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 and Minoru Kanehisa

KEGG DAS is a DAS (Distributed Annotation System) service for all organisms in the GENOME and GENES databases in KEGG (Kyoto Encyclopedia of Genes and Genomes). It started as part of the standardization efforts along with KEGG API and KGML, which are a SOAP based KEGG web service and XML representation of KEGG pathways, respectively. In addition to the genome sequences, the KEGG DAS server also provides KEGG annotations for each gene linked to the PATHWAY and LIGAND databases, as well as the SSDB database containing paralog, ortholog and motif information. We have been developing the server based on open source software including BioRuby, BioPerl, BioDAS and GMOD/GBrowse to make the system consistent with the existing open standards. Thus, the contents of the KEGG DAS server can be accessed programatically by the DAS protocol and graphically in a web browser using GBrowse. BioDAS, 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 with similar expression patterns on the same pathway can be written very easily. Furthermore, GBrowse provides a graphical web interface for the KEGG DAS server to browse, search, zoom and visualize a particular region of the genome. Thanks to the powerful feature of the GBrowse, users are also able to add their own annotations onto the KEGG DAS view by uploading their data. This enables researchers to annotate the genome by themselves, and by sharing the annotations with the community they can continuously refine the annotation. This is so-called "community annotation." KEGG DAS is available at http://das.hgc.jp/.

2. Automatic assignments of orthologs and paralogs in complete genomes

Toshiaki Katayama and Minoru Kanehisa

In addition to the human intensive efforts in the community databases, we are developing a computational method for automating genome annotations. It is based on a graph analysis of the KEGG SSDB database, containing sequence similarity relations among all the genes in the completely sequenced genomes. The nodes of the SSDB graph are genes (currently about 680,000 genes in over 200 genomes) and the edges are the Smith-Waterman sequence similarity scores computed by the SSEARCH program (currently over 300 million edges above the threshold score of 100). The edges are not only weighted but also directed, indicating the best (top-scoring) hit when a gene in an organism is compared against all genes in another organism. Thus, a highly connected cluster of nodes containing a number of bidirectional best hits might be considered an ortholog cluster consisting of functionally identical genes. Such a cluster can be found by our heuristic method for finding "quasi-cliques", but the SSDB graph is too large to perform quasi-clique finding at a time. Therefore, we introduce a hierarchy (evolutionary relationship) of organisms and treat the SSDB graph as a nested graph. The automatic decomposition of the SSDB graph into a set of quasi-cliques results in the KEGG OC (Ortholog Cluster) database, which is regularly updated and made avaiable at http://www.genome.jp/kegg-bin/OC/count/lookup.

3. SOAP/WSDL interface for the KEGG system

Shuichi Kawashima, Toshiaki Katayama and Minoru Kanehisa

We have started developing a new web service, named KEGG API, 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 universe of genes and proteins (GENES/SSDB databases), and the information about the universe of chemical compounds and reactions (LIGAND database). The KEGG API provides valuable means for accessing the KEGG system, such as searching and computing biochemical pathways in cellular processes or analyzing the universe of genes in the completely sequenced genomes. The users access the KEGG API server using the SOAP technology over the HTTP. The SOAP server also comes with WSDL, which makes it easy to build a client library for a specific computer language. This enables the users to write their own programs for many different purposes and to automate the procedure of accessing the KEGG API server and retrieving the results. Over the first year after the initial release of the KEGG API, we have received numerous feedbacks beyond our expectations. Making use of these suggestions, we are continuously developing the KEGG API to refine the interface to the KEGG system. Recent key changes include the following: (1) Methods for retrieving the pathway information have been re-organized for consistency. (2) Wrapper methods for the DBGET/LinkDB system have been fully incorporated to enable complicated database search. DBGET/LinkDB serves as a database retrieval system behind the KEGG system. (3) Several new methods (e.g. methods for retrieving a set of homolog genes) have been added. (4) To control the number of results, 'start' and 'max_results' arguments were introduced to the methods that return vast amounts of results at once. The KEGG API is available at http://www.genome.jp/kegg/soap/.

4. High performance database entry retrieval system

Kazutomo Ushijima, Chiharu Kawagoe, Toshiaki Katayama, Shuichi Kawashima, Hideo Bannai, 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 43,918,359 (Release 144). Although we have provided a simple entry retrieval system covering several major databases for several years at the Human Genome Center, searches have come to require quite a long time. Therefore, in order to continue to utilize the rapidly growing databases, it has become a critical issue to develop a new data retrieval system that can scan the whole database at a high speed. Furthermore, as the number of matched entries by a simple keyword search increases along with the growth of the database size, a new mechanism is needed to narrow down the results by applying a field specific search. To meet these needs, we have developed a new 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., and is publicly accessible at http://higet.hgc.jp/. 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 and OMIM. We are planning to include other valuable databases (such as PDB and RefSeq) and also planning to develop an interdatabase search interface and a complex search facility combining keyword search and sequence similarity search.

5. Development of algorithms for searching similar patterns in molecular structures

Tetsuo Shibuya, Michihiro Araki and Minoru Kanehisa

Finding similar patterns in biological sequences of proteins and DNAs is a critical step in identifying biological functions of these molecules. Algorithms to search sequence similarities were developed in conjunction with the availability of DNA and protein databases. In recent years carbohydrate sequence data have become publicly available, such as in the KEGG GLYCAN database, and algorithms to search similar tree structures have also been developed. Furthermore, with the emphasis on chemical genomics, databases of chemical compounds are now rapidly expanding. Thus, it is desired to find compounds with similar chemical structures from the databases and to elucidate various types of biological functions of given compounds. However, exisiting structure comparison methods for organic compounds, not to say methods for similarity search on chemical structure databases, are not well-suited for understanding biological functions. Hence we are developing new algorithms and data structures for indexing and/or searching structures of various kinds, including compounds, polyketides, nonribosomal peptides, and proteins. For polyketides and nonribosomal peptides, we already developed algorithms for searching enzymes that are used to synthesize them. For proteins, we are developing a new data structure based on geometric hashing and suffix trees to achieve fast search for structurally similar proteins.

6. Chemogenomic dissection of the biosynthetic process of medicinal natural products

Michihiro Araki, Tetsuo Shibuya, and Minoru Kanehisa

Medicinal natural products are enzymically synthesized as secondary metabolites for specific biological purposes, which have been the major sources of bioactive compounds with diverse pharmacological activities. In order to make full use of the potential of natural products as research tools as well as drug leads on the context of metabolic engineering, it is of great importance to understand the biosynthetic strategies with the integrative computational analyses of chemical and genomic information. An increasing amount of such information becomes available to allow us to extract the chemical design principles of the natural products coded on genomic information. We have thus started identifying the system structures of the biosynthetic devices by merging the chemical information of molecular building blocks and module structures into the genomic information of the biosynthetic pathway.

Publications

Kanehisa, M, Goto, S, Kawashima, S, Okuno, Y, Hattori, M: The KEGG resource for deciphering the genome. Nucleic Acids Res., 32: D277-80, 2004.

Okamoto, S, Kawashima, S, Narikawa, R, Kanehisa, M: Correlation between signal transduction domains and habitats in cyanobacteria. Genome Informatics, 15: P054, 2004.

Moriya, Y, Katayama, T, Nakaya, A, Itoh, M, Yoshizawa, A, Okuda, S, Kanehisa M: Automatic generation of KEGG OC (Ortholog Cluster) and its assignment to draft genomes, Genome Informatics, 15: P049, 2004.

Masoudi-Nejad, A, Jauregui, R, Kawashima, S, Goto, S, Kanehisa, M, Endo, TR: The kingdom of plantae EST indices: a resource for plant genomics community, Genome Informatics, 15: P102, 2004.

Yoshizawa, A, Okuda, S, Moriya, Y, Itoh, M, Katayama, T, Kawano, S, Okamoto, S, Kawashima, S, Kanehisa, M: Functional classification of coiled-coil proteins in multiple genomes, Genome Informatics, 15: P157, 2004.

Tamada, Y, Bannai, H, Imoto, S, Katayama, T, Kanehisa, M, Miyano, S: Modeling gene networks utilizing evolutionary information using Bayesian network models, Genome Informatics, 15: P020, 2004.

Okuda, S, Yoshizawa, A, Moriya, Y, Itoh, M, Katayama, T, Goto, S, Kanehisa, M: Functional categorization of multiple genomes using KEGG OC in the Genome Indecies, Genome Informatics, 15: P050, 2004.

Yuasa, T., Hayashi, T., Ikai, N., Katayama, T., Aoki, K., Obara, T., Toyoda, Y., Maruyama, T., Kitagawa, D., Takahashi, K., Nagao, K., Nakaseko, Y. and Yanagida, M. 2004. An interactive gene network for securin-separase, condensin, cohesin, Dis1/Mtc1 and histones constructed by mass transformation, Genes Cells, 9: 1069-1082, 2004.

Shibuya, T., Kashima, H. and Konagaya, A. Accurate cDNA Clustering Algorithm based on Spliced Sequence Alignment, Bioinformatics, 20: 29-39, 2004.

Shibuya, T. Generalization of a Suffix Tree for RNA Structural Pattern Matching, Algorithmica, 39: 1-19, 2004

Kwan, A., Miyano, S., Okazaki, Y. and Shibuya, T. What are the applications and limitations of microarray datamining for immunology? First International Immunoinformatics Symposium, 2004.

Araki, M, Katayama, T, Matsuura, Y, Kanehisa, M., KEGG PEPTIDE: a database for peptide structures, IBSB 2004, 1-2, 2004.

Kobayashi, H, Kaern, M, Araki, M, Chung, K, Gardner, TS, Cantor, CR, Collins, JJ., Programmable cells: interfacing natural and engineered gene networks. Proc. Natl. Acad. Sci. U.S.A., 101: 8414-8419, 2004.

渋谷哲朗，バイオインフォマティクスにおけるデータマイニングの基礎，ゲノム研究実験ハンドブック，辻本豪三，田中利男編，羊土社，pp. 45–48，2004.

片山俊明，インターネットの利用．ゲノム研究実験ハンドブック，辻本豪三，田中利男編，羊土社．pp. 20–29，2004.

川島秀一，プログラムからのデータベースアクセス．ゲノム研究実験ハンドブック，辻本豪三，田中利男編，羊土社，pp. 109–113，2004.

