Clinical Genome Informatics (CGI) and its Social Informational Infrastructure

Jun Nakaya†

†*Clinical Genome Informatics Center, School of Medicine, Kobe University, 650-0017 Japan*

Summary
CGI is an essential informatics to support genomic medicine that is a medicine based on genome information. We believe that CGI domain must be a key technological field to establish the gene based medicine or pharmacogenomics. We analyzed the requirements to the social informational infrastructure in CGI domain and have developed the required technologies. In concrete terms, the coupling technology of the electronic health record (EHR) and the database is a technological basis. The public key infrastructure (PKI) is also important to secure the genome information that is ultimate personal information. Sharing tracks with the translational research informatics (TRI) is also required to make CGI practical and public. The CGI is inevitable and essential to establish the genomic medicine that is one of the goals of the post genomic researches. This paper reviews the definition, its requirements, its technological background, and the perspectives of CGI, and then discusses about the social informational infrastructure that is demanded in CGI domain.

Key words:
Clinical Genome Informatics, CGI, TRI, Genomic Medicine, PKI, EHR, informatics, knowledge processing, GSVML

1. Introduction to CGI

We can define that CGI is an Informatics aiming to support human health (especially Genome Medicine) based on the Genome Information [1]. In current context, the CGI can be positioned as an integrated informatics of the genome informatics and the clinical informatics. Here, the genome informatics is a kind of bioinformatics that is based on genome information, and the clinical informatics is a kind of medical informatics aiming to support clinical practice. The CGI is expected to be a main fundamental academic field in the post genomic era that is an era to bridge genome information to the phenome information for the effective use of the information [2].

The objective of the CGI is to establish the informational support of the clinical medicine including the clinical practice and the developmental process of the practices. In plain words, the CGI try to offer the informational support to the genome medicine and its developmental processes.

In brief, we can define that the genome medicine is a medicine based on the genome information [3]. The genome is the coinage of gene and ome, and ome is derived from the chromosome. Then the genome means the total of masses of genes. So the genome information is the conformity of gene information and their mutual relationship information. To relate the genome information with the genome medicine, we must relate the genome information with the phenome information at the beginning [4]. The phenome information is a substantial expression result on human body, and there are many hierarchical omics information such as molecular network to relate the genome information to the phenome information. We need more information about the abnormal state information such as the disease information, medical information, and the environmental information. Here the medical information is consisting of the diagnostic information and the treatment information, and the environmental information can be positioned as the cause of the disease.

After the declaration about the completion of the human genome sequence analysis project [5], the post genomic era began. Shortly the NIH road map indicated the direction of the biomedical research [6]. The FDA covered the NIH road map in white book on 2004 [7]. We have lots of things to establish the genome medicine. The remained things are the integration of the genome information, the reconciliation of the issue of the gene distribution and the genetic selection, DNA bank establishment, laws to support the genome medicine, and many. Nowadays we can say that we are in the era of integration. From the view of the biomedical technological view, a part of the genome medicine can be realized before the 2010 [8].

This genome medicine has three characters such as the personalized medicine, the initiative medicine, and the scientific medicine. The personalized medicine is the medicine per country per locality per personal. The initiative medicine is the predictive preventive medicine. The scientific medicine is the molecular evidence based medicine. The current targets of the genome medicine are the drug discovery, the personalized medicine, and the economical medicine. Anyway, the application at real world is the key issue of the genome medicine.

In the future medicine, the genome medicine will be able to play a big role at the personalized medicine and the disease prevention, while the genome medicine will play a small role at the unpainful medicine and the total

Manuscript received December 10, 2006.
Manuscript revised January 15, 2007

personality medicine with the mind care. Here the role of informatics in the genome medicine will be the informational support to achieve the following three clinical objects:
1. Improvement of the "clinical safety"
2. Improvement of the "clinical efficiency" or the "clinical effects"
3. Improvement of the "economical efficiency"

These are the priority order, and the third one should not be the first one. So we can also define that the specific character of the CGI is the informatics that is directly associated with human life.

