in the complete *Ciona intestinalis* genome sequence (JGI version 1.95). To find potential muscle-specific genes in the genome sequence, we constructed some muscle-specific motifs combination (element) models based on the results of the DBTGR dataset analysis. These models were used to detect new muscle-specific genes in the whole genome. After prediction of TATA-box sequences in the upstream region of each gene, significant motif combinations matching the models were detected in the genome sequence. To improve these muscle-specific promoter element models, and discover potentially new ones, we are going to analyze the new *Ciona intestinalis* genome version 2.0 and to carry out comparative genomics analysis using the *Ciona savignyi* genome sequence.

3. Comparative Analysis of Firmicute Promoters

Nicolas Sierro and Kenta Nakai

With the rapid increase in the number of bacterial genome entirely sequenced, systematic function analysis projects have started to decipher the total gene activity of these organisms. 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. The previously constructed database of transcriptional regulation in *Bacillus subtilis* (DBTBS) focuses on known transcription factors, their recognition sequences and the gene they control in *B. subtilis.* However, in the recent years, more than 60 other firmicute genomes, including medically and industrially important species such as *Streptococcus pyogenes*, *Staphylococcus aureus* or *Lactococcus lactis*, have been completely sequenced. Using the information available from these genomes, as well as regulatory events reported in literature, a new insight in the evolution of regulatory networks could be obtained. For instance although the heat shock response is primordial for the survival of all bacteria, and as such is likely to have appeared early in evolution, the interaction between two of its known regulation pathways, controlled by the HrcA and CtsR transcription factors, varies depending on species and subspecies. This illustrates the importance of extending the comprehensive *B. subtilis* information with that obtained from other related bacteria in order to provide a more accurate picture of the bacterial gene regulation.

4. Update of DataBase of Transcription Start Sites

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

DBTSS was first constructed in 2002 based on precise, experimentally determined 5′-end clones. Several major updates and additions have been made since the last report. First, the number of human clones has drastically increased, going from 190,964 to 1,359,000. We checked reliability of these data with ChIP of chip experiments of the ENCODE project. Around 54% of the transcription start sites were observed in ChIP on chip positive (>=2 of ratio) region. Moreover, these TSSs correspond to 90% of 5′-end clones in ENCODE region; therefore, our data is consistent with other experimental works. Second, information about potential alternative promoters is presented because the number of 5′-end clones is now sufficient to determine several promoters for one gene. Namely, we defined putative promoter groups by clustering TSSs separated by less than 500 bases. 8,308 human genes and 4,276 mouse genes were found to have putative multiple promoters. To verify these putative alternative promoters, we obtained 138 alternative promoters (64 genes) based on other experimental methods. 74% of them corresponded to our DBTSS alternative promoters. Finally, we have added TSS information for zebrafish, malaria, and schyzon (a red algae model organism). DBTSS is accessible at http://dbtss.hgc.jp.

5. Comparative Analysis of Alternative Promoters between Human and Mouse Genes

Katsuki Tsuritani[3], Yutaka Suzuki[2], Koichi Kimura[4], Ai Wakamatsu[4], Riu Yamashita, Takao Isogai[4], Sumio Sugano[2], and Kenta Nakai: [3]Taisho Pharmaceutical Co. Ltd., [4]REPRORI Co. Ltd.

It gradually becomes clear that a large population of human genes are regulated by more than one alternative promoters (APs). In this re-

port, we focus on the comparative analysis of human/mouse APs. We extracted orthologous genes that have more than two APs in either human or mouse from DBTSS (http://dbtss.hgc.jp/) and analyzed their putative promoter regions using a local alignment program LALIGN. We classified the extracted genes into five categories: '0-0', '1-1', '1-m', 'm-1', and 'm-m' based on the conservation of their promoter regions; where '0-0' is the case that there were no conserved regions, '1-m' and 'm-1' contain some redundant orthologous promoters in mouse and human, respectively, 'm-m' is the case that both '1-m' and 'm-1' hold, and '1-1' is the remaining case which indicates that the promoter is not duplicated. The number of categories for '0-0', '1-1', '1-m', 'm-1', and 'm-m' were 538, 4364, 821, 334, and 224, respectively. We extracted 523 genes with more than two reciprocally conserved promoter regions from the major category '1-1' and defined them as the AP core dataset. Among them, the genes that have the most conserved promoter regions were 'neural precursor cell expressed, developmentally down-regulated 4-like' and 'regulator of G-protein signaling 3', both having five conserved regions. By analyzing their Gene Ontology annotation, we found that genes related to 'signal transduction' are significantly enriched in the AP core dataset. Our results suggest that APs contributing to the diversity of cellular signal transduction are well-conserved throughout the evolution.