荒木通啓，服部正泰，金久實，バイオインフォマティクスとケモインフォマティクスの融合，現代医療，Vol. 36，№5，pp. 86–92，2004

Human Genome Center

Laboratory of DNA Information Analysis
DNA情報解析分野

Professor	Satoru Miyano, Ph.D.
Research Associate	Seiya Imoto, Ph.D.
Research Associate	Hideo Bannai, Ph.D.
Instructor	Michiel J.L. de Hoon

教授	理学博士	宮野 悟
助手	理学博士	井元清哉
助手	博士 (情報理工学)	坂内英夫
特任教員	Ph.D.	Michiel J.L. de Hoon

The aim of the research at this laboratory is to establish computational methodologies for discovering and interpreting information of nucleic acid sequences, proteins and some other experimental data arising from researches in Genome Science. Our current concern is 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. Using protein-protein interactions for refining gene networks estimated from microarray data by Bayesian networks

Naoki Nariai, Sunyong Kim, Seiya Imoto, Satoru Miyano

We propose a statistical method to estimate gene networks from DNA microarray data and protein-protein interactions. Because physical interactions between proteins or multiprotein complexes are likely to regulate biological processes, using only mRNA expression data is not sufficient for estimating a gene network accurately. Our method adds knowledge about protein-protein interactions to the estimation method of gene networks under a Bayesian statistical framework. In the estimated gene network, a protein complex is modeled as a virtual node based on principal component analysis. We show the effectiveness of the proposed method through the analysis of *Saccharomyces cerevisiae* cell cycle data. The proposed method improves the accuracy of the estimated gene networks, and successfully identifies some biological facts.

b. Superiority of network motifs over optimal networks and an application to the revelation of gene network evolution

Sascha Ott, Annika Hansen[1], Sunyong Kim, and Satoru Miyano: [1]Wolfson Institute for Biomedical Research, University College London

Estimating the network of regulative interactions between genes from gene expression measurements is a major challenge. Recently, we have shown that for gene networks of up to around 35 genes, optimal network models can be computed. However, even optimal gene network models will in general contain false edges, since the expression data will not unambiguously point to a single network. In order to overcome this problem, we present a computational method to enumerate the most likely m networks and to extract a widely common subgraph (denoted as gene network motif) from these. We apply the method to bacterial gene expression data and extensively compare estimation results to knowledge. Our results reveal that gene network motifs are in significantly better agreement to biological knowledge than optimal network models. We also confirm this observation in a series of estimations using synthetic microarray data and compare estimations by our method with previous estimations for yeast. Furthermore, we use our method to estimate similarities and differences of the gene networks that regulate tryptophan metabolism in two related species and thereby demonstrate the analysis of gene network evolution.

c. Finding optimal models for small gene networks

Sascha Ott, Seiya Imoto, Satoru Miyano

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

d. Bayesian network and radial basis function network regression for nonlinear modeling of genetic network

Tomohiro Ando, Seiya Imoto, Satoru Miyano

A new method for constructing gene networks from microarray gene expression data is proposed using Bayesian networks and radial basis function network regression models. An essential point of Bayesian network modeling is the construction of the conditional distribution of each random variable. Unfortunately, many statistical methods often fail in capturing biological systems when a target gene receives complicated effects from its parent genes. To capture the complicated biological systems, radial basis function network regression models are utilized. We analyze *Saccharomyces cerevisiae* gene expression data and evaluate the resulting network by comparing with biological knowledge. We also use Monte Carlo simulations to investigate the properties of the proposed method. Real data analysis detects complicated effects as well as a set of parent genes from a large set of candidates even when there are complicated relationship between a target gene and its parent genes.

e. Functional data analysis of the dynamics of gene regulatory networks

Tomohiro Ando, Seiya Imoto, Satoru Miyano

A new method for constructing gene networks from microarray time-series gene expression data is proposed in the context of Bayesian network approach. An essential point of Bayesian network modeling is the construction of the conditional distribution of each random variable. When estimating the conditional distributions from gene expression data, a common problem is that gene expression data contain multiple missing values. Unfortunately, many methods for constructing conditional distributions require a complete gene expression value and many loose effectiveness even with a few missing value. Additionally, they treat microarray time-series gene expression data as static data, although time can be an important factor that affects the gene expression levels. We overcome these difficulties by using the method of functional data analysis. The proposed network construction method consists of two stages. Firstly, discrete microarray time-series gene expression values are expressed as a continuous curve of time. To account for the time dependency of gene expression measurements and the noisy nature of the microarray data, P-spline nonlinear regression models are utilized. After this preprocessing step, the conditional distribution of each random variable is constructed based on functional linear regression models. The effectiveness of the proposed method is investigated through Monte Carlo simulations and the analysis of *Saccharomyces cerevisiae* gene expression data.

f. Constructing biological pathway models with hybrid functional Petri nets

Atsushi Doi, Sachie Fujita[2], Hiroshi Matsuno[2], Masao Nagasaki, Satoru Miyano: Faculty of Science, Yamaguchi University

In many research projects on modeling and analyzing biological pathways, the Petri net has been recognized as a promising method for representing biological pathways. From the pioneering works by Reddy et al. (1993), and

Hofest? dt (1994), that model metabolic pathways by traditional Petri net, several enhanced Petri nets such as colored Petri net, stochastic Petri net, and hybrid Petri net have been used for modeling biological phenomena. Recently, Matsuno et al. ((PSB 8: 152-163, 2003), introduced the hybrid functional Petri net (HFPN) in order to give a more intuitive and natural modeling method for biological pathways than these existing Petri nets. Although the paper demonstrates the effectiveness of HFPN with two examples of gene regulation mechanism for circadian rhythms and apoptosis signaling pathway, there has been no detailed explanation about the method of HFPN construction for these examples. The purpose of this paper is to describe method to construct biological pathways with the HFPN step-by-step. The method is demonstrated by the well-known glycolytic pathway controlled by the lac operon gene regulatory mechanism.

g. Modeling and simulation of fission yeast cell cycle on hybrid functional Petri net

Sachie Fujita[2], Mika Matsui[2], Hiroshi Matsuno[2], Satoru Miyano

Through many researches on modeling and analyzing biological pathways, Petri net has been recognized as a promising method for representing biological pathways. Recently, Matsuno et al. (PSB 8: 152-163, 2003) introduced hybrid functional Petri net (HFPN) for giving more intuitive and natural biological pathway modeling method than existing Petri nets. They also developed Genomic Object Net (GON) which employs the HFPN as a basic architecture. Many kinds of biological pathways have been modeled with the HFPN and simulated by GON. This paper gives a new HFPN model of cell cycle of fission yeast with giving six basic HFPN components of typical biological reactions, and demonstrating the method how biological pathways can be modeled with these HPFN components. Simulation results by GON suggest a new hypothesis which will help biologists for performing further experiments.

h. Simulated cell division processes of the Xenopus cell cycle pathway by Genomic Object Net

Mika Matsui[2], Sachie Fujita[2], Shunichi Suzuki[2], Hiroshi Matsuno[2], Satoru Miyano

Matsuno et al. (PSB 8: 152-163, 2003) modeled and simulated a Drosophila multicellular patterning by Delta-Notch signaling pathway by using a software “Genomic Object Net” which is developed based on hybrid functional Petri net (HFPN) architecture. However, in this model, cellular formation is fixed throughout the simulation. Then, this paper constructs an HFPN model of *Xenopus* cell cycle pathway which includes the mechanism for cell division control as well as checkpoint processes. With this model, dynamic cell division processes of *Xenopus* early embryo including the changes in cell division cycles from synchronous to asynchronous are simulated.

i. A versatile Petri net based architecture for modeling and simulation of complex biological processes

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

The research on modeling and simulation of complex biological systems is getting more important in Systems Biology. In this respect, we have developed Hybrid Function Petri net (HFPN) that was newly developed from existing Petri net because of their intuitive graphical representation and their capabilities for mathematical analyses. However, in the process of modeling metabolic, gene regulatory or signal transduction pathways with the architecture, we have realized three extensions of HFPN, (i) an entity should be extended to contain more than one value, (ii) an entity should be extended to handle other primitive types, e.g. Boolean, string, (iii) an entity should be extended to handle more advanced type called object that consists of variables and methods, are necessary for modeling biological systems with Petri net based architecture. To deal with it, we define a new enhanced Petri net called hybrid functional Petri net with extension (HFPNe). To demonstrate the effectiveness of the enhancements, we model and simulate with HFPNe four biological processes that are difficult to represent with the previous architecture HFPN.

j. Integrating biopathway databases for large-scale modeling and simulation

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

Biopathway databases have been developed, such as KEGG and EcoCyc that compile interaction structures of biopathways together with biological annotations. However, these biopathways are not directly editable and simulatable. Thus, we have made an application, the Biopathway Executer (BPE) that reconstructs these two major biopathway databases to XML formats of modeling and simulation platforms. BPE is developed with JAVA and has a database of executable biopathways that integrates some parts of biopathway information, KEGG and BioCyc, and other databases, e.g. MIPS and BRENDA. Currently, BPE employs the XML format (GONML) of a Hybrid Functional Petri net (HFPN) for the output. The features of HFPN are: (i) biopathways that contain discrete and continuous processes can be modeled, (ii) all biopathways that are modeled with ordinary differential equations (ODEs) can be remodeled, (iii) biopathways can be modeled while keeping readability by human. Other XML formats of biopathways, SBML and CellML are subsets of GONML. Thus, BPE can bridge major biopathway databases and major modeling and simulating softwares. To demonstrate the effectiveness/ usability of BPE, four examples are created and simulated on Genomic Object Net which is based on the HFPN architecture; (i) executable KEGG maps while keeping the features of original maps, (ii) executable BioCyc maps while keeping the features of original maps, (iii) large-scale editable and simulatable KEGG metabolic pathways, (iv) a metabolic pathway with gene regulatory networks. These examples show that BPE is a useful tool for integrating biopathway databases for large scale modeling and simulation.

k. Predicting the operon structure of *Bacillus subtilis* using operon length, intergene distance, and gene expression information

Michiel J.L. de Hoon, Seiya Imoto, Kazuo Kobayashi[3], Naotake Ogasawara[3], Satoru Miyano: [3]Nara Institute of Science and Technology

We predict the operon structure of the *Bacillus subtilis* genome using the average operon length, the distance between genes in base pairs, and the similarity in gene expression measured in time course and gene disruptant experiments. By expressing the operon prediction for each method as a Bayesian probability, we are able to combine the four prediction methods into a Bayesian classifier in a statistically rigorous manner. The discriminant value for the Bayesian classifier can be chosen by considering the associated cost of misclassifying an operon or a nonoperon gene pair. For equal costs, an overall accuracy of 88.7% was found in a leave one-out analysis for the joint Bayesian classifier, whereas the individual information sources yielded accuracies of 58.1%, 83.1%, 77.3%, and 71.8% respectively. The predicted operon structure based on the joint Bayesian classifier is available from the DBTBS database (http://dbtbs.hgc.jp).

l. Predicting gene regulation by sigma factors in *Bacillus subtilis* from genome-wide data

Michiel J.L. de Hoon, Yuko Makita, Seiya Imoto, Kazuo Kobayashi[3], Naotake Ogasawara[3], Kenta Nakai, Satoru Miyano

Sigma factors regulate the expression of genes in *Bacillus subtilis* at the transcriptional level. We assess the accuracy of a fold-change analysis, Bayesian networks, dynamic models and supervised learning based on coregulation in predicting gene regulation by sigma factors from gene expression data. To improve the prediction accuracy, we combine sequence information with expression data by adding their log-likelihood scores and by using a logistic regression model. We use the resulting score function to discover currently unknown gene regulations by sigma factors. The coregulation-based supervised learning method gave the most accurate prediction of sigma factors from expression data. We found that the logistic regression model effectively combines expression data with sequence information. In a genome-wide search, highly significant logistic regression scores were found for several genes whose transcriptional regulation is currently unknown. We provide the corresponding RNA polymerase binding sites to enable a straightforward experimental verification of these predictions.

m. A neural network method for identification of RNA-interacting residues in protein

Euna Jeong, I-Fang Chung, Satoru Miyano

Identification of the most putative RNA-interacting residues in protein is an important and challenging problem in a field of molecular recognition. Structural analysis of protein-RNA complexes reveals a strong correlation between interaction residues and their structure. Building on this viewpoint, we have developed a neural network predictor to correctly identify residues involved in protein-RNA interactions from protein sequence and its structural information. The system has been exhaustedly cross-validated with various strategies differing in input encoding, amount of input information, and network architectures. In addition, we have evaluated performance among functional subsets of complexes. Finally, to reflect the properties of protein-RNA complexes in our dataset, two kinds of post-processing method are adopted. The experimental result shows that our system yields not-trivial performance although the residues in interaction sites are too scarce.

