A System for Describing Cis-Regulatory Machinery Unit

Tsuguchika Kaminuma[1] Takako Takai-Igarashi[2] Masumi Yukawa[3] Yoshitomo Tanaka[3] Hiroshi Tanaka[3]
[1]*Center for Quantum Life Sciences, Graduate School of Science, Hiroshima University, Japan*
[2]*Graduate School of Information Science and Technology, University of Tokyo, Japan*
[3]*Tokyo Medical and Dental University, Japan*
Email : kaminuma@cbi.or.jp

ABSTRACT: Studies on cellular pathways and networks are now one of the most actively researched topics in all fields of biomedicine ranging from developmental biology to etiology. Many databases have been developed and quantitative simulation models have been proposed. One of the eventual goals of pathway/network studies is to integrate different types of pathway/network models and databases to simulate overall cellular responses. A bottleneck to this goal is modeling gene expression since the mechanism of this process is not yet fully unveiled. We are developing a small scale computer program called CiRMU (Cis-Regulatory Machinery Unit model) for describing, viewing, analyzing, and modeling the process of gene expression. A prototype system is being designed and implemented for analyzing functions of nuclear receptors.

1 INTRODUCION

Active studies on pathways and networks and development of their databases are being carried out in the fields of bioinformatics and biomedicine. Development of computerized databases of pathways and networks started in early 1990s, when many biomolecular structure data and their analysis packages were made available to the biological research communities. Today more than hundred pathway/network databases are listed in public websites [1]. These pathways/networks are classified into three categories; cell signaling pathway, metabolic pathway (metabolic map), and gene regulatory network. (We may add protein-protein databases as the fourth category.) We have developed one of the first-generation cell signaling pathway databases, CSNDB [2], and are interested in more innovative application of such databases. In particular we are interested in how we can go from pathway/network to disease, i.e., to explain a disease with intra-cellular pathways/networks and cell-cell molecular communications.

However as a relatively small research group we are not intending to provide large-scale, general-purpose standard reference type databases, which are usually supported by big national grants. Rather we are interested in small-scale handy software that can easily be used and modified by individual researchers to compile their own data and knowledge of gene expression related to a specific problem domain. The CiRMU (a Cis-Regulatory Machinery Unit) system emerged from such an idea. It is a handy computerized memo system to elucidate gene expression process. Development of the CiRMU system is a part of our Nuclear Receptor-Syndrome X (NR-SX) Project, a project to approach nuclear receptors and (metabolic) syndrome X problems from informatics and computing methodology [3]. This paper describes the concept and primary implementation of the system.

2 SYSTEM DESIGN CONSIDERATION

2.1 Purpose of the system

Gene expression is considered as one of the most important research area in the so-called post-genomic era [4]. Knowledge of genes has a limited value without the knowledge of how these genes are expressed and how their products (synthesize proteins) are produced. The main purpose of developing the CiRMU system is to provide individual researchers a tool to describe the expression process of a gene. A researcher can use the CiRMU system as a computerized memo that allows him/her to describe the logical relations among transcription factors, DNA biding sites (response elements), cofactors, and target genes as well as the ligands, product proteins and functions of transcription factors.

Unfortunately not all players in this arena have been identified but there are many that are only guesses and need to be further investigated. The system will be used not only to describe what has been identified but also what is guessed or speculated. For example, we know that nuclear receptors are transcription factors, but we do not know all target genes of a nuclear receptor yet. The system must help researchers to identify target genes of a given transcription factor. For that purpose the system not only helps researchers to find existing data and knowledge of known target genes but also provide some computer-based algorithmic tools to search for putative target genes. Same policies are applied to the other elements of gene expressions, namely, ligands of transcription factors, their cofactors, binding sites and so on.

2.2 Modeling gene expression

Gene expression is a complex process consisting of various steps ranging from the activation of a transcriptional machinery to the synthesis of a functional protein [5]. These steps are grouped into two main events: transcription and translation. In eukaryotes transcription occurs in the nucleus, while translation is carried out in the cytoplasm. Transcription is initiated by the activation of transcription factors that bind to specific DNA sequences, called response elements, situated upstream of the site at which transcription is started. Traditionally different steps of gene expression were considered to be independent events. In a contemporary view all steps involved in regulating gene expression from transcription to translation are believed to constitute a continuous process. Transcription produces