6. Comparative Sequence Analysis of Human and Mouse Promoter Regions

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

Computational sequence analysis of promoter regions is essential to elucidate the mechanism of transcriptional regulation. The accumulation of experimentally validated TSS data made possible a large scale promoter comparison not only possible but also an effective approach. Based on the database of transcriptional start sites (DBTSS) developed in our laboratory, we carried out the most comprehensive comparison to date for promoter regions of human and mouse genes, aiming at elucidating the relationship between gene function and promoter conservation. For 70% of the orthologous gene pairs, promoters were detected to be evolutionary related when compared to promoters from unrelated genes. Conservation levels in a wide range from gene to gene. The conservation level of genes with specific functions was examined based on GO slim categories, and new functional categories with high promoter conservation levels were identified in addition to the ones reported previously. Furthermore, a similar analysis regarding protein conservation levels was carried out, and evolutional constraints on genes were discussed from the points of view of protein sequence and promoter sequence. For signal transducer, evolutionary constraints tend to be on promoter sequences rather than protein sequences, while for enzymes, they are on protein sequences rather than promoter sequences.

7. Comprehensive Analysis of Triplet Repeats in Vertebrate Genomes

Shigeo Okada, Riu Yamashita, Kengo Kinoshita, and Kenta Nakai

About 3% of the human genome is composed of triplet repeats, and their expansion in specific genes is associated with at least 42 human diseases. However, the definition of triplet repeats has not been settled since it may contain one or more potentially important interruptions. For example, Huntington gene has one CAA interruption in its CAG repeat regions. According to the current Repeat Evolution Model, interruptions may have appeared by point mutations during evolution. In addition, the possible functions of the interruptions are to energetically stabilize repeats and to prevent their expansions. However, interruptions are not considered by the conventional *in silico* researches. Therefore we reinvestigated triplet repeats taking interruptions into account. We defined the repeats with interruptions using triplet repeat disease genes and statistical significance. While there are 6,015 amino acid repeats without interruptions in human coding sequences (CDSs), we can find 22,581 repeats with interruptions. From the investigation of repeats in human and mouse orthologous genes, 4,996 repeats exist only in human genes, and 3,522 repeats only in mouse genes. The percentage of a specific amino acid repeat and the average length of the repeats do not vary much between the two species. We also find that the distribution of the differences between mouse and human glutamine-repeat lengths from orthologous genes is almost symmetrical. We investigated the mutation rate of triplet repeats between human, chimpanzee and mouse CDSs. Interruptions caused by one point mutation are the most abundant ($>40\%$). We also find that the distribution of point mutations resulting in an interruption does not vary much between species, but varies widely depending on the repeated triplet. This indicates that the formation of interruptions depends on the repeated triplet, but is independent of the species

and triplet repeat length. In summary, it appears that repeats and interruptions are formed at the same rate, independently of the species and despite the diversification of the genomes. Our results are in agreement with the Repeat Evolution Model and imply that interruptions could play some roles in the repeats because they are relatively frequent. We expect that our results will be of great help to understand the relationship between triplet repeats and diseases.

8. ATTED-II: A Database of Co-expressed Genes And *cis-elements for Identifying Co -Regulated Gene Groups in Arabidopsis*

Takeshi Obayashi[5,7], Kenta Nakai, Daisuke Shibata[6], Kazuki Saito[7], Hiroyuki Ohta[5]: [5]Tokyo Institute of Technology, [6]Kazusa DNA Research Institute, [7]Chiba University

Finding out the combination of co-expressed gene sets would be valuable for a wide variety of experimental design, such as targeting the genes for functional identification and for investigation of possible cis-elements in promoter sequences. Here, we report the construction of *Arabidopsis thaliana* trans-factor and cis-element prediction database (ATTED-II), which provides co-regulated gene relationships based upon co-expressed genes deduced from microarray data with the predicted cis elements. ATTED-II includes following features: (i) lists of co-expressed genes calculated with 771 publicly available microarray data in *A. thaliana* or with the subsets of these data, (ii) prediction of cis regulatory elements in the 200-bp region upstream of transcription start site to estimate co-regulated genes from the co-expressed genes, (iii) prediction of subcellular localizations and functional groups of proteins to support the estimation of the co-regulated genes. ATTED-II can thus provide the clues for researchers to clarify the function and regulation of particular genes and the networks of gene-to-gene relationships (http://www.atted.bio.titech.ac.jp).