2. Algorithmic and Statistical Methods for Bioinformatics

a. Finding optimal pairs of patterns

Hideo Bannai, Heikki Hyyrö[4], Ayumi Shinohara[4], Masayuki Takeda[4], Kenta Nakai, Satoru Miyano: [4]Department of Informatics, Kyushu University

We consider the problem of finding the optimal pair of string patterns for discriminating between two sets of strings, i.e. finding the pair of patterns that is best with respect to some appropriate scoring function that gives higher scores to pattern pairs which occur more in the strings of one set, but less in the other. We present an $O(N^2)$ time algorithm for finding the optimal pair of substring patterns, where N is the total length of the strings. The algorithm looks for all possible Boolean combination of the patterns, e. g. patterns of the form p ˆ _q, which indicates that the pattern pair is considered to match a given string s, if p occurs in s, AND q does NOT occur in s. The same algorithm can be applied to a variant of the problem where we are given a single set of sequences along with a numeric attribute assigned to each sequence, and the problem is to find the optimal pattern pair whose occurrence in the sequences is correlated with this numeric attribute. An efficient implementation based on suffix arrays is presented, and the algorithm is applied to several nucleotide sequence datasets of moderate size, combined with microarray gene expression data, aiming to find regulatory elements that cooperate, complement, or compete with each other in enhancing and/or silencing certain genomic functions.

b. Case-control study of binary trait considering interactions between SNPs and environmental effects using logistic regression

Reiichro Nakamichi, Seiya Imoto, Satoru Miyano

In this paper, we propose a combination of logistic regression and genetic algorithm for the association study of the binary disease trait. We use a logistic regression model to describe the relation of multiple SNPs, environments and the target binary trait. 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 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 propose 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 patients. We succeeded in finding an efficient set of to predict types of diabetes and some SNPs which have strong interaction to age while it is not significant as a single locus.

c. Kernel mixture survival models for identifying cancer subtypes, predicting patient's cancer types and survival probabilities

Tomohiro Ando, Seiya Imoto, Satoru Miyano

One important application of microarray gene expression data is to study the relationship between the clinical phenotype of cancer patients and gene expression profiles on the whole-genome scale. The clinical phenotype includes several different types of cancers, survival times, relapse times, drug responses and so on. Under the situation that the subtypes of cancer have not been previously identified or known to exist, we develop a new kernel mixture modeling method that performs simultaneously identification of the subtype of cancer, prediction of the probabilities of both cancer type and patient's survival, and detection of a set of marker genes on which to base a diagnosis. The proposed method is successfully performed on real data analysis and simulation studies.

d. A mixed factors model for dimension reduction and extraction of a group structure in gene expression data

Ryo Yoshida[5], Tomoyuki Higuchi[5], Seiya Imoto: [5]Institute of Statistical Mathematics

When we cluster tissue samples on the basis of genes, the number of observations to be grouped is much smaller than the dimension of feature vector. In such a case, the applicability of conventional model-based clustering is limited since the high dimensionality of feature vector leads to overfitting during the density estimation process. To overcome such difficulty, we attempt a methodological extension of the factor analysis. Our approach enables us not only to prevent from the occurrence of overfitting, but also to handle the issues of clustering, data compression and extracting a set of genes to be relevant to explain the group structure. The potential usefulness is demonstrated with the application to the leukemia dataset.

Publications

Akutsu, T., Bannai, H., Miyano, S., Ott, S. On the complexity of deriving position specific score matrices from positive and negative sequences. Discrete Applied Mathematics. In press.

Ando, T., Imoto, S., Miyano, S. Bayesian network and radial basis function network regression for nonlinear modeling of genetic network. Proc. Third International Conference on Information. 561-564, 2004.

Ando, T., Imoto, S., Miyano, S. Functional data analysis of the dynamics of gene regulatory networks. Proc. Knowledge Exploration in Life Science Informatics KELSI2004. Lecture Notes in Artificial Intelligence. 3303: 69-83, 2004.

Ando, T., Imoto, S., Miyano, S. Kernel mixture survival models for identifying cancer subtypes, predicting patient's cancer types and survival probabilities. Genome Informatics. 15 (2): 201-210, 2004.

Araki, Y., Konishi, S., Imoto, S. Functional discriminant analysis for time-seriese gene expression data via radial basis function expansion. Proc. COMPSTAT 2004, 613-620, 2004. Physica-Verlag/Springer.

Bannai, H., Hyyrö, H., Shinohara, A., Takeda, M., Nakai, K., Miyano, S. An $O(N^2)$ algorithm for discovering optimal Boolean pattern pairs. IEEE/ACM Transactions on Computational Biology and Bioinformatics. In press.

Bannai, H., Hyyrö, H., Shinohara, A., Takeda, M., Nakai, K., Miyano, S. Finding optimal pairs of patterns. Proc. 4th International Workshop on Algorithms in Bioinformatics (WABI 2004). Lecture Notes in Bioinformatics. 3240: 450-462, 2004.

Bannai, H., Inenaga, S., Shinohara, A., Takeda, M., Miyano, S. Efficiently finding regulatory elements using correlation with gene expression. J. Bioinformatics and Computational Biology. 2(2): 273-288, 2004.

De Hoon, M.J.L., Imoto, S., Kobayashi, K., Ogasawara, N., Miyano, S. Predicting the operon structure of *Bacillus subtilis* using operon length, intergene distance, and gene expression information. Pacific Symposium on Biocomputing. 9: 276-287, 2004.

De Hoon, M.J.L., Imoto, S., Nolan, J., Miyano, S. Open source clustering software. Bioinformatics. 20(9): 1453-1454, 2004.

De Hoon, M.J.L., Makita, Y., Imoto, S., Kobayashi, K., Ogasawara, N., Nakai, K., Miyano, S. Predicting gene regulation by sigma factors in *Bacillus subtilis* from genome-wide data. Bioinformatics, 20 (Suppl. 1): i101-i108, 2004.

Doi, A., Fujita, S., Matsuno, H., Nagasaki, M., Miyano, S. Constructing biological pathway models with hybrid functional Petri nets. In Silico Biology. 4(3): 271-291, 2004.

Fujita, S., Matsui, M., Matsuno, H., Miyano, S. Modeling and simulation of fission yeast cell cycle on hybrid functional Petri net. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences. E87-A(11): 2919-2928, 2004.

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. 1st International Workshop on Data Mining and Bioinformatics. Lecture Note in Comupter Science. In press.

Imoto, S., Higuchi, T., Goto, T., Tashiro, K., Kuhara, S., Miyano, S. Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. J. Bioinformatics and Computational Biology. 2(1): 77-98, 2004.

Imoto, S., Higuchi, T., Kim, S., Jeong, E., Miyano, S. Residual bootstrapping and median filtering for robust estimation of gene networks from microarray data. Proc. 2nd Computational Methods in Systems Biology. Lecture Note in Bioinformatics. In press.

Inenaga, S., Bannai, H., Hyyrö, H., Shinohara, A., Takeda, M., Nakai, K., Miyano, S. Finding optimal pairs of cooperative and competing patterns with bounded distance. Proc. 7th International Conference on Discovery Science (DS 2004). Lecture Notes in Artifitial Intelligence. 3245: 32-46, 2004.

Jeong, E., Chung, I., Miyano, S. A neural network method for identification of RNA-interacting residues in protein. Genome Informatics. 15(1): 105-116, 2004.

Kim, S., Imoto, S., Miyano, S. Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data. Biosystems. 75(1-3): 57-65, 2004.

Konishi, S., Ando, T., Imoto, S. Bayesian information criteria and smoothing parameter selection in radial basis function networks. Biometrika. 91(1): 27-43, 2004.

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

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. Proc. The 3 rd Asia-Pacific Bioinformatics Conference 2005, Imperial College Press, In press.

Miyano, S. Computational systems biology. Proc. Third International Conference on Information (Li, L. and Yen, K.K., eds.). 9-14, 2004.

Nagasaki, M., Doi, A., Matsuno, H., Miyano, S. Genomic Object Net: a platform for modeling and simulating biopathways. Applied Bioinformatics. 2: 181-184, 2004.

Nagasaki, M., Doi, A., Matsuno, H., Miyano, S. Integrating biopathway databases for large-scale modeling and simulation. Proc. Second Asia-Pacific Bioinformatics Conference (APBC 2004) (Y.P. Chen, ed). Conferences in Research and Practice in Information Technology. 29: 43-52, 2004.

Nagasaki, M., Doi, A., Matsuno, H., Miyano, S. A versatile Petri net based architecture for modeling and simulation of complex biological processes. Genome Informatics, 15(1): 180-197, 2004.

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. Case-control study of binary trait considering interactions between SNPs and environmental effects using logistic regression. Proc. 4th IEEE Bioinformatics and Bioengineering. 73-78, 2004. IEEE Press.

Nakano, M., Noda, R., Kitakaze, H., Matsuno, H., Miyano, S. XML pathway file conversion between Genomic Object Net and SBML. Proc. The Third International Conference on Information. 585-588, 2004.

Nariai, N., Kim, S., Imoto, S., Miyano, S. Using protein-protein interactions for refining gene networks estimated from microarray data by Bayesian networks. Pacific Symposium on Biocomputing. 9: 336-347, 2004.

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., Imoto, S., Miyano, S. Finding optimal models for small gene networks. Pacific Symposium on Biocomputing. 9: 557-567, 2004.

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

Takei, Y., Inoue, K., Ogoshi, M., Kawahara, T., Bannai, H., and Miyano, S. Identification of novel adrenomedullin in mammals: a potent cardiovascular and renal regulator. FEBS Letters, 556: 53-58, 2004.

Yoshida, R., Higuchi, T., Imoto, S. A mixed factors model for dimension reduction and extraction of a group structure in gene expression data. Proc. 3rd Computational Systems Bioinformatics. 161-172, 2004. IEEE Press.