2. Domain analysis on CGI

As discussed in Translational Research Informatics (TRI) [9], three dimensions are important to evaluate the technologies that would be used in the clinical domain [10]. Considering that CGI domain is a part of clinical domain, the CGI domain should follow the discipline in the clinical domain. Based on this three dimensional analysis in CGI domain, the basic important technology is defined as the integration/systematization of information and knowledge [1]. In other words the collection or assemble of information and knowledge is the basic issue. Based on the integrated information, we can establish more sophisticated informational infrastructure for CGI.
Considering that the difference between the current medicine and the genome medicine is caused just by the quantity of data. The precise and systematized integration of information, data, and knowledge are essential fundamental of the CGI infrastructure.

The systematized information/knowledge also can be basis of the cause and effect relational model. The specific technologies are the ontology as a technology for index or knowledge or terminology, the Markup Language as a technology of data exchanging format, and the Integrated database as a technology for data storing. Here the compatibility with the other Markup Languages must be considered because the database technology premises the data exchanging. The data and knowledge derived from these technologies must be sharable internationally from a view of the availability of the data. This implies that these technologies must link up with the international standardization activities such as International Organization for Standardization (ISO) [11], Health Level Seven (HL7) [12].

The methodology for the prediction is essential in CGI domain. Because CGI should stand on the point that genome medicine is based on the larger quantity of data derived from genome information. The target of the prediction is changed currently to the clinical path from the diagnosis. Here the clinical path is a path that a patient will track back clinically. The molecular target also can be the target.
Actually the development of the prediction methodology is not easy, in addition, the basic data infrastructure has not constructed yet. So the practical way to establish the prediction in current context is to develop the cause and effect prediction model. It can be used to predict the results of human based clinical trial or economical effect. An examples is a cause and effect model of "biosystem - human clinical result correlation table" at neoplastic patient at prototype design or discovery phase. The object of this technology is to avoid the critical and unfruitful trial or the unproductive investment based on the prediction. This prediction technology does not need to stand on the deterministic model, if the prediction model has enough accuracy. Considering the context that the current deterministic technology does not have enough technological level that can predict the complicated phenomenon as clinical facts with enough exactness, adopting the cause and effect models that have enough accuracy is a reasonable choice to establish the prediction model.

The bench mark that can test the accuracy of the prediction is essential to establish the prediction. The benchmarks for causal effect models and the logical background of evaluations are required. The novel biomarker or the statistic marker and measurement technology having causal relation with the human clinical result or the final economical effect of the probes can be the basis to establish the prediction.

The quantification of knowledge is inevitable to calculate the knowledge and the prediction. To achieve the precision prediction, we need to quantify the prediction to compare the accuracy of the prediction. To quantify the prediction, we need the quantified knowledge basically to calculate the predictions. The biomedical knowledge that is described with the multi dimensional features [13] may have the potential of the quantification.
The establishment of the personalized statistics is desired to analyze the personalized medicine.

To earn the social consensus to use CGI technologies in clinical practice, the CGI must be coupled with the TRI. In this way, the TRI is essential to establish the CGI. The derived technologies from TRI can be a basis of CGI. Some examples are the clinical protocol management methodology, the optimized clinical planning, and the TR management methodology from informatics aspect.