9. Computational Analysis of microRNA Recognition Site

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

microRNAs (miRNAs) are noncoding RNAs about 22 nucleotides that suppress translation of target genes by binding to their mRNA, and thus have a role in gene regulation. Recently Lim et al. (2005 Nature) transfected miRNAs into human cells and used microarrays to examine changes in the messenger RNA profile. These microarray profiles indicate that the 3′-untranslated regions of down-regulated messages have a significant propensity to pair to the 5′-region of the miRNAs, as expected if many of these messages are the direct targets of the miRNAs. However miRNA target prediction* search complementary sequence of Seed (SeedCS) results contain many false positives. We need to define more fine Seed and create the target prediction algorism to reduce false positive. Here we present a visualization method to discover out the miRNAs recognition site by computational analysis for identify the location of miRNAs recognition site and its required length. And present method to reduce false positive at target prediction. The sequence data of miRNAs were obtained from Lim et al., and the mRNA sequence data were obtained from the Ensembl database. The complementation at each position of mRNA sequence by a miRNA was considered. The length and start position within the miRNA sequence of any nucleotide stretch longer than 2based were recorded. We thereby obtained length vs position matrices for each miRNA-mRNA pair. Matrices obtained with the same miRNA were added together, yielding general miRNA matrices. The microarray data were obtained from NCBI GEO (GSE2075). miRNA specific gene expression profiles were linked to Ensembl transcripts based on the mapping of microarray oligo nucleotides on Ensembl mRNAs. For consider both sensitivity and specificity, we use maximum Matthews correlation coefficient for score. That score are ploted as dot color and aligned matrix shape. It indicates miRNA's length 8 position 0 and length 7 position 1 is important to recognize mRNA. Next, to predict significant down-regulated mRNAs, we picked up top mRNAs that down-regulated at microarray experiment. It has many SeedCS compare with ambiguously down-regulated mRNAs. Then we picked mRNA that has plural SeedCS. Then, false positives were reduesed. mRNAs which has more than two Seed complementary sequences have highly down-regulated and can be easily predicted.

*miRNA target prediction: Prediction of mRNAs that are degraded by miRNA

10. Sequence-Based Analyses of Biosynthesis Rate Limiting Factors in Wheat Germ Cell-Free System

Naoya Fujita, Motoaki Seki[8], Kazuo Shinozaki[8], Kengo Kinoshita, Tatsuya Sawasaki[9], Yaeta Endo[9], Kenta Nakai: [8]RIKEN, [9]Faculty. of Eng., Ehime Univ.

The production of proteins themselves is essential for their structural and functional analyses in the post-genome era. A wheat germ cell-free system is a helpful method for that purpose, as it is able to produce proteins from various sources' mRNAs using the translational machinery of wheat germs. Only one 5'-UTR was used to obtain similar biosynthesis yields regardless of the coding sequence. However, the range of yields observed is very wide. We therefore investigated the causes of the yield variation based on the protein sequences. The dataset used consists of 425 protein kinases from *Arabidopsis thaliana.* 4 candidates were considered as yield variation factors: disordered region, coiled coil structure, codon usage and mRNA secondary structure. Disordered regions and coiled coil structures were predicted using DISOPRED2 and COILS respectively. A specific negative correlation between N terminal disorders and yields was found. If a protein has a high disorder, its yield will be low. However, the yield range of proteins with a lower disorder percentage remains wide. Focusing on the region with less than 20% disorder and considering coiled coil structures, we found that if a protein kinase has coiled coils, its biosynthesis yield will be low. mRNA structure also relate to the yield decrease, consistent with the observed *in vivo* gene expression in *E. coli.* Codon usage has no influence on our dataset, although this lack of effect may be restricted to the protein kinase family. We interpreted the disorder and coiled coil factors and proposed an entanglement model in which potential interactions and entanglements between neighboring synthesized polypeptides could occur, resulting in a lower protein production. Further investigations regarding biosynthesis yields of other protein families as well as protein kinase from other organisms are currently underway.