Yoshida, R., Imoto, S., Higuchi, T. A penalized likelihood estimation on transcriptional module-based clustering. Proc. 1st International Workshop on Data Mining and Bioinformatics. Lecture Note in Comupter Science. In press.

土井淳，長崎正朗，松野浩嗣，宮野悟．Genomic Object Netによるパスウエイの表現とシミュレーション．ゲノミクス・プロテオミクスの新展開～生物情報の解析と応用～(今中忠之編)．エヌ・ティー・エス．930–937，2004.

土井淳，長崎正朗，松野浩嗣，宮野悟．生命パスウエイのモデル化・可視化技術と創薬研究への応用．月刊薬事．46(7)：1265–1272，2004.

宮野悟．ゲノムからバイオインフォマティクスへ．現代医療．36(5)：1018–1021，2004.

宮野悟，松野浩嗣，倉田博之．システム生物学．ゲノム研究実験ハンドブック(辻本豪三，田中利男編集)，49–54，羊土社，2004.

宮野悟．バイオパスウェイシミュレーション—医学とコンピュータ．Molecular Medicine．41：328–331，2004.

Human Genome Center

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

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

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 playing significant roles in human cancer

a. Genes that are inducible by p53

Yoshio Anazawa, Park Woong Ryeon, Chizu Tanikawa, Kyongah Yoon, Hirohumi Arakawa, Koichi Matsuda and Yusuke Nakamura

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

dependent apoptosis, and might be a good candidate for next-generation gene

The *p53AIP1* gene, which we recently identified as a novel p53-target, mediates p53-dependent apoptosis. We evaluated the effects of adenovirus-mediated introduction of *p53AIP1* (Ad-*p53AIP1*) on 30 human cancer-cell lines *in vitro*, and two cell lines *in vivo*, in comparison with the effects of p53 (Ad-*p53*). In 20 of the 30 cell lines, p53AIP1-induced apoptosis was observed, and in 12 of these p53AIP1-sensitive cancer cell lines, the apoptotic effects of p53AIP1 were greater than those of p53 itself. Cancers with wild-type p53, which were thought to be p 53-resisitant, were likely to be sensitive to p53 AIP1-induced apoptosis. p53-resistant cancers such as LS174T (p53+/+) and A549 (p53+/+), in which no increase of *p53AIP1* mRNA expression was observed when Ad-*p53* was introduced, were killed effectively by Ad-*p53AIP1*. Furthermore, co-introduction of *p53* and *p53AIP 1* had synergistic effects on the induction of apoptosis regardless of *p53* status. Finally, adenovirus-mediated introduction of *p53AIP1* suppressed tumor growth *in vivo*. These results suggested that *p53AIP1* gene transfer might become a new strategy for the treatment of p53-resistant cancers.

We also identified the $_1$aldehyde dehydrogenase 4 (ALDH4) gene as a direct target of p53. ALDH4 is a mitochondrial matrix NAD$^+$-dependent enzyme catalyzing the second step of the proline degradation pathway. The expression of *ALDH4* mRNA was induced by various cellular stresses, including hydrogen peroxide, UV, g-irradiation or adriamycin treatment in a p 53-dependent manner. p53-dependent transcriptional activity of the binding site was confirmed by a reporter assay. Inhibition of ALDH4 expression by antisense oligonucleotides could enhance cell death induced by infection of Ad-p53. ALDH4 over-expressing cells showed significantly lower intracellular ROS level than control cells after treatment of hydrogen peroxide or UV. Those cells were also resistant to cell damage by hydrogen peroxide. These results suggest that ALDH4 is a direct-target of p53, and that p 53 might play a protective role against cell damage induced by intracellular ROS generation through transcriptional activation of ALDH4.

Although a number of p53-target genes have been identified, the mechanisms of p53-dependent activities that determine cellular survival or death are still not fully understood. Here we report 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 over-expression 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. p53CSV 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 p53CSV is one of important players in the p53-mediated cell survival.

b. Colon, Liver, and Gastric cancers

Yoichi Furukawa, Ryuji Hamamoto, Hiroshi Okabe, Li Meihua, Meiko Takahashi, Takashi Shimokawa, Kazutaka Obama, Michihiro Sakai, Katsuaki Ura, Takaaki Kobayashi, Pittella Fabio, Natini Jinawath and Yusuke Nakamura

Through a genome-wide cDNA microarray, we identified *SMYD3*, a novel gene overexpressed in the great majority of colorectal carcinomas and hepatocellular carcinomas. Introduction of *SMYD3* into NIH3T3 cells enhanced the cell growth, and its reduced expression by siRNAs in various cancer cells resulted in significant growth suppression. We found that through an interaction with an RNA helicase HELZ, SMYD3 formed a complex with RNA polymerase II and transactivated a set of genes including oncogenes, homeobox genes, and genes associated with cell cycle regulation. Subsequent analysis revealed that SMYD3 bound to a “5’-CCCTCC-3’” motif present in the promoter region of its downstream genes such as *Nkx2*. 8. A SET domain of this protein revealed histone H3-lysine 4 (H3-K4)-specific methyltransferase activity, which was enhanced in the presence of 90-kD heat shock protein (HSP90A). Our findings suggest that SMYD3 has a histone methyltransferase activity and plays an important role in transcriptional regulation as a member of RNA polymerase complex, and that its activation is one of key factors in human carcinogenesis.

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

Among the genes that were frequently transactivated in colorectal tumors, we identified a novel human gene termed *LEMD1* (*LEM domain containing1*) whose expression was elevated in 17 of 18 CRCs compared with their corresponding non-cancerous mucosae. Multiple-tissue northern blot analysis revealed that the gene was abundantly expressed in the testis in a tissue specific manner among the 16 adult normal tissues examined. Subsequent analysis identified six alternatively spliced forms of transcripts, of which one transcript of 201 bp was expressed exclusively in the CRC tissues but not in their corresponding normal tissues. Since proteins that are not expressed in normal tissues except in the testis and expressed in cancer tissues, it should serve as a promising therapeutic target for CRC treatment like immunotherapy.

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

Among the genes commonly transactivated in the cancers, we identified a novel human gene, which we termed *CLUAP1* (Clusterin Associated Protein 1). It encodes a nuclear protein of 413 amino acids containing a coiled-coil domain. To investigate its function, we searched for CLUAP 1-interacting proteins using yeast two-hybrid system and identified nuclear Clusterin. Expression of CLUAP1 was gradually increased in the late S to G2/M phases of cell cycle and it returned to the basal level in the G0/G1 phases. Suppression of this gene by siRNAs resulted in growth retardation in the transfected cells. These data provide better understanding of colorectal carcinogenesis, and inactivation of CLUAP1 may conceivably serve in the future as a novel therapeutic intervention for treatment of colon cancer.

We characterized biological importance of a novel human gene termed *RNF43* (RING finger protein 43) in colorectal carcinogenesis. Its exogenous expression conferred growth-promoting effect in COS7 and NIH3T3 cells, and suppression of its expression by specific short interfering RNAs resulted in growth suppression of colon cancer cells. Interestingly, RNF43 protein was shown to be a secreted protein and addition of the conditioned media of the RNF43-transfected cells into culture media of NIH3T3 cells revealed a significant enhancement of cell growth. We also found that RNF43 protein interacted with an extracelluler domain of Notch2 *in vitro* and *in vivo*, suggesting that RNF43 may exert its growth promoting effect through modulation of Notch2-signaling pathway in an antocrine manner. These data imply that RNF43 plays an important role in colorectal carcinogenesis and may serve as a good target molecule for development of diagnosis and novel drugs for CRCs.

Among the genes up-regulated in hepatocellular carcinomas, we focused on a novel gene, termed *WDRPUH* and characterized its biological function. *WDRPUH* encodes a predicted 620 amino acid protein containing 11 highly conserved WD40 repeat domains. Multiple-tissue northern blot analysis revealed its specific expression in the testis among the 16 normal tissues examined. Transfection of plasmids designed to express *WDRPUH*-specific siRNA significantly reduced its expression in HCC cells and resulted in growth suppression of the transfected cells. Interestingly, we found that WDRPUH associated directly with HSP70, proteins of the chaperonin-containing TCP-1 (CCT) complex, as well as BRCA2. These findings have disclosed a novel insight into hepatocarcinogenesis and suggested that WDRPUH may be a molecular target for development of new strategies to treat HCCs.

Gastric cancer is the fourth leading cause of cancer-related death in the world. Two histologically distinct types of gastric cancers, namely intestinal type and diffuse type, have different epidemiological and pathophysiological features, suggesting different mechanisms of carcinogenesis. A number of studies have been carried out to determine the molecular mechanisms of intestinal-type gastric cancer, whereas little is known about those of diffuse-type gastric cancer that has a more invasive phenotype and poorer prognosis. To clarify the mechanisms that underlie development and/or progression of diffuse-type gastric cancer, we compared the expression profiles of 20 diffuse-type gastric cancer tissues with their corresponding non-cancerous mucosae by means of cDNA microarray containing 23,040 genes in combination with laser-microbeam microdissection. We identified 153 genes commonly up-regulated and more than 500 genes commonly down-regulated. Furthermore, comparison of the expression profiles of diffuse-type with those of intestinal-type gastric cancers identified 46 genes that may represent the distinct molecular signature of each type of gastric cancer. The signature of diffuse-type cancer exhibits altered expression of genes related to cell-matrix interaction and extracellular matrix components. Although further investigation of their functions is essential, these data should help to better understand the different mechanisms underlying gastric carcinogenesis and may also provide clues to the identification of novel diagnostic markers and/or therapeutic targets of diffuse-type gastric cancer. In addition, among the 153 genes whose expression levels were elevated in cancers compared to non-cancerous mucosae, we identified a gene termed *NOL8* that encodes a putative 150-kDa protein with an RRM domain in its amino-acid terminal region. Northern blot analysis revealed that *NOL8* was abundantly expressed in skeletal muscle but not expressed in other 22 tissues examined. Immunocytochemical staining of NOL8 showed its localization in the nucleolus and subsequent protein phosphatase analysis revealed that it was present as the phophorylated form. In addition, transfection of short-interfering RNA specific to *NOL8* into three diffuse-type gastric cancer cells, St-4, MKN45 and TMK-1, effectively reduced expression of this gene and induced apoptosis in these cells. These findings provide a new insight into diffuse-type gastric carcinogenesis and may contribute to development of new therapeutic strategies for diffuse-type gastric cancer.