2. Social Informational Infrastructure of CGI

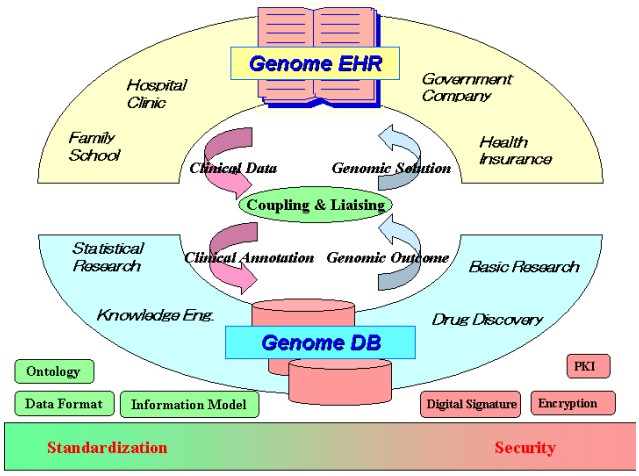

Fig. 1 Social Informational Infrastructure of CGI

The fundamental information infrastructure for the genome medicine can be classified into two categories such as the social informational foundations and the informational Infrastructure for Translational Research (TR). The social informational infrastructure includes the standardization and the security. The standardization is essential for sharing, exchanging, and effective use of the data. The security is essential to protect the ultimate personal information. The informational fundamentals for TR are essential as a social pipeline to realize the genome medicine. This means same as the word "bench to bed side". The informatics for TR is defined as the TRI in previous paper [9].

In concrete terms, the main two technologies of the social informational fundamentals for genome medicine are the genome Electronic Health Record (EHR) and the clinical genome database (Fig. 1). I must add that these two technologies are complimentary and must work coupled together. By engaging these coupled two technologies, the social informational infrastructure can offer the personalized medicine at hospitals or schools, while it can offer the clinical information to the meta analysis. This infrastructure can collect the clinical genomic information, while the infrastructure reflects the research fruits in the clinical practice.

More specifically three supporting elemental techniques for these two technologies must be the ontology, the information model, and the data format. According to the security, the current elemental techniques for the data security must be the public key infrastructure (PKI), the digital signature, and the encryption.

The practical pattern to apply the genome scientific findings to the clinical practice must pass the social pipeline that is called as the Translational Research phase (TR). All novel applications derived from the genomic researches must pass the social pipeline including the TR phase and the clinical trials to earn the legal background and social consensus. Here the essences of what pass the pipeline are data, information, and knowledge. Then the informational fundamentals that can support the transition process and the data passing are significant.

In this way, the CGI must be based on the TRI which needs above informational infrastructure that should have an information cycle as for the utilization and the collection of information. This infrastructure is required as a social platform to open the door to the genome medicine.

3. Current Status

The current status of the CGI is in the stage of initial booting world-widely. Presumably it is in the construction phase of the data collection platform. The elemental techniques of the data collection platform are the information model, the ontology, and the data format. These are important to establish the integration of the genome EHR, the clinical genome database, and the ontologies. As for the ontology, the Gene Ontology (GO) is the biological ontology mainly for the genome [14], while the SNOMED-CT (Systematized Nomenclature of Medicine for Clinical Terms) is the clinical terminology for clinical use [15]. With holding many discussions, these go near to be an international standard. In US and UK, the national projects as the integration of these ontologies and the connection with the messaging models have started. In Japan, there is a project of clinical omics ontology framing to support the TR [16]. The basic platform for the CGI system is going ahead. As for the data exchanging format, the Genomic Sequence Variation Markup Language (GSVML) is in process of the international standardization at ISO [17]. The GSVML is the data exchanging format of genomic sequence variation data to use it mainly in human health. The object of GSVML is to enhance the data exchanging of genomic sequence variation data with focusing on SNP mainly for human based clinical use internationally. The GSVML receives an expectation to contribute the international collection of clinical genomic data. Recently the GSVML developing team started the development of the interfaces to the HL7 ver3 and the statistical program packages such as SAS [18].

In Health Level Seven (HL7), the Clinical Genomics Special Interest Group is working on the CGI. They are in the development of the genetic information model through discussions. This information model will be a part of the HL7 RIM (Reference Information Model) [19]. The international standardization of the HL7 RIM is in process. The GSVML project is in collaboration with the HL7 CG SIG activity, and the data structure of the standardized genome EHR will reflect these efforts. Currently the interface between the HL7 ver3 genotype model [20] and the GSVML is in development.