11. Construction and Analyses of A Non-interacting Sites Database for Protein-protein Interaction

Miho Higurashi, Kenta Nakai, and Kengo Kinoshita

Many of the celluer events are regulated through the interaction between protein and protein. Recent growth of PDB entry enables us to reveal characteristics of protein-protein interacting sites from the viewpoint of structural genomics. In the past studies, comparison of protein-protein interacting sites with whole surface of protein was done to extract features of protein-protein interacting sites. However, the feature of protein-protein interacting site may be obscured because whole surface contain interacting sites. To extract the characteristics of protein -protein interacting sites, comparison of protein-protein interacting sites with non-interacting sites should be done. Refer to protein-protein interacting site of homologue proteins in PDB, we constructed database for sites with which interacts nothing, as far as we know.

12. Modeling Tertiary Structure of Complementarity-Determining Region of Antibodies

Toru Hosokawa, Kenta Nakai, and Kengo Kinoshita

The tertiary structure of protein is important to predict protein function and achieve rational drug design. Therefore, the establishment of protein tertiary structure prediction by computer is needed. We focused on improving prediction accuracy of the complementarity determining region (CDR) of antibody. It has been reported that CDRs can be modeled by using templates from other antibodies. However, the precise prediction of the position of the CDR and the selection of an appropriate template for the target CDR are necessary. We developed a prediction method which uses hidden markov model and achieved almost 100% prediction accuracy. In previous works, several relations between the amino acid sequence and the canonical structure (called "rule") were suggested. While validating these rules and found 35% of CDR structures for which template less than RMSD 1Å cannot find by known rules. Therefore we created new rules by clustering structures and patterns of sequences. By implementing a new prediction method for localization of the CDR and for rule matching, we successfully automated the structure prediction process of antibodies and improved the overall accuracy of structure prediction from 2.5Å to 1.9Å RMSD. A web server which predicts tertiary structure of antibodies from amino acid sequence was built based on these results and existing homology modeling method.

13. Identification of The Ligand Binding Sites On The Molecular Surface of Proteins

Kengo Kinoshita and Haruki Nakamura[10]: [10]Osaka Univ.

Identification of protein biochemical functions based on their three-dimensional structures is now required in the post-genome-sequencing era. Ligand binding is one of the major biochemical functions of proteins, and thus the identification of ligands and their binding sites is the starting point for the function identification. Here we describe our trial on structure-based function prediction, based on the similarity searches of molecular surfaces against the functional site database, eF-site.

14. PreDs: A Server for Predicting dsDNA-binding Site on Protein Molecular Surfaces

Yuko Tsuchiya[10], Kengo Kinoshita, Haruki Nakamura[10]

PreDs is a WWW server that predicts the dsDNA-binding sites on protein molecular surfaces generated from the atomic coordinates in a PDB format. The prediction was done by evaluating the electrostatic potential, the local curvature and the global curvature on the surfaces. Results of the prediction can be interactively checked with our original surface viewer.

15. P-cats: prediction of catalytic residues in proteins from their tertiary structures.

Kengo Kinoshita and Motonori Ota[5]

P-cats is a web server that predicts the catalytic residues in proteins from the atomic coordinates. P-cats receives a coordinate file of the tertiary structure and sends out analytical results via e-mail. The reply contains a summary and two URLs to allow the user to examine the conserved residues: one for interactive images of the prediction results and the other for a graphical view of the multiple sequence alignment.

16. Analysis of the three-dimensional structure of the "C-X-G-X-C" motif in the CMGCC and CAGYC regions of α- and β-subunits of human chorionic gonadotropin: Importance of Glycine Residue (G) in the Motif

Kengo Kinoshita, Masami Kusunoki[10], and Kiyoshi Miyai[10]

The "C-X-G-X-C" motif of human glycoprotein hormones, hTSH, hLH, hFSH and hCG, are strictly conserved. These proteins form dimer with the identical α-subunit and specific β-subunit for each glycoprotein hormones. Some mutational studies have shown the importance of the Gly residues in the β-subunit. In this study, using the recently solved structure of hCG, we have analyzed role of the glycine residue in the α- and β-subunits by the conformational energy calculation with loop closure algorithm. As a result, in the α-subunit, only a Gly residue is allowed at the third site due to the steric hiderance within the subunit. On the other hand in the β-subunit, Ala residue is also acceptable in the monomer structure, but the Ala is also forbidden when the dimer structure is formed. This different role of the Gly in each subunit can be a possible explanation of the importance of Gly residue in the β-subunit as in the case of TSH deficiency disease, which is caused by the mutation from Gly to Ala in β-subunit.