c. Epigenetic control of cancer

Motoko Unoki and Yusuke Nakamura

ICBP90, (inverted CCAAT box-binding protein of 90 kDa), has been reported as a regulator of *topoisomerase IIa* expression. We present evidence here that ICBP90 binds to methyl-CpG when at least one symmetrically methylated-CpG dinucleotides is presented as its recognition sequence. A SET and RING finger-associated (SRA) domain accounts for the high binding affinity of ICBP90 for methyl-CpG dinucleotides. This protein constitutes a complex with HDAC1 also via its SRA domain, and bound to methylated promoter regions of various tumor suppressor genes, including *p16*^{INK4A} and *p14*^{ARF}, in cancer cells. It has been reported that expression of ICBP90 was up-regulated by E2F-1, and we confirmed the up-regulation was caused by binding of E2F-1 to the intron1 of *ICBP90*, which contains two E2F-1 binding motifs. Our data also revealed accumulation of ICBP90 in breast-cancer cells, where it might suppress expression of tumor suppressor genes through deacetylation of histones after recruitment of HDAC1. The data reported here suggest that ICBP90 is involved in cell proliferation by way of methylation-mediated regulation of certain genes.

d. cDNA microarray analysis of cancer

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

(1) Chemosensitivity

Gefitinib (Iressa, ZD1839), an inhibitor of epidermal growth factor receptor-tyrosine kinase (EGFR-TK), has shown potent anti-tumor effects and improved symptom and quality-of-life of a subset of patients with advanced NSCLC. However, a large portion of the patients showed no effect to this agent. To establish a method to predict the response of NSCLC patients to gefitinib, we used a genome-wide cDNA microarray to analyze 33 biopsy samples of advanced NSCLC from patients who had been treated with an identical protocol of second-to 7th-line gefitinib monotherapy. We identified 51 genes whose expression differed significantly between 7 responders and 10 non-responders to the drug. We selected the 12 genes that showed the most significant differences to establish a numerical scoring system (GRS, gefitinib response score), for predicting response to gefitinib treatment. The GRS system clearly separated the two groups without any overlap, and accurately predicted responses to $_2$the drug in sixteen additional NSCLC cases. The system was further validated by the semi-quantitative RT-PCR, immunohistochemistry, and ELISA for serological test. Moreover, we proved that the anti-apoptotic activity of amphiregulin (AREG), a protein that was significantly over-expressed in non-responders but undetectable in responders, leads to resistance of NSCLC cells to gefitinib *in vitro*. Our results suggested that sensitivity of a given NSCLC to gefitinib can be predicted according to expression levels of a defined set of genes that may biologically affect drug sensitivity and survival of lung-cancer cells. Our scoring system might eventually lead to achievement of personalized therapy for NSCLC patients.

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. Hence, 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.

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

(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. A gene encoding ADAM8, a disintegrin and metalloproteinase domain-8 protein was selected as a candidate for such molecule. Tumor-tissue microarray was applied to examine expression of ADAM8 protein in archival lung-cancer samples from 363 patients. Serum ADAM 8 levels of 105 lung-cancer patients and 72 controls were also measured by ELISA. A role of ADAM8 in cellular motility was examined by Matrigel assays. ADAM8 was abundantly expressed at both transcriptional and protein levels in the great majority of lung-cancer tissues and cell-lines examined. A high level of ADAM8 protein expression was significantly more common in advanced stage-IIIB/IV adenocarcinomas (ADCs) than in ADCs at stages I-IIIA. Serum levels of ADAM8 protein detected by ELISA were significantly higher in lung-cancer patients than in healthy controls. The proportion of the serum ADAM8-positive cases defined by our criteria was 63% and that for carcinoembryonic antigen (CEA) was 57%, indicating equivalent diagnostic power of these two markers. A combined assay using both ADAM8 and CEA increased sensitivity, as 80% of the patients with lung cancer were then diagnosed as positive while only 11% of 72 healthy volunteers were falsely diagnosed as positive. In addition, exogenous expression of ADAM8 increased the migratory activity of mammalian cells, an indication that ADAM8 may play a significant role in progression of lung cancer.

(3) Testicular cancer

To identify new diagnostic markers for testicular germ cell tumors (TGCTs), including seminomas, as well as potential targets of new drugs for treating the disease, we compared gene-expression profiles of cancer cells from 13 seminomas with normal human testis using laser-capture microdissection and a cDNA microarray representing 23,040 genes. We identified 349 genes that were commonly up-regulated in seminoma cells. The functions of 227 were known to some extent; the remaining 122 included 57 ESTs. On the list were cyclin D2 (CCND2), prostate cancer over-expressed gene 1 (POV1), and junction plakoglobin (JUP), all of which were already known to be over-expressed in seminomas. On the other hand, our protocol selected 593 genes as being commonly down-regulated in seminoma cells. That list included 340 functionally characterized genes; the other 253 included 131 ESTs. To confirm the expression data, we performed semi-quantitative RT-PCR experiments with nine highly up-regulated genes, and the results supported those from of our microarray analysis. The information provided here should prove useful for identifying genes whose products might serve as molecular targets for treatment of TGCTs. We subsequently analyszed NACHT, leucine rich repeat and PYD containing 7 (*NALP7*), that was significantly transactivated in testicular seminomas. Northern blot analyses confirmed an approximately 3.3-kb transcript that was expressed exclusively in testis although the expression level of this gene in normal testis was much lower than in tumor cells, suggesting an important role of this gene in germ-cell proliferation. Immunohistocheminal analysis using anti-NALP7 polyclonal antibody detected the endogenous NALP7 protein in the cytoplasm of embryonal carcinoma cells and testicular seminoma tissues. Transfection of small interfering RNA (siRNA) for *NALP7* significantly reduced the *NALP7* expression and resulted in growth suppression of testicular germ-cell tumor. These findings imply that *NALP7* may play crucial roles in cell proliferation as well as testicular tumorigenesis, and represents a promising candidate for development of targeted therapy for TGCTs.

(4) Pancreatic cancer

To characterize molecular mechanism involved in pancreatic carcinogenesis and apply the information toward development of novel tumor markers and therapeutic targets, we analyzed gene-expression profiles of 18 pancreatic tumors using a cDNA microarray representing 23,040 genes. As pancreatic ductal adenocarcinomas usually contain a low proportion of cancer cells in the tumor mass, we prepared 95% pure populations of pancreatic-cancer cells by means of laser microbeam microdissection (LMM), and compared their expression profiles to those of similarly purified, normal pancreatic ductal cells. Because of the high degree of purity in the cell populations, a large proportion of genes that we detected as up-regulated or down-regulated in pancreatic cancers were different from those reported in previous studies. We identified 260 genes that were commonly up-regulated in pancreatic cancer cells; the functions of 66 of them (including 31 ESTs) are currently unknown. The up-regulated genes included some that were previously reported to be over-expressed in pancreatic cancers, such as interferon-induced transmembrane protein 1 (IFITM1), plasminogen activator, urokinase (PLAU), prostate stem cell antigen (PSCA), S100 calcium binding protein P (S 100P), and baculoviral IAP repeat-containing 5 (BIRC5). On the other hand, 346 genes were commonly down-regulated in the cancer cells. Of those, 211 had been functionally characterized; this group included tumor suppressor genes such as AXIN1 up-regulated 1 (AXUD1), deleted in liver cancer 1 (DLC1), growth arrest and DNA-damage-inducible beta (GADD45B), and p53-inducible p53DINP1 (p53DINP1). These data should provide useful information for finding candidate genes whose products might serve as specific tumor markers and/or as molecular targets for treatment of patients with pancreatic cancer.

Through functional analysis of genes that were transactivated in 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*, *DLG5* 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 subcellular 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.

(5) Prostate cancer

To characterize the molecular mechanisms involved in prostate carcinogenesis and to evaluate a controversial hypothesis about the putative transition from prostatic intraepithelial neoplasia (PIN) to invasive prostate cancer (PC), we analyzed gene-expression profiles of 20 PCs and 10 high-grade PINs, using a cDNA microarray representing 23,040 genes. Considering the histological heterogeneity of PCs and the minimal nature of PIN lesions, we applied laser microbeam microdissection (LMM) to purify populations of PC and PIN cells, and then compared their expression profiles with those of corresponding normal prostatic epithelial cells that were also purified by LMM. A hierarchical clustering analysis clearly separated the PC group from the PIN group except for three tumors that were morphologically defined as one very-high-grade PIN and two low-grade PCs. The findings suggested that PINs and PCs share some molecular features, and supported the hypothesis of PIN-to-PC transition. Based on this hypothesis, we identified 21 genes that were up-regulated and 63 that were down-regulated commonly in PINs and PCs comparing with normal epithelium, which were considered to be involved in the presumably early stage of prostatic carcinogenesis. The altered genes included *AMACR, OR51E2, RODH* and *SMS*. Furthermore, comparing the expression profiles of PCs with those of PINs, we identified 41 genes that were up-regulated and 98 that were down-regulated in the transition from PINs to PCs; those altered genes included elements that are likely to be involved in cell adhesion or motility of invasive PC cells, such as *POV1, CDKN2C, FASN, ITGB2, LAMB2, PLAU*, and *TIMP1*. These data provide clues to the molecular mechanisms underlying prostatic carcinogenesis, and suggest candidate genes whose products might serve as molecular targets for prevention and treatment of prostate cancers.

(6) Breast cancer

Breast carcinoma is a complex disease characterized by accumulation of multiple genetic alterations, and investigators are far from having a full understanding of the molecular basis of mammary tumorigenesis. In this study we analyzed gene-expression profiles of 81 surgical specimens of 12 ductal carcinomas in situ (DCIS) and 69 invasive ductal carcinomas (IDC). After applying laser-microbeam microdissection to all samples we achieved 98-99% pure populations of breast-cancer cells, and of normal breast epithelial cells used as controls. A cDNA-microarray analysis of 23,040 genes in these samples and a subsequent unsupervised hierarchical clustering distinguished two tumor groups, mainly in terms of estrogen-receptor (ER) status. We then undertook a supervised analysis and identified 325 genes that were commonly either up-or down-regulated in both pathologically discrete stages (DCIS and IDC), indicating that these genes might play important roles in malignant transformation of breast ductal cells. In addition, we searched invasion-associated gene candidates whose expression was altered in IDC, but not in DCIS, and identified 24 up-regulated genes and 41 down-regulated genes. Furthermore, we identified 34 genes that were expressed differently in tumors from patients with lymph-node metastasis as opposed to no metastasis. On that basis we developed a scoring system that correlated well with the metastatic status. Tumors from all of the 37 test patients with lymph-node metastasis yielded positive scores by our definition, whereas 38 of the 40 tumors (95%) without lymph-node metastasis had negative scores. Our data should provide useful information for identifying predictive markers for invasion or metastasis, and suggest potential target molecules for treatment of breast cancers.

