

THE GRID AS A “Ba” FOR BIOMEDICAL KNOWLEDGE CREATION

AKIHIKO KONAGAYA

*Advanced Genome Information Technology Research Group
RIKEN GSC*

*1-7-22, Suehiro-cho, Tsurumi, Yokohama, Kanagawa, Japan
Email: konagaya@gsc.riken.jp*

Data-driven biology, typified by the Human Genome Project, has produced an enormous amount of experimental data, including data on genomes, transcriptomes, proteomes, interactomes, and phenomes. The next pursuit in genome science concentrates on elucidating relations among the data. Key to understanding these relations are bio-networks that incorporate information on protein-protein interactions, metabolic pathways, signal transduction pathways, and gene regulatory networks. Development of these bio-networks, however, requires biomedical knowledge of life phenomena in order to add biological interpretations. In this sense, data creation and knowledge creation play complementary roles in the development of genome science. As for knowledge creation, Ikujiro Nonaka proposed the importance of “ba”, that is, a time and place in which people share knowledge and work together as a community. Grid computing offers great potential to extend the concept of “ba” to networks, especially in terms of deepening the understanding and use of bio-networks by means of sharing explicit knowledge represented by ontology, mathematical simulation models and bioinformatics workflows.

1. INTRODUCTION

In April 2003, HUGO proudly announced the completion of the Human Genome Project [1]. The success of that project has opened doors for further post-genome sequence projects that produce genome-wide data at multiple levels, for example, on transcriptomes, proteomes, metabolomes, and phenomes, to name a few. The next challenge is to elucidate bio-networks that incorporate information on protein-protein interactions, metabolic pathways, signal transduction pathways, gene regulatory networks, and so on. To understand these bio-networks, it is necessary to introduce biological interpretations to molecular-molecular interactions and pathways. Scientific experts provide such interpretations implicitly. However, the explicit representation of biological knowledge through such structures as ontologies

and mathematical simulation models is necessary in order to be able to analyze bio-networks from a computational point of view.

Knowledge creation requires a time and place in which people share knowledge and work together as a community. Ikujiro Nonaka called this place “ba” [2], as originally proposed by the Japanese philosopher Kitaro Nishida [3]. “Ba” can be considered a type of superstructure similar to a virtual organization or community based on mutual trust. This paper discusses how to organize a “ba” for grid computing [4] from the viewpoint of biomedical knowledge creation, that is, for developing and interpreting computational bio-networks.

Section 2 of this paper describes the differences between data, information and knowledge using gene annotation as an example. Section 3 discusses the role of tacit knowledge in bioinformatics. Section 4 introduces knowledge-intensive approaches to drug-drug interaction based on ontology and mathematical modeling. Finally, Section 5 discusses the superstructure of grid computing necessary for creating and sharing biomedical knowledge.

2. DATA, INFORMATION AND KNOWLEDGE

Although the boundaries between data, information and knowledge are somewhat unclear, each differs from the viewpoint of “interpretation”. “Data” is self-descriptive in nature. We can transfer data from person to person without explanation, for example, the nucleotide sequence “atcg”. “Information” is objective in the sense that its interpretation is almost unique. Given a nucleotide sequence, you may find a coding region using the gene structure information together with information on the translation initiation and termination. On the other hand, knowledge is subjective in the sense that its interpretation depends on the individual’s background. For this reason, we need ontologies and mathematical models to represent “descriptive” knowledge that is shared or is supposed to be shared like information.

Genome annotation is a knowledge-intensive bioinformatics application that maps the structural and functional information of genes. To annotate a gene, in-depth understanding of the functionality of the gene is required, by integrating information from genome, transcriptome, proteome and other databases. Terms play an essential role in genome annotation. Imagine that we are given homologous sequences by BLAST search for a specific coding region. If the sequences are all ESTs (expressed sequence tags) or unannotated genes, our understanding of the coding regions is limited. However, when we are given terms such as ‘biotin carboxylase’ and ‘campylobacter’ in the annotation of the sequences, we obtain

more knowledge associated with the terms. The terms may remind biochemical experts of proteins and pathways related to the carbon dioxide fixation reaction. They may also remind medical or pharmaceutical experts of diseases, for example, cheilitis caused by the deficiency of biotin carboxylase and enteritis caused by campylobacter [5,6]. This example suggests two important aspects about terms. First, terms play a role in linking knowledge to other knowledge. Second, the semantics of a term depend completely on the expertise and knowledge of the scientist. We will discuss the characteristics of personal knowledge from the viewpoint of knowledge creation in the next section.