5. Perspectives

Getting into the post genomic era, many germ of efforts for CGI have started globally. The genomic data had already overswollen internationally, and they are waiting for the effective use.

The knowledge informational infrastructure project had started as a toolbox for the researchers corresponding to the term of "new pathway to discovery" in the NIH roadmap [6]. The main focus is in the integration and the utilization of the knowledge for the effective use of the scattered data on this Internet society. This infrastructure is principal to promote the molecular targeting development. This infrastructure is also important to establish the CGI.

As noted in previous section, the GSVML project is going on. This project is international and tries to support the data exchange of the clinical genomics data internationally. The other project such as the CGI unit [21], the omics based medical informatics unit [22], or ClinicalBioInformatics unit [23] are going on in Japan. These projects try to train the experts for the CGI domain in practical way. Like other countries, many projects that try to reveal the meaning of the genome information are going on. As for the TRI, a NTRSC (National Translational Research Support Center) concept is in under contemplation [24] in Japan. Basically this project is prototyping the informational platform infrastructure to collect the nation-wide information.

Generally the IT system is recognized as the infrastructure to enhance the availability of data in other countries such as Australia. In those countries, they started with the integration of the distributed patient databases [25] to utilize the accumulated patient data that is a kind of legacy. These approaches are reasonable to enforce the foundation of the CGI. In collaboration with these efforts, we must put forward the development of the social informational infrastructure of CGI to move forward internationally.

7. Conclusion

The CGI is the key informational technological domain to open any future medicines. The point and the difference of the CGI domain are in a massive amount of genomic data. All requirements and technological backgrounds are derived from this fact. To achieve the strategic utilization of the overswollen genomic data in CGI domain, we must promote the continued effort to integrate the internationally distributed data as social informational infrastructure. Thus we also should develop the practical technologies based on the clinical demands. The CGI is an essential informatics to fruit the genomic researches practically. I believe that the CGI must be a social informational infrastructure with the TRI at the end of the post genomic era.

Acknowledgment

I wish to thank Tetsuo Shimizu and Hiroshi Tanaka who are the core members of the CGI and TRI study group.

References

[1] Jun Nakaya : "Orientation of the Clinical Genome Informatics", lecture text of the Clinical Genome Informatics Educational Unit, Kobe, 2005, pp1-27

[2] Lawrie W Powell : Bridging the gap between basic science and clinical medicine: mentors and memories, MJA 177 (11/12): 657-660, 2002

[3] Shigetaka Asano, Jun Nakaya : Key Note "Hurdles to establish Genome Medicine and the NTRSC", Proceedings of the 1st Symposium on Genome Medical Informatics, Tokyo, 2004, pp15-18

[4] Jun Nakaya, Tetsuo Shimizu : Knowledge Architecture based on Evidence Based Logical Atomism for Translational Research, International Journal of Computer Science and Network Security, 6 (2), ISSN: 1738-7906, 175-179, 2006

[5] Francis S. Collins, Michael Morgan, Aristides Patrinos, The Human Genome Project: Lessons from Large-Scale Biology, Science, 2003, 286

[6] Elias, Z. The NIH Roadmap. Science. Oct Vol. 302. 3, 63-72, 2003

[7] U.S.A., DHHS, FDA (eds.): Innovation or Stagnation; Challenge and Opportunity on the critical path to new medical products. FDA Report, US Department of Health and Human Services Food and Drug Administration, Washington, Mar. 2004

[8] Ok Baek, Theresa Gaffney, Kris Joshi, Barry Robson, David Rosen, Cathi Stahlbaum, Ruth Taylor, Pnina Vortman.: Personalised healthcare 2010: Are you ready for information-based medicine?, Executive strategy report, 2004. <http://www-935.ibm.com/services/au/index.wss/ibvstudy/igs/a1007746?cntxt=a1005069> (URL) [accessed November 28, 2006]

[9] Nakaya, J. The Translational Research Informatics (TRI) (Leadoff Article) International Journal of Computer Science and Network Security. 6(7A), 2006, pp117-122