2. Genes responsible for other diseases

a. Bone development

Mitsuhito Doi and Yusuke $_3$Nakamura

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

b. 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: Department of Medicine, Kidney Center, Tokyo Women's Medical University, Tokyo, Japan; Laboratory for Genotyping, Laboratory for Cardiovascular Diseases, Laboratory for Rheumatic Diseases, Laboratory for Diabetic Nephropathy, and Laboratory for Medical Informatics, SNP Research Center, The Institute of Physical and Chemical Research (RIKEN), Tokyo, Japan; Department of Urology, Iwate Medical University, Iwate, Japan; Division of Clinical Nephrology and Rheumatology, Niigata University Graduate School of Medical and Dental Sciences, Niigata, Japan; Department of Urology, Iwate Prefectural Ofunato Hospital, Iwate, Japan; 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 m-binding protein 2 (IGHMBP2) at chromosome 11 q13.2-q13.4. The association ($c^2 = 17.1$, $p = 0.00003$; 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 (G34448 A) 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

c. Crohn 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: Department of Medicine, Division of Gastroenterology, Social Insurance Chuo General Hospital, Tokyo, Japan, SNP Research Center, the Institute of Physical and Chemical Research (RIKEN), Kanagawa, Japan

Crohn disease (CD) is an inflammatory bowel disease as characterized by chronic transmural, segmental and typically granulomatous inflammation of the gut. Recently, two novel candidate gene loci associated with CD, *SLC22A4* and *SLC 22A5* on chromosome 5 known as *IBD5*, and *DLG5* on chromosome 10, were identified through association analysis of Caucasian CD patients. We validated these candidate genes in Japanese patients with CD and found a weak but possible association with both *SLC22A4* ($p = 0.028$) and *DLG5* ($p = 0.023$), although the reported genetic variants that were indicated to be causative in Caucasian population were completely absent in or were not associated with Japanese CD patients. These findings imply significant differences in genetic background with CD susceptibility among different ethnic groups and further indicate some difficulty of population-based studies.

Publications

Okabe, H., Furukawa, Y., Kato, T., Hasegawa, S., Yamaoka, Y. and Nakamura, Y. Isolation of development and differentiation enhancing factor-like 1 (DDEFL1) as a drug target for hepatocellular carcinomas. Int. J. Oncology. 24: 43-48, 2004.

Anazawa, Y., Arakawa, H., Nakagawa, H. and Nakamura, Y. Identification of STAG1 as a key mediator of a p53-dependent apoptotic pathway. Oncogene 23: 7621-7627, 2004.

Yoshida, K., Monden, M., Nakamura, Y. and Arakawa, H. Adenovirus-mediated p53AIP1 gene transfer as a new strategy for treatment of p53-resistant tumors. Cancer Sci. 95: 91-97, 2004.

Li, M., Lin, Y.-M., Hasegawa, S., Shimokawa, T., Murata, K., Kameyama, M., Ishikawa, O., Katagiri, T., Tsunoda, T., Nakamura, Y. and

Furukawa, Y. Genes associated with liver metastasis of colon cancer, identified by genome-wide cDNA microarray. Int. J. Oncology. 24: 305-312, 2004.

Nagayama, S., Iiizumi, M., Katagiri, T., Toguchida, J. and Nakamura, Y. Identification of PDZK4, a novel human gene with PDZ domains, that is up-regulated in synovial sarcomas. Oncogene 23: 5551-5557, 2004.

Ochi, K., Daigo, Y., Katagiri, T., Nagayama, S., Tsunoda, T., Myoui, A., Naka, N., Araki, N., Kudawara, I., Ieguchi, M., Toyama, Y., Toguchida, J., Yoshikawa, H. and Nakamura, Y. Prediction of response to neoadjuvant chemotherapy for osteosarcoma by gene-expression profiles. Int. J. Oncol. 24: 647-655, 2004.

Nakamura, Y. Isolation of p53-target genes and their functional analysis (Review). Cancer Sci. 95: 7-11, 2004.

Sekiya, T., Adachi, S., Kohu, K., Yamada, T., Higuchi, O., Furukawa, Y., Nakamura, Y., Nakamura, T., Tashiro, K., Kuhara, S., Ohwada, S. and Akiyama, T. Identification of BMP and activin membrane-bound inhibitor (BAMBI), an inhibitor of TGF-β signaling, as a target of the β-catenin pathway in colorectal tumor cells. J. Biol. Chem. 279: 6840-6846, 2004.

Doi, M., Nagano, A. and Nakamura, Y. Molecular cloning and characterization of a novel gene, EMILIN-5, and its possible involvement in skeletal development. Biochem. Biophy. Res. Commun. 313: 888-893, 2004.

Yoon, K., Nakamura, Y. and Arakawa, H. Identification of ALDH4 as a p53-inducible gene and its protective role in cellular stresses. J. Hum. Genet. 49: 134-140, 2004.

Nakamura, T., Furukawa, Y., Tsunoda, T., Ohigashi, H., Murata, K., Ishikawa, O., Ohgaki, K., Kashimura, N., Miyamoto, M., Hirano, S., Kondo, S., Katoh, H., Nakamura, Y. and Katagiri, T. Genome-wide cDNA microarray analysis of gene-expression profiles in pancreatic cancers using populations of tumor cells and normal ductal epithelial cells selected for purity by laser microdissection. Oncogene 23: 2385-2400, 2004.

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

Iida, A., Saito, S., Sekine, A., Kataoka, Y., Tabei, W. and Nakamura, Y. Catalog of 300 SNPs in 23 genes encoding G-protein coupled receptors. J. Hum. Genet. 49: 194-208, 2004.

Kochi, Y., Yamada, R., Kobayashi, K., Takahashi, A., Suzuki, A., Sekine, A., Mabuchi, A., Akiyama, F., Tsunoda, T., Nakamura, Y. and Yamamoto, K. Analysis of single-nucleotide polymorphisms in Japanese rheumatoid arthritis patients shows additional susceptibility markers besides the classic shared epitope susceptibility sequences. Arthritis & Rheumatism 50: 63-71, 2004.

Harima, Y., Togashi, A., Horikoshi, K., Imamura, M., Sougawa, M., Sawada, S., Tsunoda, T., Nakamura, Y. and Katagiri, T. Prediction of outcome of advanced cervical cancers to thermoradiotherapy according to expression profiles of 35 genes selected by cDNA microarray analysis. Int. J. Radiat. Oncol. Biol. Phys. 60: 237-248, 2004.

Yuki, D., Lin, Y.-M., Fujii, Y., Nakamura, Y. and

Furukawa, Y. Isolation of LEM domain-containing 1, a novel testis-specific gene expressed in colorectal cancers. Oncol. Rep. 12: 275-280, 2004.

Jinawath, N., Furukawa, Y. and Nakamura, Y. Identification of NOL8, a nucleolar protein containing an RNA recognition motif (RRM), which was overexpressed in diffuse-type gastric cancer. Cancer Science 95: 430-435, 2004.

Iida, A. and Nakamura, Y. Identification of 45 novel SNPs in the 83-kb region containing peptidylarginine deiminase types 1 and 3 loci on chromosomal band 1p36. 13. J. Hum. Genet. 49: 387-390, 2004.

Ozaki, K., Inoue, K., Sato, H., Iida, A., Ohnishi, Y., Sekine, A., Sato, H., Odashiro, K., Nobuyoshi, M., Hori, M., Nakamura, Y. and Tanaka, T. Functional variation in LGALS2 confers risk of myocardial infarction and regulates lymphotoxin-alpha secretion in vitro. Nature 429: 72-75, 2004.

Kamatani, N., Sekine, A., Kitamoto, T., Iida, A., Saito, S., Kogame, A., Inoue, E., Kawamoto, M., Harigai, M. and Nakamura, Y. Large-scale single-nucleotide polymorphism (SNP) and haplotype analyses, using dense SNP maps, of 199 drug-related genes in 752 subjects: The analysis of the association between uncommon SNPs within haplotype blocks and the haplotypes constructed with haplotype-tagging SNPs. Am. J. Hum. Genet. 75: 190-203, 2004.

Iida, A., Saito, S., Sekine, A., Tabei, W., Kataoka, Y. and Nakamura, Y. Identification of 20 novel SNPs in the guanine nucleotide binding protein alpha 12 gene locus. J. Hum. Genet. 49: 445-448, 2004.

Jinawath, N., Furukawa, Y., Hasegawa, S., Li, M., Tsunoda, T., Satoh, S., Yamaguchi, T., Imamura, H., Inoue, M., Shiozaki, H. and Nakamura, Y. Genome-wide analysis of gene expression in diffuse-type 4 gastric cancers using cDNA microarray; comparative expression profiles between diffuse-type and intestinal-type gastric cancer. Oncogene 23: 6830-6844, 2004.

Tsunoda, T., Lathrop, G.M., Sekine, A., Yamada, R., Takahashi, A., Ohnishi, Y., Tanaka, T. and Nakamura, Y. Variation of gene-based SNPs and linkage disequilibrium patterns in the human genome. Hum. Mol. Genet. 13: 1623-1632, 2004.

Hamamoto, R., Furukawa, Y., Morita, M., Iimura, Y., Silva, F.P., Li, M., Yagyu, R. and Nakamura, Y. SMYD3 encodes a novel histone methyltransferase involved in the proliferation of cancer cells. Nat. Cell Biol. 6: 731-740, 2004.

Ashida, S., Nakagawa, H., Katagiri, T., Furihata, M., Tsunoda, T., Takata, R., Kasahara, K., Miki, T., Fujioka, T., Shuin, T. and Nakamura, Y. Molecular features of the transition from prostatic intraepithelial neoplasia (PIN) to prostate cancer : Genome-wide gene-expression profiles of prostate cancers and PINs. Cancer Res. 64: 5963-5972, 2004.

Nishidate, T., Katagiri, T., Lin, M.-L., Mano, Y., Miki, Y., Kasumi, F., Yoshimoto, M., Tsunoda, T., Hirata, K. and Nakamura, Y. Genome-wide gene-expression profiles of breast-cancer cells purified with laser microbeam microdissection: Identification of genes associated with progression and metastasis. Int. J. Oncol. 25: 797-819, 2004.

Mizuguchi, T., Collod-Beroud, G., Akiyama, T., Abifadel, M., Harada, N., Morisaki, T., Allard, D., Varret, M., Claustres, M., Morisaki, H., Ihara, M., Kinoshita, A., Yoshiura, K., Junien, C., Kajii, T., Jondeau, G., Ohta, T., Kishino, T., Furukawa, Y., Nakamura, Y., Niikawa, N., Boileau, C. and Matsumoto, N. Heterozygous TGFBR2 mutations in Marfan syndrome. Nat. Genet. 36: 855-860, 2004.

Unoki, M., Nishidate, T. and Nakamura, Y. ICBP90, an E2F-1 target, recruits HDAC1 and binds to methyl-CpG through its SRA domain. Oncogene 23: 855-860, 2004.

Cha, P.-C., Yamada, R., Sekine, A., Nakamura, Y. and Koha, C.-L. Inference from the relationships between linkage disequilibrium and allele frequency distributions of 240 candidate SNPs in 109 genes of drug-related genes importance in 4 Asian populations. J. Hum. Genet. 49: 558-572, 2004.

Takahashi, M., Lin, Y.-M., Nakamura, Y. and Furukawa, Y. Isolation and characterization of a novel gene CLUAP1 whose expression is frequently upregulated in colon cancer. Oncogene 23: 9289-9294, 2004.