3. TACIT KNOWLEDGE, EXPLICIT KNOWLEDGE AND KNOWLEDGE SPIRAL

Michael Polanyi, a 20th-century philosopher, commented in his book, *The Tacit Dimension*, that we should start from the fact that 'we can know more than we can tell'. This phrase implies that computers are limited in their ability to represent knowledge, no matter how fast they can calculate and no matter how much storage they may have. Furthermore, in his book *The Knowledge-creating Company*, Ikujiro Nonaka observed that the strength of Japanese companies does not result simply from the solid management of explicit knowledge but from the establishment of common tacit knowledge. This does not, however, indicate the superiority of tacit knowledge over explicit knowledge. Explicit knowledge is important for analyzing and interpreting huge data sets such as genome sequences. To clarify this issue, Ikujiro Nonaka developed the concept of the "knowledge spiral," which turns tacit knowledge into explicit knowledge (externalization) and explicit knowledge into tacit knowledge (internalization), as shown in Figure 1.

Consider how the knowledge spiral could be applied to bioinformatics applications. A gene ontology has been developed to control the terminology used for genome annotation by strictly defining the relationship of terms [7]. From the viewpoint of knowledge spiral theory, genome annotation can be considered a type of knowledge transfer from tacit knowledge to explicit knowledge (externalization). Gene ontology also serves to label gene clusters through gene annotations (combination). These annotations then help biologists understand the functionality of genes (internalization). Finally, the process can be extended so that everyone in the community can share the same understanding of gene functionality (socialization). In this way, we can create new explicit knowledge (the annotation of genes) and tacit knowledge (an understanding of gene functionality) by repeating the above process throughout the community.

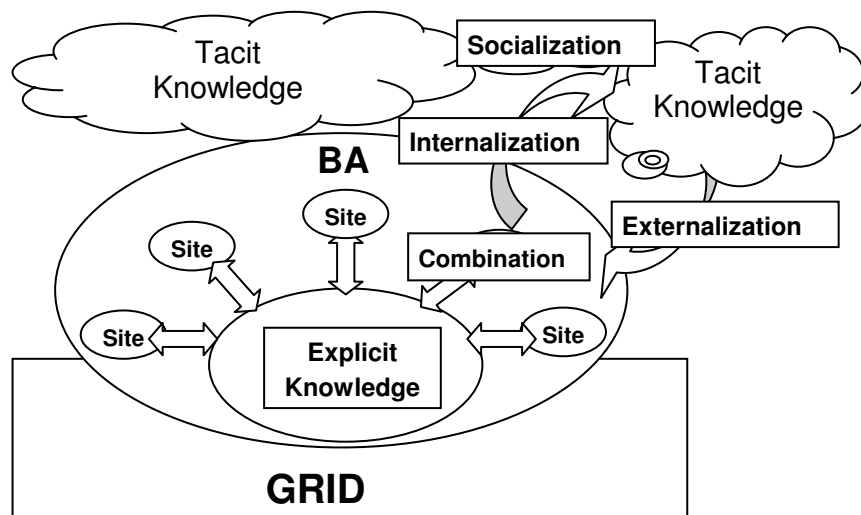


Figure 1. Knowledge Spiral on a Grid

Mathematical modeling of biological phenomena is another interesting application of the knowledge spiral to bioinformatics. Mathematical models model the dynamic behavior of biological phenomena, that is, the time-dependent state transition of molecular interactions. To design mathematical models, information on biochemical reactions and kinetic parameters is needed. Many efforts have been made to extract this information from literature databases [8, 9]. However, the information in the literature is fragmented and sometimes contradictory in terms of bionetwork modeling. Consider, for example, protein-protein bindings and mRNA expression profiles. Mathematical modeling also requires in-depth biological knowledge and strong mathematical skills in order to integrate information obtained from biological experiments. The development of mathematical models is still a state-of-the-art process performed by human experts knowledgeable in the life phenomena of interest [10]. Mathematical models, once established, however, can be extended to gene disruption models and over-expression models. They can assist our understanding of phenomena more deeply, helping us better understand the many efforts involved in biological experimentation. In this way, new explicit and tacit knowledge can be created.

4. KNOWLEDGE-INTENSIVE APPROACH TO DRUG-DRUG INTERACTION

Drug-drug interaction is a significant clinical problem. It was recently recognized that drug response strongly depends on the polymorphism of drug response genes such as cytochrome P450 (CYPs) [11]. Severe drug side effects may occur when a patient is a poor or extensive metabolizer of a given drug. This problem becomes more complicated when more than two drugs are concomitantly administered. To address this issue, we developed a drug-drug interaction prediction system based on a drug interaction ontology and a stochastic particle simulation system that incorporate drug-molecule interaction information extracted from the literature and sequence databases. The system will incorporate personal genome information in the near future.