messenger RNAs, which are then used as templates for

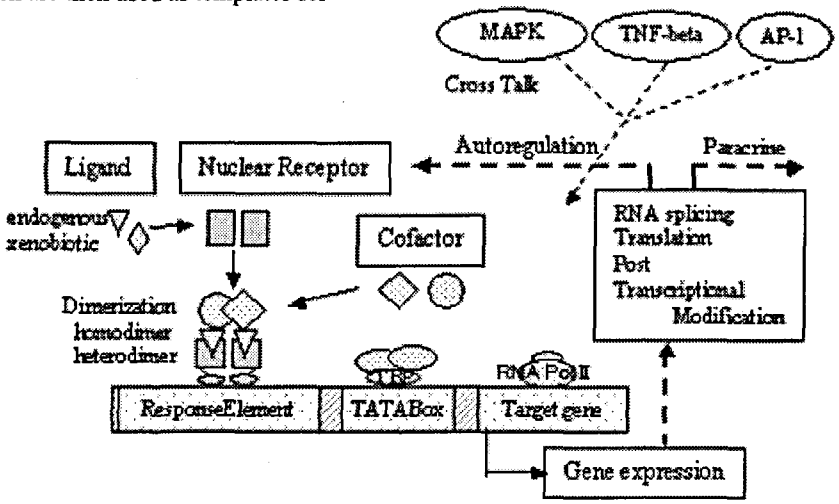

Figure 1: Nuclear receptor signal transduction pathway

synthesizing proteins. Proteins often experience the so called post translation modifications. The proteins thus produced may also influence the initial transcription step either by modifying (metabolizing) ligands of the transcription factors or changing the population of the transcription factors (receptors). In case of nuclear receptor, target genes sometimes include the activated nuclear receptor itself, forming a feedback loop (see Fig. 1).

In a typical case there are a number of factors that bind at multiple sites residing on both proximal and distal promoter (and enhancer) regions of a gene. Sometimes regulatory elements map at regions far from the gene. Such elements are called trans-acting elements. Often more than one transcription factor share the same binding site, and competition occurs. In addition cofactors/coregulators (co-activators and co-repressors) may also bind to these proteins. Since transcription is the terminus of a cell signaling pathway, cofactors/coregulators may transduce signals from cell signal pathways activated by phosphorylation cascade. Transcription factors and co-regulators interact and form a very complex functional aggregate. Some coregulators seem to be ligand-independent. Many new putative coregulators were identified by new technologies enabling more systematic search for protein-protein interactions. Yeast-two hybrid is such a method.

The basic function of the CiRMU system is to edit various data and knowledge related to gene expression by taking the above complexities into account. Special care must be paid for the rapid accumulation of data and discoveries, different modes of molecular interactions, and heterogeneous and complicated knowledge structure. The data and knowledge should be retrieved per gene, per cofactor, and per transcription factor base. Of course some other view points such as per ligand are also possible.

2.3 Analytical tools

In order to elucidate the gene expression process, the CiRMU system should also provide effective tools to analyze assembled data and knowledge. Although there are broad range of tools for this purpose, we are specifically focusing on two kinds of tools in developing our prototype. One is an algorithmic tool and the other is for network-based simulation.

The former aims to elucidate the binding sites of transcription factors and identify all target genes of a transcription factor. There are already computer systems to identify binding sites of transcription factors [6]. However nuclear factors are transcription factors whose biding sites are very difficult to predict. Such prediction methods are also known to produce too much false positives. Nevertheless they are still useful auxiliary tools, and the CiRMU system has interfaces to such prediction systems. Tools for network simulation are software that can represent biological pathways and networks in both diagrams and equations. The diagrams show how elements of gene expression relate causally, and the equations represent dynamics of pathways and networks mathematically. Analytical tools of the CiRMU system are highly application-dependent. Therefore we are adding applications one by one each time we apply the system to solve a specific problem.

2.4 Visualization Tools

Visualization methods such as tables, graphs, charts, diagrams, and 2D/3D images are powerful means for both checking and analyzing the data and knowledge contents stored in a system. In addition to visualization tools used in standard data analysis packages, the CiRMU system is equipped with interfaces to molecular visualization tools including 2D/3D molecular images and tools for visualizing pathways and networks. The latter are used to draw diagrams of pathways and networks and check the logical (causal) relations among elements involved in gene expression.

3 SYSTEM IMPLEMENTATION

We assume that years will be needed to fulfill our eventual

goal. This is due not to the lack of computational capacity or software tools but to the lack of data and knowledge. The pace of data and knowledge production in biomedicine is ever accelerated in an exponential manner. Thus researchers must struggle with a vast amount of data and knowledge. However if one focuses on a specific problem, for example aiming to identify the target genes of nuclear receptors, the binding sites for a specific gene to be expressed, or precise conditions that trigger the mRNA synthesis, one finds that data and knowledge are very sparse and that it is very difficult to elucidate appropriate logical scheme of the phenomenon. We therefore divided our implementation work into three stages.