Yoshitake, Y., Nakatsura, T., Monji, M., Senju, S., Matsuyoshi, H., Tsukamoto, H., Hosaka, S., Komori, H., Fukuma, D., Ikuta, Y., Katagiri, T., Furukawa, Y., Ito, H., Shinohara, M., Nakamura, Y. and Nishimura, Y. Proliferation potential-related protein, an ideal esophageal cancer antigen for immnotherapy, identified using cDNA microarray analysis. Clin. Cancer Res. 10: 6437-6448, 2004.

Yagyu, R., Furukawa, Y., Lin, Y.-M., Shimokawa, T., Yamamura, T. and Nakamura, Y. A novel oncoprotein RNF43 functions as an autocrine manner in colorectal cancer. Int. J. Oncol. 25: 1343-1348, 2004.

Yang, L., Leung, A.C.C., Ko, J.M.Y., Lo, P.H.Y., Tang, J.C.O., Srivastava, G., Oshimura, M., Stanbridge, E.J., Daigo, Y., Nakamura, Y., Tang, C.M.C., Lau, K.W., Law, S. and Lung, M. L. Tumor suppressive role of a 2.4 Mb 9q

33-q34 critical region and DEC1 in esophageal squamous cell carcinoma. Oncogene, Dec 6, 2004 [Epub ahead of print].

Ishikawa, N., Daigo, Y., Yasui, W., Inai, K., Nishimura, H., Tsuchiya, E., Kohno, N. and Nakamura, Y. ADAM8 as a novel serological and histochemical marker for lung cancer. Clin. Cancer Res. 10: 8363-8370, 2004.

Okada, K., Hirota, E., Mizutani, Y., Fujioka, T., Shuin, T., Miki, T., Nakamura, Y. and Katagiri, T. Oncogenic role of NALP7 in testicular seminomas. Cancer Sci. 95: 949-954, 2004.

Yamazaki, K., Takazoe, M., Tanaka, T., Ichimori, T., Saito, S., Iida, A., Onouchi, Y., Hata, A. and Nakamura, Y. Association analysis of SLC 22A4, SLC22A5 and DLG5 in Japanese patients with Crohn disease. J. Hum. Genet. 49: 664-668, 2004.

Minaguchi, T., Yoshikawa, H., Nakagawa, S., Yasugi, T., Yano, T., Iwase, H., Mizutani, K., Shiromizu, K., Ohmi, K., Watanabe, Y., Noda, K., Nishiu, M., Nakamura, Y. and Taketani, Y. Association of PTEN mutation with HPV-negative adenocarcinoma of the uterine cervix. Canacer Letters 210: 57-62, 2004.

Nakatsuda, T., Komori, H., Kudo, T., Yoshitake, Y., Senju, S., Katagiri, T., Furukawa, Y., Ogawa, M., Nakamura, Y. and Nishimura, Y. Mouse homologue of a novel human oncofetal antigen, Glypican-3, evokes T cell-mediated tumor rejection without autoimmune reactions in mice. Clin. Cancer Res. 10: 8630-8640, 2004.

Uchida, N., Tsunoda, T., Wada, S., Furukawa, Y., Nakamura, Y. and Tahara, H. Ring finger protein (RNF) 43 as a new target for cancer immunotherapy. Clin. Cancer Res. 10: 8577-8586, 2004.

Kakiuchi, S., Daigo, Y., Ishikawa, N., Furukawa, C., Tsunoda, T., Yano, S., Nakagawa, K., Tsuruo, T., Kohno, N., Fukuoka, M., Sone, S. and Nakamura, Y. Prediction of sensitivity of advanced non-small cell lung cancers to gefitinib. Hum. Mol. Genet. 13: 3029-3043, 2004.

Kamiyama, H., Kurosaki, K., Kurimoto, M., Katagiri, T., Nakamura, Y., Kurokawa, M., Sato, H., Endo, S. and Shiraki, K. Herpes simplex virus induced death receptor-dependent apoptosis and regression of transplanted human cancers. Cancer Sci. 95: 990-998, 2004.

Onouchi, Y., Onoue, S., Tamari, M., Wakui, K., Fukushima, Y., Yashiro, M., Nakamura, Y., Yanagawa, H., Kishi, F., Ouchi, K., Terai, M., Hamamoto, K., Kudo, F., Aotsuka, H., Sato, Y., Nariai, A., Kaburagi, Y., Miura, M., Saji, T., Kawasaki, T., Nakamura, Y. and Hata, A. CD 40 ligand gene and Kawasaki disease. Eur. J. Hum. Genet. 12: 1062-1068, 2004.

Hirota, T., Obara, K., Matsuda, A., Akahoshi, M., Nakashima, K., Hasegawa, K., Takahashi, N., Shimizu, M., Sekiguchi, H., Kokubo, M., Doi, S., Fujiwara, H., Miyatake, A., Fujita, K., Enomoto, T., Kishi, F., Suzuki, Y., Saito, H., Nakamura, Y., Shirakawa, T. and Tamari, M. Association between genetic variation in the gene for death-associated protein-3 (DAP3) and adult asthma. J. Hum. Genet. 49: 370-375, 2004.

Kanazawa, A., Tsukada, S., Sekine, A., Tsunoda, T., Takahashi, A., Kashiwagi, A., Tanaka, Y., Babazono, T., Matsuda, M., Kaku, K., Iwamoto, Y., Kawamori, R., Kikkawa, R., Nakamura, Y. and Maeda, S. Association of the gene encoding wingless-type mammary tumor virus integration-site family member 5B (WNT5B) with type 2 diabetes. Am. J. Hum. Genet. 75: 832-843, 2004.

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. and Nakamura, Y. Association of a single-nucleotide polymorphism in the immunoglobulin u-binding protein 2 gene with immunoglobulin A nephropathy. J. Hum. Genet., Dec. 14, 2004 [Epub ahead of print].

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

Iida, A., Ozaki, K., Tanaka, T. and Nakamura, Y. Fine-scale SNP map of an 11-kb genomic region at 22q13.1 containing the galectin-1 gene. J. Hum. Genet., in press.

Park, W.-R. and Nakamura, Y. p53CSV, a novel p53-inducible gene involved in the p53-dependent cell-survival pathway. Cancer Res., in press.

Taniuchi, K., Nakagawa, H., Nakamura, T., Eguchi, H., Ohigashi, H., Ishikawa, O., Katagiri, T. and Nakamura, Y. Over-expressed P-cadherin/CDH3 promotes motility of pancreatic cancer cells by interacting with p120ctn and activating rho-family GTPases. Cancer Res., in press.

Takata, R., Katagiri, T., Kanehira, M. Tsunoda, T., Shuin, T., Miki, T., Namiki, M., Kohri, K., Matsushita, Y., Fujioka, T. and Nakamura, Y. Predicting response to M-VAC neoadjuvant chemotherapy for bladder cancers through genome-wide gene expression profiling. Clin. Cancer Res., in press.

Nishiu, M., Tomita, Y., Nakatsuka, S., Takakuwa, T., Iizuka, N., Hoshida, Y., Ikeda, J., Iuchi, K., Yanagawa, R., Nakamura, Y. and

Aozasa, K. Distinct pattern of gene expression in pyothorax-associated lymphoma (PAL), a lymphoma developing inlong-standing inflammation. Cancer Sci. 95: 828-834, 2004.

Kizawa, H., Kou, I., Iida, A., Sudo, A., Miyamoto, Y., Fukuda, A., Mabuchi, A., Kotani, A., Kawakami, A., Yamamoto, S., Uchida, N., Nakamura, K., Notoya, K., Nakamura, Y. and Ikegawa, S. An aspartic acid repeat polymorphism in asporin negatively affects chondrogenesis and increases susceptibility to osteoarthritis. Nat. Genet., in press.

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	博士(理学)	中　井　謙　太
Associate Professor	Kengo Kinoshita, Ph.D.	助教授	博士(理学)	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 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 regulatory information, which will lead us to the understanding of its cellular role represented by the networks of inter-gene interaction.

1. Regulon prediction based on sigma factor prediction and DBTBS (Database of Transcriptional Regulation in *Bacillus subtilis*)

Yuko Makita, Michiel J.L. de Hoon[1], Naotake Ogasawara[2], Satoru Miyano[1], and Kenta Nakai: [1]Laboratory of DNA Information Analysis; [2]Nara Institute of Science and Technology

Sigma factors, often in conjunction with other transcription factors, regulate gene expression in prokaryotes at the transcriptional level. Specific transcription factors tend to co-occur with specific sigma factors. To predict new members of the transcription factor regulons, we applied Bayes' rule to combine the Bayesian probability of sigma factor prediction calculated from microarray data and the sigma factor binding sequence motif, the motif score of the transcription factor associated with the sigma factor, the empirically determined distance between the transcription start site to the *cis*-regulatory region, and the tendency for specific sigma factors and transcription factors to co-occur. By combining these information sources, we improve the accuracy of predicting regulation by transcription factors, and also confirm the sigma factor prediction. We applied our proposed method to all genes in *Bacillus subtilis* to find currently unknown transcriptional regulations by transcription factors and sigma factors.

2. Comprehensive analysis of alternative promoters

Tsuritani Katuski[3], Riu Yamashita, Yutaka Suzuki[4], Sumio Sugano[4] and Kenta Nakai: [3]Taisho Pharmaceutical Co. ltd.; [4]Grad. School New Frontier Sci.

It is getting clearer that a single gene produces multiple transcripts in many ways. In mammals, an example of such mechanisms is the use of alternative promoters, whose involvement is implicated in the spatial/temporal regulation of their downstream gene expression as well as the isoformal production of their products. To further understand this phenomenon, we comprehensively analyzed a large amount of human/mouse 5' ESTs that have intact transcriptional start sites (TSSs), obtained with the oligo-capping or related methods. Namely, we mapped the ESTs on the genome and classified them into clusters that are separated each other for more than 500 bp on it. If we assume that these clusters are produced from alternative promoters, 7,635 of 14,312 human genes seem to have alternative TSSs, which means that about 50%of mammalian genes may be regulated by alternative promoters. We further created a more convincing 'alternative promoter core dataset' (AP core dataset) from the alternative TSS dataset, based on the conservation of their upstream regions between human and mouse. The AP core dataset contains 523 genes, which include genes, such as *shc* and AKAP1, that are known to have alternative promoters. Moreover, we observed that some genes have tissue-specific TSS clusters. By analyzing their Gene Ontology annotation, we also found that genes related to 'signal transduction' are significantly enriched in the AP core dataset. Our results give a firm foothold for the clarification of the alternative transcription mechanism.