When designing a drug-interaction ontology (DIO), we focused on a triadic relation consisting of *input*, *effector* and *output* [12]. The triadic relation represents the causality of molecular-molecular interaction in a cell. *Input* indicates a trigger of this molecular interaction. *Effector* indicates the central player in this molecular interaction. *Output* indicates the result of this molecular interaction. *Input*, *effector* and *output* for drug metabolism consist of a drug, an enzyme and a drug metabolite. Note that an *output* can be used as an *input* for a succeeding triadic relation. In this way, we are able to represent a drug metabolic pathway as a chain of triadic relations. Drug-drug interaction can be represented as triadic relations that share the same *input* or *effector* in their metabolic pathways.

A triadic relation can be extended to incorporate an indirect molecular interaction such as a metabolic pathway as well as a direct molecular reaction such as an enzymatic reaction. In other words, our system is able to represent metabolic pathways as a single triadic relation by ignoring intermediate reactions. The system is also able to represent the causality of a high-level molecular reaction, such as the inactivation of enzymatic function and the inactivation of drug function in cases for which biological observation is available but the molecular mechanism is unknown.

To date, we have extracted more than 3,000 interactions from the literature and entered these into the molecular interaction knowledge base. We have also developed a prototype system that infers the occurrence of drug-drug interaction in triadic relations.

A triadic relation is sufficiently powerful to represent drug-biomolecular interactions qualitatively, but is limited in its ability to analyze the dynamic behavior of quantitative information. Drug metabolism is highly non-linear, and drug response sometimes becomes sensitive to the initial drug dosage. This situation

becomes more complex when more than two drugs are concomitantly used, as shown in Figure 2. In the figure, 6-mercaptopurine (6MP) is an anti-cancer drug, and allopurinol is an anti-hyperuricemia drug used to reduce the purine bodies that often result from cancer therapy. It is well known, however, that allopurinol inactivates xanthine oxidase (XO), which metabolizes 6MP to thiourea, which ultimately is excreted in urine. Severe side effects may occur in patients unable to excrete 6MP [13].

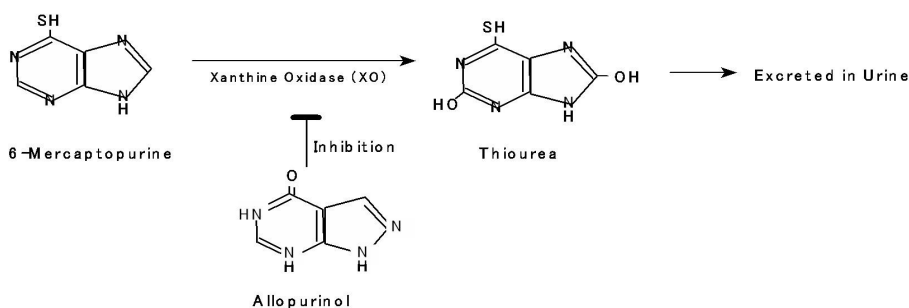


Figure 2. Drug-Drug Interaction between 6MP and Allopurinol

Although an ordinal difference equation system is the leading approach to analyzing drug metabolism processes, we adopted a spatio-temporal stochastic particle simulation to analyze trafficking processes, localizations, and membrane penetration [14]. The particle simulation system simulates a molecule as a particle that walks randomly on a 2D or 3D lattice grid unless boundary conditions, such as a membrane, are provided. Each molecular interaction may occur with a specified probability when reactive particles pass the same node of the grid. Our particle simulation is sound in the sense that the average behavior of a kinetic reaction is the same as that obtained from ordinal rate equations when the number of particles is sufficient and the particles are distributed uniformly. Our particle simulation can also account for the non-uniform distribution of particles and the behavior of particular molecules such as DNA, membrane structures, and receptor complexes. The particle simulation has good potential to overcome the limitations of conventional simulations that are based on ordinal differential equation systems, partial differential equation systems, and other deterministic simulation systems.

5. SUPERSTRUCTURE FOR BIOMEDICAL KNOWLEDGE CREATION

The concept of the grid computing has great potential to accelerate knowledge creation by extending knowledge spirally throughout the network community. Knowledge spiraling requires a platform upon which people share knowledge and work together. A grid enables users to share data, information and knowledge as well as computational resources. It should also emphasize the social aspects, or the superstructures constructed on the IT platform, that play an essential role in collaborative works in virtual organizations [15, 16]. We will discuss these issues from the viewpoints of community formulation, service interoperability and intellectual property development.