Initial Stage

In the first stage, we survey basic literatures and existing informatics and computational resources that seem to be relevant to our problems. Today almost all of these informatics and computational resources are on the World Wide Web [7], and are available by on-line retrieval. These resources are expanding rapidly, but at the same time some of them are fading away. The entire information landscape is changing rapidly. This means that it is not easy to maintain the interfaces for accessing relevant outer resources. Therefore we are cautiously working on building the environment that allows us to search, identify, and retrieve relevant data easily. The literatures are made available in a text search environment. At this stage the CiRMU system seems to be just a mere unstructured collection of simple tables, charts and diagrams that describe gene regulatory processes.

Intermediate Stage

In the second stage, we try to describe gene expression and related processes using an explicit network structure. Here we use Petri net for representing causal relations among molecular interactions in gene expression processes. Petri net is suitable for representing both static and dynamic features of a network, and has been widely used in bioinformatics particularly for modeling metabolic pathways [8]. The Petri net subsystem must have an interface to the source data and knowledge body. The basic architecture of the latter is a collection of documents written in eXtensible Markup Language (XML) format. XML documents are textual data structured by tags. Like Petri net XML is also widely used in bioinformatics [9]. However all user interfaces of our system are written in the standard HyperText Markup Language (HTML), and the system has a CGI that converts HTML to XML and vice versa (see Fig. 2).

Final Stage

In the final stage, the CiRMU system will be integrated into a larger scale informatics and computing infrastructure for the Nuclear Receptor-Syndrome X project [3]. The resources will consist of (1) a computational chemistry workbench for simulating the docking of nuclear receptors to ligands and the interactions between a ligand-activated nuclear receptor (and co-factor) complex with DNA (response elements), (2) pathways and networks that contain nuclear receptors as their nodes, (3) cell-cell communication models via key metabolic syndrome molecules such as insulin, and (4) digital physiological models of Syndrome X (such as obesity, diabetes, hypertension, and atherosclerosis)

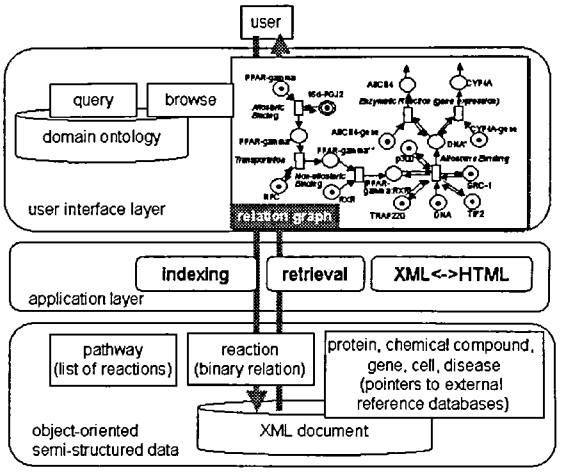

Figure 2: Image of system implementation

So far as hardware is concerned, the first to second stage systems are run on PCs, while the last stage system will be on a server type machine.

4 PRIMARY APPLICATION

We chose nuclear receptors as our primary application field. Nuclear receptors are ligand-activated transcription factors that form a superfamily. In case of humans there are 48 nuclear receptors almost half of whose ligands have been identified, leaving some as true orphans. These superfamily proteins play important roles in cell proliferation, differentiation, reproduction, and metabolism. They become the target of extensive studies in relation to endocrine and metabolic diseases. Nuclear receptors are important for pharmacology and drug design since many important drug metabolic enzymes such as cytochrome P-450s (CYPs) and transporters are coded by genes that are targets of (orphan) nuclear receptors. These enzymes and transporters are considered to be the main players of the so called ADME/Tox (absorption, distribution, metabolism, and excretion/toxicology). Therefore nuclear receptors are also important in terms of chemical safety (toxicology) as well as for nutritional study.

Considering these importances, one of the authors has proposed a large theoretical and computational research team be organized for collaborating with experimental and clinical researchers in attacking the problems of nuclear receptors and their related diseases. The proposal was named the NR-SX (Nuclear Receptor and Syndrome X) Project. Several computational chemistry and bioinformatics software tools and algorithms are concurrently being developed in this project. Development of the CiRMU system is a part of this project.

Right now basic data and knowledge have been accumulated manually based on about 400 selected papers on the topic of nuclear receptors and metabolic syndromes (see Fig. 3). Here the CiRMU system is used as means to elucidate causal relationships among ligands, nuclear receptors/cofactors, DNA response elements (binding sites), target genes, synthesized proteins and their functional feedbacks. Such data and knowledge are useful for

developing therapeutic agents called selective nuclear receptor modulators. They are also useful for predicting drug-drug interactions and realizing the so called personalized medicine.

One of the interesting applications of the CiRMU system is retrieval of feedback loops of some nuclear receptor signalings. It is known that some nuclear receptors are autoregulatory, in that, their target genes include the gene of the activated nuclear receptor itself. Liver X Receptor (LXR) is one such example. There are also chain reactions. Some such chain reactions involve more than three genes. However details of such a network are not yet fully unveiled. The CiRMU system is now used to extract such looped pathways.