Publications

Kato, K., Yamashita, R., Matoba, R., Monden, M., Noguchi, S., Takagi, T., and Nakai, K. Cancer gene expression database (CGED): a database for gene expression profiling with accompanying clinical information of human cancer tissues, Nucl. Acids Res., 33: D533-D 536, 2005.

Makita, Y., De Hoon, M.J.L., Ogasawara, N., Miyano, S., and Nakai, K. Bayesian joint prediction of associated transcription factors in Bacillus subtilis, Pacific Symposium on Biocomputing 2005 (Altman et al ed.), 507-518, World Scientific, 2005.

Poluliakh, N., Konno, M., Horton, P., and Nakai, K. Parameter landscape analysis for common motif discovery programs, in Eskin, E. & Workman, C. (eds.), Regulatory Genomics, RECOMB 2004 International Workshop, RRG 2004, San Diego, CA, USA, March 26-27, 2004, Revised Selected Papers. Lecture Notes in Computer Science 3318, pp. 79-87, Springer, 2005.

Yamashita, R., Suzuki, Y., Sugano, S., and Nakai, K. Genome-wide analysis reveals strong correlation between CpG islands with nearby transcription start sites of genes and their tissue-specificity, Gene, 350(2): 129-136, 2005.

Nakao, M., Barrero, R.A., Mukai, Y., Motono, C., Suwa, M., and Nakai, K. Large-scale analysis of human alternative protein isoforms: pattern classification and correlation with subcellular locallization signals, Nucl. Acids Res., 33(8): 2355-2363, 2005.

De Hoon, M.J.L., Makita, Y., Nakai, K., and Miyano, S. Prediction of transcriptional terminators in Bacillus subtilis and related species, PLoS Comput. Biol., 1(3): e25, 2005.

Horton, P., Park, K.-J., Obayashi, T., and Nakai, K., Protein subcellular localization prediction with WoLF PSORT, Proc. APBC, in press.

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., in press

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., in press

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., in press.

Kinoshita, K. and Nakamura, H. Identification of the ligand binding sites on the molecular surface of proteins. *Protein Sci* 14: 711-718, 2005.

Tsuchiya, Y., Kinoshita, K, Nakamura, H. PreDs: a server for predicting dsDNA-binding site on protein molecular surfaces. *Bioinformatics* 21: 1721-1723, 2005.

Kinoshita, K. and Ota, M. P-cats: prediction of catalytic residues in proteins from their tertiary structures. *Bioinformatics* 21: 3570-3571, 2005.

Kinoshita, K., Kusunoki, M., Miyai, K. Analysis of the three-dimensional structure of the "C-X-G-X-C" motif in the CMGCC and CAGYC regions of α- and β-subunits of human chorionic gonadotropin: Importance of Glycine Residue (G) in the Motif, *Endocrine J., in press*

中井謙太，法澤公寛，ゲノム解析とテラヘルツ技術，大森豊明監修，テラヘルツテクノロジー：発生・計測・応用技術・展望，エヌ・ティー・エス，pp. 405–415，2005.

中井謙太，ヒト遺伝子転写開始点のゲノムワイド解析，小原・菅野・小笠原・高木・藤山・辻編集，ゲノムから生命システムへ，蛋白質核酸酵素増刊，50⑯：p. 2083，2005.

木下賢吾，プロテインインフォマティクスによるタンパク質の機能推定，日本結晶学会誌，47：341–347，2005.

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

Professor	Katsuhisa Horimoto, Ph.D.	特任教授　理学博士	堀　本　勝　久
Research Associate	Sachiyo Aburatani, Ph.D.	特任助手　農学博士	油　谷　幸　代

The main projects of our laboratory are to reveal new biological meanings at molecular level by various statistical approaches, and to train the researchers for the appropriate use of statistical techniques. The subject under investigation focuses on devising the methods for inferring the association between biological phenomena from gene expression profiles by statistical models. In addition, the algebraic approaches are applied to some issues on theoretical biology.

1. Association Inference by Statistical Models

a. A graphical chain model for inferring regulatory systems network from gene expression profiles.

Sachiyo Aburatani, Shigeru Saito[1], Katsuhisa Horimoto: [1]INFOCOM CORP.