3. Comprehensive analysis of CpG islands in promoter regions

Riu Yamashita, Yutaka Suzuki[4], Sumio Sugano[4] and Kenta Nakai

It has been envisaged that CpG islands are often observed near the transcriptional start sites (TSSs) of house-keeping genes. However, neither the precise positions of CpG islands relative to TSSs of genes nor the correlation between the presence of the CpG islands and the expression specificity of these genes are well-understood. Using thousands of sequences with known TSSs in human and mouse, we found that there is a clear peak in the distribution of CpG islands around TSSs in the genes of these two species. Thus, we classified human (mouse) genes into 6,600 (2,948) CpG+genes and 2,619 (1,830) CpG-ones, based on the presence of a CpG island within the −100: +100 region. We estimated the degree of each gene being a housekeeper by the number of cDNA libraries where its ESTs were detected. Then, the tendency that a gene lacking CpG islands around its TSS is expressed with a higher degree of tissue specificity turned out to be evolutionarily conserved. We also confirmed this tendency by analyzing the gene ontology annotation of classified genes. Since no such clear correlation was found in the control data (mRNAs, pre-mRNAs, and chromosome banding pattern), we concluded that the effect of a CpG island near the TSS should be more important than the global GC content of the region where the gene resides.

4. Global detection of human genes that have the terminal oligo-pyrimidine (TOP) motif

Riu Yamashita, Yutaka Suzuki[4], Sumio Sugano[4] and Kenta Nakai

It is known that the expression of many genes coding ribosomal proteins and translation factors is regulated at the time of translation. Those genes have terminal oligo-pyrimidine sequence at the 5'-end of their mRNA and are called TOP genes. However, how many TOP genes exist in human/mouse genomes is still unknown. By using the accurate information of TSSs (transcriptional start sites), higher accuracy to detect them is expected. As a first step, we focused on genes that have fixed TSS positions, which will also reduce the number of potential false positives. By using a weight matrix constructed from known TOP genes, we could screen 511 candidate TOP genes. In them, all of 9 previously characterized TOP genes were included. Moreover, 78 of 80 ribosomal protein genes were also included. There were 99 human TOP gene candidates that have also mouse orthologs. Our result suggests that TOP genes are not only translation-related but may include a wider variety of genes such as chaperones and transport proteins.

5. *Cis* element analysis using DNA array data in *Arabidopsis*

Takeshi Obayashi[5], Kenta Nakai, and Hiroyuki Ohta[5]: [5]Tokyo Institute of Technology

Although *cis*-element prediction methods from gene expression data are generally used, there are few reports in plant science for the comprehensive analysis of regulatory mechanisms of gene expression using bioinformatics approaches. Thus, we comprehensively predicted *cis* elements from gene expression data obtained by our microarray analyses of *Arabidopsis thaliana* genes. Based on the results of *cis* element prediction, we estimated the regulatory mechanisms of gene expression. The analysis was performed as follows: First, many candidates of *cis* elements were selected by several methods. After the evaluation of their effects on gene expression, a set of non-redundant *cis* elements was created. Finally, the regulatory mechanisms of gene expression were inferred from these results.

6. Large-scale analysis of human alternative protein isoforms: pattern classification and correlation with subcellular localization signals

Mitsuteru Nakao[6], Roberto A. Barrero[7], Paul Horton[6], and Kenta Nakai: [6]CBRC, Nat. Inst.

Advanced Industrial Sci. Tech.; [7]Nat. Inst. Genet.

We investigated human alternative protein isoforms of more than 2,600 genes based on full-length cDNA clones and SwissProt. We classified the isoforms and examined their co-occurrence for each gene. Further, we investigated potential relationships between these changes and differential subcellular localization. The two most abundant patterns were the one with different C-terminal regions and the one with an internal insertion, which together account for 43% of the total. Although changes of the N-terminal region are less common than those of the C-terminal region, extension of the C-terminal region is much less common than that of the N-terminal region, probably because of the difficulty of removing stop codons in one isoform. We also found a tendency for the members of a gene family to show the same combination of co-occurrence in alternative isoforms. We interpret this as evidence that there is some structural relationship which produces a repertoire of isoformal patterns. Finally, many terminal changes are predicted to cause differential subcellular localization, especially in targeting to either peroxisomes or mitochondria. Our study sheds new light on the enrichment of the human proteome through alternative splicing and related events. Our database of alternative protein isoforms is available through the Internet.

7. Identification of the ligand binding sites on the molecular surface of proteins.

Kengo Kinoshita and Haruki Nakamura[8]: [8]Institute for Protein Research, Osaka University

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. Previously we reported our first trial on structure based function prediction, based the similarity searches of molecular surfaces against the functional site database. Here we describe the extension of our first trial by expanding the search database to whole hetero atom binding sites appearing within the PDB with the new analysis protocol. In addition, we have determined the similarity threshold line, by using 10 structure pairs with solved free and complex structures. Finally, we extensively applied our method to newly determined hypothetical proteins including some without annotations, and evaluated the performance of our methods.

8. Prediction of DNA binding sites on molecular surface of proteins

Yuko Tsuchiya[8], Kengo Kinoshita, Haruki Nakamura[8]

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, local curvature and global curvature on the surfaces. Results of the prediction can be interactively checked with our original surface viewer.

Publications

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

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

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

Imanishi T, Itoh T, Suzuki Y, O’Donovan C, Fukuchi S, Koyanagi KO, Barrero RA, Tamura T, Yamaguchi-Kabata Y, Tanino M, Yura K, Miyazaki S, Ikeo K, Homma K, Kasprzyk A, Nishikawa T, Hirakawa M, Thierry-Mieg J, Thierry-Mieg D, Ashurst J, Jia L, Nakao M, Thomas MA, Mulder N, Karavidopoulou Y, Jin L, Kim S, Yasuda T, Lenhard B, Eveno E, Suzuki Y, Yamasaki C, Takeda JI, Gough C, Hilton P, Fujii Y, Sakai H, Tanaka S, Amid C, Bellgard M, Bonaldo Md M, Bono H, Bromberg SK, Brookes AJ, Bruford E, Carninci P, Chelala C, Couillault C, Souza SJ, Debily MA, Devignes MD, Dubchak I, Endo T, Estreicher A, Eyras E, Fukami-Kobayashi K, R Gopinath G, Graudens E, Hahn Y, Han M, Han ZG, Hanada K, Hanaoka H, Harada E, Hashimoto K, Hinz U, Hirai M, Hishiki T, Hopkinson I, Imbeaud S, Inoko H, Kanapin A, Kaneko Y, Kasukawa T, Kelso J, Kersey P, Kikuno R, Kimura K, Korn B, Kuryshev V, Makalowska I, Makino T, Mano S, Mariage-Samson R, Mashima J, Matsuda H, Mewes HW, Minoshima S, Nagai K, Nagasaki H, Nagata N, Nigam R, Ogasawara O, Ohara O, Ohtsubo M, Okada N, Okido T, Oota S, Ota M, Ota T, Otsuki T, Piatier-Tonneau D, Poustka A, Ren SX, Saitou N, Sakai K, Sakamoto S, Sakate R, Schupp I, Servant F, Sherry S, Shiba R, Shimizu N, Shimoyama M, Simpson AJ, Soares B, Steward C, Suwa M, Suzuki M, Takahashi A, Tamiya G, Tanaka H, Taylor T, Terwilliger JD, Unneberg P, Veeramachaneni V, Watanabe S, Wilming L, Yasuda N, Yoo HS, Stodolsky M, Makalowski W, Go M, Nakai K, Takagi T, Kanehisa M, Sakaki Y, Quackenbush J, Okazaki Y, Hayashizaki Y, Hide W, Chakraborty R, Nishikawa K, Sugawara H, Tateno Y, Chen Z, Oishi M, Tonellato P, Apweiler R, Okubo K, Wagner L, Wiemann S, Strausberg RL, Isogai T, Auffray C, Nomura N, Gojobori T, and Sugano S., Integrative annotation of 21,037 human genes validated by full-length cDNA clones, PLoS Biol., 2 (6): E162-, 2004.

Horton, P., Mukai, Y., and Nakai, K., Chapter 14: Protein subcellular localization prediction, in L. Wong (ed.), Practical Bioinformatician, World Scientific Publishing Co., pp. 193-216, 2004.

De Hoon, M.J.L., Makita, Y., Imoto, S., Kobayashi, K., Ogasawara, N., Nakai, K., and Miyano, S., Predicting gene regulation by sigma factors in *Bacilllus subtilis* from genome-wide data, Bioinformatics, 20 (Supp. 1): I101-I108, 2004.

Suzuki, Y., Yamashita, R., Shirota, M., Sakakibara, Y., Chiba, J., Mizushima-Sugano, J., Kel, A. E., Arakawa, T., Caminci, P., Kawai, J., Hayashizaki, Y., Takagi, T., Nakai, K., and Sugano, S., Large-scale collection and characterization of promoters of human and mouse genes, In silico Biol., 4: 0036-, 2004.

Suzuki, Y., Yamashita, R., Shirota, M., Sakakibara, Y., Chiba, J., Mizushima-Sugano, J., Nakai, K., and Sugano, S., Sequence comparison of human and mouse genes reveals a homolgous block structure in the promoter regions, Genome Res., 14: 1711-1718, 2004.

Poluliakh, N., Konno, M., Horton, P., and Nakai, K., Parameter landscape analysis for common motif discovery programs, Proc. 1st Ann. RECOMB Satellite Workshop on Regulatory Genomics, in press.

Bannai, H., Hyyro, H., Shinohara, A., Takeda, M., Nakai, K., and Miyano, S., Finding optimal pairs of patterns, Lecture Notes in Comp. Sci. (WABI 2004), 3240: 450-, 2004.

Inenaga, S., Bannai, H., Hyyro, H., Shinohara, A., Takeda, M., Nakai, K., and Miyano, S., Finding optimal pairs of cooperative and competing patterns with bounded distance, Lecture Notes in Comp. Sci. (DS 2004), 3245: 32-, 2004.

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.

Bannai, H., Hyyro, H., Shinohara, A., Takeda, M., Nakai, K., and Miyano, S., An O(N^2) algorithm for discovering optimal boolean pattern pairs, TCBBSI, in press.

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, in press.

Kinoshita, K., and Nakamura, H., Identification of the ligand binding sites on molecular surface of proteins, Protein Science, in press.

Tsuchiya, Y., Kinoshita, K., and Nakamura, H., PreDs: A server for predicting dsDNA-binding site on protein molecular surface, Bioinformatics, in press.

山下理宇・中井謙太，第３編第１節　ゲノムのデータベース，今中忠行監修　ゲノミクス・プロテオミクスの新展開：生物情報の解析と応用，エヌ・ティー・エス：786–791，2004.

中井謙太，ゲノムデータベース入門：つくる人と使う人のために，高木利久編　東京大学生物情報科学学部教育特別プログラム：バイオインフォマティクス概論，羊土社：100–110，2004.

鈴木穣，山下理宇，中井謙太，菅野純夫，トランスクリプトーム解析とトランスクリプトームデータベース，小原・谷口・市川・猪飼編，バイオ高性能機器・新技術利用マニュアル　蛋白質核酸酵素８月号増刊49⑾：1859–1865，2004.

中井謙太，遺伝子の機能予測概論，松原謙一監修，情報生物学講義Ｉ—配列情報の科学財団法人国際高等研究所，55頁，2004.