As described in the introduction, development of community is a basis of “ba” for knowledge creation. Grid users with network accounts establish a community to share data, software and computers over the network. In the case of a grid that uses Globus Tool Kit, the boundary of the community is restricted by the availability of Globus accounts to access the remote computers. The boundary can be relaxed if a single representative login account is provided for access to the remote computers as shown in the Open Bioinformatics Environment (OBIEEnv) [17]. OBIEEnv enables a local user to access the remote computers through the representative account if the local user has a valid account on a local machine. The community can thus be extended to the total number of local accounts on the grid.

The use of a reverse proxy server is another approach to extend the community. The reverse proxy server can provide the portal accounts needed to control access to the necessary web pages. The portal accounts enable the extension of the community to non-computing professionals, such as experimental biologists who prefer to access data and tools through web pages. As of July 2005, more than 400 portal users had registered for the bioinformatics applications on the Open Bioinformatics Grid (OBIGridⁱ). One successful example is the genome annotation support system (OBITco) developed for the *Thermus thermophilus* research community in Japan [18].

Grid services and workflows enable portal users to automate bioinformatics activities performed using hands-on web applications. Most public databases and well-known freeware bioinformatics applications are already available as web servicesⁱⁱ. Knowledge management, that is, the sharing and reproduction of bioinformatics workflows, is a key challenge for knowledge creation.

ⁱ <http://www.obigrid.org/>

ⁱⁱ <http://industry.ebi.ac.uk/soaplab/ServiceSets.html>

Interoperability is important to ensure proper data transfer among applications. XML formats enable input and output data to be described in an architecturally independent manner. However, bioinformatics workflows require interoperability of semantics as well as data format. Let us consider a simple bioinformatics workflow for the annotation of a microbial genome: Glimmer2 [19] for gene finding, BLAST [20] for homology search, and CLUSTAL W [21] for multiple alignment. This workflow seems reasonable, but would produce unsatisfactory results if the commands were consecutively executed. This is because Glimmer2 may produce too many gene candidates. BLAST may return very similar but unimportant sequences such as EST (expressed sequence tag) or annotated sequences. We therefore require filtering processes that are able to eliminate redundant and irrelevant data from the computational results [22]. A bioinformatics service ontology is therefore needed in order for the community to make use of and share bioinformatics workflows.

Intellectual property on a grid is another important issue to be resolved from the viewpoint of knowledge creation. Who owns the intellectual property when new knowledge is created? How are copyrights, database rights, patent law, ethics, personal and human rights to be considered? It may be possible to apply a general public license (GPL)ⁱⁱⁱ or other freeware license to research products created on a grid. A new licensing framework may be necessary for commercial products. Either way, this important issue needs to be resolved over the long term.

6. CONCLUSION

This paper discussed our experience with and perspectives on biomedical knowledge creation on the Open Bioinformatics Grid (OBIGrid) from the viewpoint of community development, service interoperability and intellectual property development as well as web services and workflow made available through grid computing.

Grid computing has great potential to become a platform for biomedical knowledge creation. The key to this knowledge creation is the transfer of knowledge from tacit knowledge to explicit knowledge. Ontology and mathematical modeling play essential roles in the representation of descriptive biomedical knowledge.

The knowledge spiral also requires “ba,” a place in which people share knowledge and work together. Web services and workflows for bioinformatics help extend the community for the purposes of knowledge sharing. However, much remains to be done in terms of enhancing the interoperability of the services and protecting intellectual property rights that arise from the development of new knowledge.

ⁱⁱⁱ <http://www.gnu.org/licenses/>

ACKNOWLEDGEMENT

The author expresses special thanks to Dr. Sumi Yoshikawa, Dr. Ryuzo Azuma and Dr. Kazumi Matsumura of RIKEN GSC for intensive discussions on drug interaction ontology and stochastic particle simulation. He also thanks Dr. Fumikazu Konishi, Mr. Ryo Umetsu and Mr. Shingo Ohki of RIKEN GSC and his students at Tokyo Institute of Technology for fruitful discussions and the implementation of the grid and web services on OBIGrid.