A procedure for graphical chain modeling has been designed for analyzing the expression profiles of genes that can be classified into several blocks in a natural order. Since the gene expression profiles often share similar patterns, the genes within a block are grouped into some clusters, as a prerequisite for the modeling. Then, the clusters in the naturally ordered blocks are regarded as variables. Finally, the associations of the variables within and between blocks are inferred by the covariance selection in graphical Gaussian modeling. The newly designed procedure for graphical chain modeling was applied to 619 expression profiles of cell-cycle related genes in yeast, which were selected from 792 genes experimentally identified as being transcribed in the order of four cell-cycle phases, G_1, S, G_2, and M. By the application of the procedure, the 619 genes were classified into 50 clusters, and a chain graph was fitted for 50 clusters in the four phases. By focusing on the clusters including transcription factors, characteristic relationships between the clusters emerged from the associations of the clusters within and between the four phases; one of the remarkable features is the distinctive relationships of the clusters between neighboring and non-neighboring phases. The merits and pitfalls of the graphical chain model are discussed in terms of its application to the field of molecular biology.

b. Algorithm for predicting co-expression genes by improvement of path consistency algorithm.

Shigeru Saito[1], Sachiyo Aburatani, Katsuhisa Horimoto: [1]INFOCOM CORP.

We design a simple algorithm to predict the co-expressed genes from expression profiles. The

path consistency (PC) algorithm, which is one of constraint-based method for inferring the causal graph, is modified by considering the nature of actual expression profiles, and further improved by interpolating the biological knowledge of operons. The algorithm was applied to the expression profiles of known operons and of all genes in *Escherichia coli*, and then about 90% and 70% of known operons were correctly detected with small errors. In addition, a reasonable operon candidate in terms of gene function was presented.

c. Orchestration of Gene Systems Inferred from Expression Profiles in Hepatocellular Carcinoma.

Sachiyo Aburatani, Shigeru Saito[1], Masao Honda[2], Shu-ichi Kaneko[2], Katsuhisa Horimoto: [1]INFOCOM CORP., [2]Kanazawa Univ.

Hepatitis C virus (HCV) is the major etiologic agent of non-A non-B hepatitis and chronically infects about 170 million people worldwide. Many HCV carriers develop chronic hepatitis C (CH-C), finally complicated with hepatocellular carcinoma (HCC) in a liver with advanced stage CH-C.

Here, we analyzed the gene expression profiles of HCC and its background liver with advanced stage of CH-C, by a statistical method recently devised to infer the gene systems network from the gene expression profiles, based on the graphical Gaussian model.

Recently, we have developed an approach to infer a regulatory network, which is based on graphical Gaussian modeling (GGM). Our method provides a framework of gene regulatory relationships by inferring the relationship between the clusters, and provides clues toward estimating the global relationships between genes on a genomic scale. Also, we have devised a procedure, named ASIAN (Automatic System for Inferring A Network), to apply GGM to gene expression profiles, by a combination of hierarchical clustering. In this study, the previous version of the ASIAN web server has been improved to facilitate its utilization.

The ASIAN system is composed of four parts: 1) the calculation of a correlation coefficient matrix for the raw data, 2) the hierarchical clustering, 3) the estimation of cluster boundaries, and 4) the application of GGM to the clusters. In the GGM, the network is inferred by the calculation of a partial correlation coefficient matrix from the correlation coefficient matrix, and the partial correlation coefficient matrix can only be obtained if the correlation coefficient matrix is regular. Since the gene expression profiles on a genomic scale often include many profiles sharing similar expression patterns, the correlation coefficient matrix is not always regular. Therefore, the first three parts (1)-3)) are prerequisite for analyzing the redundant data, including many similar patterns of expression profiles, by the last part (4)), the network inference by GGM. All calculations were performed by ASIAN web site and "Auto Net Finder", PC version of ASIAN, from INFOCOM CORPORATION.

The expression profiles of 8516 genes were monitored in 17 samples in HCC. By application of ASIAN to the profiles, a graph of the gene systems is inferred, and it provides snapshot for orchestrating the gene systems in HCC.

d. Causal inference of gene systems network in hepatocellular carcinoma progression by graphical chain model.

Sachiyo Aburatani, Shigeru Saito[1], Masao Honda[2], Shu-ichi Kaneko[2], Katsuhisa Horimoto: [1]INFOCOM CORP., [2]Kanazawa Univ.

We analyzed the gene expression profiles of HCC and its background liver with advanced stage of CH-C, by a statistical method recently devised to infer the gene systems network from the gene expression profiles, based on the graphical chain model (GCM).