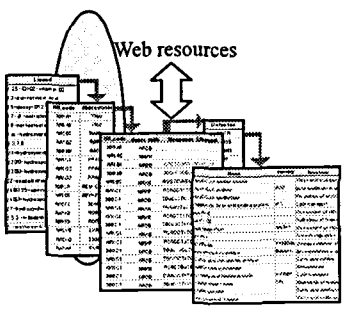

Figure 3: Reference databases and files

5 DISCUSSION

At present we are working on the first stage of the CiRMU implementation and concurrently working on some second stage designing. We are focusing our attention on gene expression process of nuclear receptors and their relations to some metabolic syndromes, particularly obesity. Thus adipocytes (adipose tissue cells) are chosen as a specific example.

There are several directions for improving our model and extending its application. The most promising approach is to investigate the dynamical features of gene expression. Advances in gene expression researches have unveiled very dynamic and pivotal roles of chromatin, which was once considered to be rather static during gene expression. Our present model of cis-regulatory machinery unit is rather static and logical. In this model, the bindings of transcription factors to DNA sequences are represented by simple on-off operations as in electric logical devices. We assumed that the target genes are expressed when the conditional transcriptional factors bind to pre-assigned response elements and auxiliary controlling cofactors bind to specific transcription factors just like a control input into a logical circuit. However this model cannot explain how the amount of a mRNA is controlled. Some transcription factors are competitive to each other in sequence binding, for they share the same response elements. It is assumed that the transcription factor that can bind tighter than the others may take the initiative, but we do not yet know the general rule of such mechanisms. For realistic simulation of such a process, one must wait advances in theoretical calculation and affinity binding experiments of protein-DNA interaction.

The second direction to deepen our system model is to elucidate the detailed features of the feedback circuitry of signals, in which nuclear receptors serve as the main regulators. For this purpose we must accelerate the compilation of knowledge on nuclear receptors using the CiRMU system. These feedback loop pathways (networks) plays a central role in cells serving as sensors and processors controlling the material metabolism and thus the overall systemic energy balance of the cell and the organism. Therefore accumulation of such knowledge will contribute to better usage of drugs, better understanding of metabolic syndromes, and better understanding of toxicology of endocrine disruptors.

The third direction is to apply our system to transcription factors other than nuclear receptors. We are particularly interested in applying our system for describing gene expression processes regulated by arylhydrocarbon receptors (AhR), NF-κB, and forkhead proteins (FOX) [10]. The first proteins are known as the dioxin receptor, the second proteins play a pivotal role in various immunological signal transductions, and the third proteins are involved in a wide range of developmental phenomena and some brain disorders (speech and language) [11]. We hope to improve our model and to increase the knowledge contents of our system through these applications.

Since pathway/network research is rapidly advancing, the environment of the CiRMU system is rapidly changing. Especially there is always a tendency to integrate existing databases and other information sources [12]. We must carefully watch such movements.

REFERENCES

[1] The Pathway Resource List (http://cbio.mskcc.org.prl).

[2] T. Igarashi and T. Kaminuma. Development of a Cell Signaling Networks Database. Pacific Symposium on Biocomputing '97, World Scientific, pages187-197, 1997.

[3] T. Kaminuma. Pathways and Networks of Nuclear Receptors and Modeling of Syndrome X. CBI Journal, 3: 130-156, 2003.

[4] F. Collins et al. A vision for the future of genomic research. Nature, 422:835-847, 2003.

[5] G. Orphanides and D. Reinberg. A Unified Theory of Gene Expression. Cell, 108:439-451, 2002.

[6] A. Sandelin. JASPAR: An open-access database for eukaryotic transcription factor binding profiles. Nucleic Acid Research. 32(Database issue):D91-4, 2004.

[7] G. Wei, D. Liu, and C. Liang. Charting gene regulatory networks: strategies, challenges and perspectives. Biochem. J. 381:1-12, 2004.

[8] M. Chen, A. Freier, J. Köhler, and A. Rüegg. The Biology Petri Net Markup Language. In Lecture Notes in Informatics: Desel J. et al. (eds.), Proceedings of Promise, pages 150-161, Potsdam, 2002.

[9] F. Achard, G. Vaysseix, and E. Barillot. XML, bioinformatics and data integration. Bioinformatics, 17:115-125, 2001.

[10] A useful information source of forkhead proteins is (http://www.biology.pomona.edu/fox.html).

[11] C. S. L. Lai et al. A forkhead-domain gene is mutated in a severe speech and language disorder. Nature, 413: 519-523, 2001.

[12] J. S. Luciano. PAX of mind for pathway researchers. Drug Dicovery Tody, 10(13):837-942, 2005.