REFERENCES

- [1] Collins F. S., Morgan M., and Patrinos A., The Human Genome Project: Lessons from Large-Scale Biology, *Science*, p.286, (11 April 2003).
- [2] Nonaka I., Toyama R., and Konno N., SECI, Ba and leadership: a unified model of dynamic knowledge creation, *Long Range Planning*, Vol. 33, pp. 5-34, (2000).
- [3] Kitaro Nishida. *An Inquiry into the Good*, translated by Masao Abe and C Ives. New Haven, USA: Yale University Press, (1990/1911).
- [4] Konagaya Akihiko and Satou Kenji (Eds), *Grid Computing in Life Science. Lecture Notes in Bioinformatics*, Vol. 3370, (2005).
- [5] Forbes GM, Micronutrient status in patients receiving home parenteral nutrition, *Nutrition*, Vol. 13, pp. 941-944, (1977).
- [6] Melamed I., Bujanover Y., Igra Y. S., Schwartz D., Zakuth V., and Spierer Z., *Campylobacter enteritis in normal and immunodeficient children*, *Am. J. Dis. Child*, No. 137, pp. 752-753, (1983).
- [7] The Gene Ontology Consortium, *Gene Ontology: tool for the unification of biology*, *Nature Genetics*, Vol. 25, pp. 25-29, (2000).
- [8] Nagashima T., Silva D.G., Petrovsky N., Socha L.A., Suzuki H., Saito R., Kasukawa T., Kurochkin I.V., Konagaya A., and Schoenbach C., *Inferring higher functional information for RIKEN mouse full-length cDNA clones with FACTS*, *Genome Research* Vol. 13, pp. 1520-1533, (2003).
- [9] Martin-Sanchez F., et al., *Synergy between medical informatics and bioinformatics: facilitating genomic medicine for future health care*, *J. Of Biomedical Informatics*, Vol. 37, pp. 30-42, (2004).
- [10] Hatakeyama M., Kimura S., Naka T., Kawasaki T., Yumoto N., Ichikawa M., Kim J.H., Saito K., Saeki M., Shirouzu M., Yokoyama S., and Konagaya A., *A computational model on the modulation of mitogen-activated protein kinase (MAPK) and Akt pathways in heregulin-induced ErbB signaling*, *Biochemical Journal*, No. 373, pp. 451-463, (2003).
- [11] Keuzenkamp-Jansen C.W., DeAbreu R.A., Bokkerink J.P., Lambooy M.A., and Trijbels JM., *Metabolism of intravenously administered high-dose 6-mercaptopurine with and without allopurinol treatment in patients with non-Hodgkin lymphoma*, *J Pediatr Hematol Oncol.*, Vol. 18, No. 2, pp. 145-150, (1996).

- [12] Ingelman-Sundberg M., The human genome project and novel aspects of cytochrome P450 research., *Toxicol Appl Pharmacol.*, (29 June 2005).
- [13] Yoshikawa S., Satou K., and Konagaya A., Drug Interaction Ontology (DIO) for Inferences of Possible Drug-drug Interactions. In: *MEDINFO 2004*, M. Fieschi et al. (Eds), IOS Press, pp. 454-458, (2004).
- [14] Azuma R., Yamaguchi Y., Kitagawa T., Yamamoto T., and Konagaya A., Mesoscopic simulation method for spatio-temporal dynamics under molecular interactions, *HGM2005* (Kyoto, Japan), (2005).
- [15] Kecheng L., Incorporating Human Aspects into Grid Computing for Collaborative Work, Keynote at ACM International Workshop on Grid Computing and e-Science (San Francisco), (21 June 2003).
- [16] Konagaya A., Konishi F., Hatakeyama M., and Satou K., The Superstructure toward Open Bioinformatics Grid, *New Generation Computing*, No. 22, pp. 167-176, (2004).
- [17] Satou K., Nakashima Y., Tsuji J., Defago X., and Konagaya A., An Integrated System for Distributed Bioinformatics Environment on Grids., Springer, LNBI, Vol. 3370, pp. 8-18, (2005).
- [18] Fukuzaki A., Nagashima T., Ide K., Konishi F., Hatakeyama M., Yokoyama S., Kuramitsu S., Konagaya A., Genome-wide functional annotation environment for *Thermus thermophilus* in OBIGrid, LNBI, Springer, Vol. 3370, pp. 32-42, (2005).
- [19] Delcher A.L., Harmon D., Kasif S., White O., and Salzberg S.L., Improved microbial gene identification with GLIMMER *Nucleic Acids Res.*, Vol. 27, No.23, pp. 4636-4641, (1999).
- [20] Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., and Lipman D.J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Research.*, Vol. 25, pp. 3389-3402, (1997).
- [21] Chenna R., Sugawara H., Koike T., Lopez R., Gibson T.J., Higgins D.G., and Thompson, J.D., Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Research*, Vol. 31, pp. 3497-3500, (2003).
- [22] Umetsu R., Ohki S., Fukuzaki A., Shinbara D., Kitagawa T., Hoshino T., and Konagaya A., An Architectural Design of Open Genome Services (OGS), *Life Science Grid 2005* (Biopolis, Singapore), (2005).