The GCM infers a causal relationship between variables that can be naturally grouped (blocks) and ordered from prior knowledge. In a GCM, any direct association between two variables in the same block is assumed to be non-causal, and any direct association between two variables from different blocks is assumed to be potentially causal. Thus, the GCM is one of the suitable models to infer the causal network between gene systems in distinct biological stages.

The procedure is as follows: 1) The genes that express characteristically in distinct stages are selected from all genes.; 2) In each stage, the profiles of genes are subjected to a clustering analysis, and the gene groups (systems) are defined in terms of biological function.; 3) The gene systems are subjected to GCM, to infer the causal network between gene systems in different stages.; 4) The causal network between gene systems are evaluated by the biological knowledge.

The expression profiles of 8516 genes were monitored in 27 samples in CH-C and 17 samples in HCC. By application of GCM to the profile data, a causal graph for the gene systems in CH-C and HCC progression were inferred from the profiles.

f. ASIAN: a web server for inferring a regulatory network framework from gene expression profiles.

Shigeru Saito[1], Kousuke Goto[1], Sachiyo Aburatani, Hiroyuki Toh[2], Katsuhisa Horimoto: [1]INFOCOM CORP., [2]Kyushu University

The standard workflow in gene expression profile analysis to identify gene function is the clustering by various metrics and techniques, and the following analyses, such as sequence analyses of upstream regions. A further challenging analysis is the inference of a gene regulatory network, and some computational methods have been intensively developed to deduce the gene regulatory network. Here, we describe our web server for inferring a framework of regulatory networks from a large number of gene expression profiles, based on graphical Gaussian modeling (GGM) in combination with hierarchical clustering (http://eureka.ims.u-tokyo.ac.jp/asian). GGM is based on a simple mathematical structure, which is the calculation of the inverse of the correlation coefficient matrix between variables, and therefore, our server can analyze a wide variety of data within a reasonable computational time. The server allows users to input the expression profiles, and it outputs the dendrogram of genes by several hierarchical clustering techniques, the cluster number estimated by a stopping rule for hierarchical clustering, and the network between the clusters by GGM, with the respective graphical presentations. Thus, the ASIAN web server provides an initial basis for inferring regulatory relationships, in that the clustering serves as the first step toward identifying the gene function.

2. Algebraic Approaches

a. Symbolic-numeric optimization for biological kinetics by quantifier elimination.

Shigeo Orii[1], Hazuhiro Anai[1], Katsuhisa Horimoto: [1]FUJITSU LTD.

We introduce a new approach to optimization for biological kinetics that deals with numerical data by symbolic quantifier elimination (QE). In this study, we illustrate the feasibility of the symbolic-numeric method in comparison with previous numerical methods.

The symbolic-numeric approach is applied to an optimization problem estimating five reaction parameters to fit a simulated signal with the five parameters to observed one, in the model described by ODE for the mechanism of irreversible inhibition of HIV protainase.

The reaction parameters $k''(k''_{22}, k''_{3}, k''_{42}, k''_{52}, k''_{6}$ with minimum SSq is the same magnitude as those by the numerical optimization methods in previous studies. Furthermore, the present method has the following merits: 1) The model parameters $\overline{k'}(i, j)$ and $\overline{k''}$ are estimated with a few points (e.g. two points) of the observed signal. 2) Feasible ranges of $\overline{k'}(i, j)$ and $\overline{k''}(i, J)$ are selected because unfeasible region can be confirmed exactly by the result "false" obtained by QE. 3) Our method enables us to estimate exactly how much the uncertainties of numerical simulation and observation should be so that the constraints become feasible. 4) The symbolic-numeric approach provides feasible ranges of reaction parameters.

b. On conditions for morphogenetic diversity of multicellular organisms.

Hiroshi Yoshida, Shigeo Orii[1], Hazuhiro Anai[1], Katsuhisa Horimoto: [1]FUJITSU LTD.

In a multicellular organism, a single cell-an egg-or a group of cells develops into a certain pattern with a variety of cell types. The variety of cell types are created through cell differentiation; the differentiation starts from an initial type, and then the initial type changes into several intermediate types before differentiating into the final type. Theoretical study of cell differentiation and morphogenesis was pioneered by Alan Turing, who showed that a reaction-diffusion system can produce an inhomogeneous stable pattern. Turing's theory provides a basis of dynamical system for morphogenesis and potentiality of cell differentiation. Embryogenesis with *increases of cell numbers* was, however, not studied. By considering Turing's study and intracellular dynamics, together with the cell division process to increase the cell numbers, Kaneko and Yomo have proposed *isologous diversification.* This allows the spontaneous cell differentiation through cell division processes and cell-cell interactions. These studies provide a basis for morphogenetic diversity of multicellular organisms. However, relevance of proliferation rates and transition rates between cell types to morphogenetic diversity has not been studied. In this paper, we shall answer this question by constructing the model based on probabilistic Lindenmayer system with interaction and by using quantifier elimination (abbreviated to QE).

Publications

Aburatani, S., Goto, K., Saito, S., Toh, H. and Horimoto, K. ASIAN: A Web Server for Inferring a Regulatory Network Framework from Gene Expression Profiles. *Nucleic Acids Res.*, 33, W659-W664, 2005.

Aburatani, S., Sakai, H., Murakami, H. and Horimoto, K. Elucidation of the relationships among the LexA-regulated genes in SOS response. *Proceedings of the 9th World Multi-Conference on Systemics, Cybernetics and Informatics*, 8, 1-5, 2005.

Aburatani, S., Sakai, H., Murakami, H. and Horimoto, K. Elucidation of the relationships among the LexA-regulated genes in SOS response. *Genome Informatics.* 16, 95-105, 2005.

Aburatani, S., Saito, S., Toh, H. and Horimoto, K. Graphical Models for Gene Expression Analyses. *Algebraic Biology 2005*, 1, 157-171, 2005.

Aburatani, S., Saito, S., Toh, H. and Horimoto, K. A Graphical Chain Modeling Approach for Analyzing Gene Expression Profiles. *Stat. Method.*, (in press).

Aburatani, S. and Horimoto, K. Statistical analysis of the relationships between LexA-regulated genes from expression profiles. *Res. Commun. Biochem. Cell & Mol. Biol.* (in press).

Dukka Bahadur K.C., Tomita, E., Suzuki, J., Horimoto, K. and Akutsu, T. Protein Threading with Profiles and Distance Constraints Using Clique Based Algorithms. *J. Bioinfo. Comput. Biol.* (in press).

Murakami, H., Sakai, H., Aburatani, S. and Horimoto, K. Relationship between Segmental Duplications and Repeat Sequences in Human Chromosome 7. *Genome Informatics*, 16, 13-21, 2005.

Murakami, H., Sakai, H., Aburatani, S. and Horimoto, K. Relationship between Segmental Duplications and Repeat Sequences in Human Chromosome 7. *Proceedings of the 9th World Multi-Conference on Systemics, Cybernetics and Informatics*, 8, 10-14, 2005.

Orii, S. Anai, H. and Horimoto, K. Symbolic-Numeric Optimization by Quantifier Elimination: an Application to Biological Kinetic Model. *Proceedings of the 9th World Multi-Conference on Systemics, Cybernetics and Informatics*, 8, 15-20, 2005.

Orii, S., Anai, H. and Horimoto, K. Symbolic-numeric estimation of parameters in biochemical models by quantifier elimination. *Bioinfo 2005*, 272-277, 2005.

Orii, S., Horimoto, K. and Anai, H., A New Approach for Symbolic-Numeric Optimization in Biological Kinetic Models. *Algebraic Biology 2005*, 1, 85-95, 2005.

Sakai, H., Murakami, H., Aburatani, S., Horimoto, K. and Kanehisa, M. Bayesian Approach for Sequence Pattern Search in Tissue Specific Alternative Splicing. *Proceedings of the 9th World Multi-Conference on Systemics, Cybernetics and Informatics*, 8, 25-30, 2005.

Saito, S., Aburatani, S. and Horimoto, K. Network Inference Tool on Personal Computer. *Proceedings of the 9th World Multi-Conference on Systemics, Cybernetics and Informatics*, 8, 21-24, 2005.

Sato, K and Horimoto, K. Comparison between Profile-Profile Methods based on Hidden Markov Models and Multiple Alignments. *Proceedings of the 9th World Multi-Conference on Systemics, Cybernetics and Informatics*, 8, 31-37, 2005.

Yoshida, H, Anai, H., Orii, S. and Horimoto, K. Inquiry into conditions for cell-type diversity of multicellular organisms by quantifier elimination. *Algebraic Biology 2005*, 1, 105-113, 2005.