[10] Tetsuo Shimizu, Jun Nakaya, et al : Report on RR2002 Project, Tetsuo Shimizu (eds.), Ministry of ECSST（Tokyo), 2004, pp4-45

[11] International Organization for Standardization (ISO) : TC215 Health Informatics . ＜ http://www.iso.org/iso/en/ISOOnline.frontpage＞（URL）[accessed July 11, 2006]

[12] Health Level Seven : What is HL7 ＜http://www.hl7.org/＞（URL）[accessed July 11, 2006]

[13] Nakaya, J., Sasaki, K., and Tanaka, H. Condensed Cross-Clinical Knowledge, Computer Science, International Journal of Computer Science and Network Security, 6 (7A), 2006, pp6-11.

[14] Ashburner, M, Ball, CA, Blake, JA, Botstein, D, Butler, H, Cherry, JM, Davis, AP, Dolinski, K, Dwight, SS, Eppig, JT, Harris, MA, Hill, DP, Issel-Tarver, L, Kasarskis, A, Lewis, S, Matese, JC, Richardson, JE, Ringwald, M, Rubin, GM, Sherlock, G.: The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. Nat Genet 25, 25-29, 2000)

[15] Stearns, MQ., Price, C., Spackman, KA., Wang, AY.: SNOMED clinical terms: overview of the development process and project status. Proc AMIA Symp. , 662-666, 2001

[16] Cyclopedic disease molecular pathological database project (in Japanese). ＜http://bioinfo.tmd.ac.jp/omix/db/＞（URL）[accessed November 28, 2006]

[17] Jun Nakaya : Genomic Sequence Variation Markup Language (GSVML), ISO/#25720 , International Standard Organization TC 215 WG2, N422 Committee Draft, 1-132, 2005

[18] Nakaya, J., Genomic Sequence Variation Markup Language (GSVML) : The First International Standard in Clinical Genome Informatics (CGI) Domain. CJKMI 2006 Proceedings, Plenary Meeting key note, 2006, pp5-8

[19] Health Level Seven, Inc. HL7 Reference Information Model. Ann Arbor, ML: Health Level Seven, Inc., 1994. ＜ http://www.hl7.org/library/data-model/RIM/modelpage_non.htm＞ (URL) [accessed July 11, 2006]

[20] Clinical Genomics Domain Information Model (HL7 POCG_DM000020). ＜ http://www.hl7.org/v3ballot/html/domains/cg/editable/POCg_DM000020.htm.＞ (URL) [accessed July 11, 2006]

[21] Clinical Genome Informatics research unit . ＜ http://www.med.kobe-u.ac.jp/cgi/torikumi/index.html ＞（URL）[accessed November 28, 2006]

[22] Bio-Medical-Omics Informatics research unit (in Japanese). ＜http://bio-omix.tmd.ac.jp/index.php＞（URL）[accessed November 28, 2006]

[23] ClinicalBioInformatics research unit . ＜ http://cbi.umin.ne.jp/＞（URL）[accessed November 28, 2006]

[24] Nakaya J, Shimizu T, Tanaka H, Asano S : Current Translational Research in Australia and Translational Research Supporting Center (TRSC) in Japan, Chem-Bio Informatics Journal, CBIJ1595- 5(2), 27-38, 2005

[25] BIO21 Australia ltd.: BIO21 member institutions <http://www.bio21.com.au/members.asp> (URL) [accessed July 11, 2006]

Jun Nakaya received the B.S. and M.S. degrees in Mech. Eng. from Hokkaido Univ. in 1985 and 1987, respectively. After staying in IBM, he received M.D. and Ph.D. degrees from Hokkaido Univ. in 1995 and 1999, respectively. He stayed in M.I.T., Institute of Medical Science, Univ. of Tokyo, and Tokyo Medical and Dental Univ.. Now he is an Associate Professor of graduate school of medicine, Kobe Univ.. He is a member of ISO, HL7, AMIA, ISMH, MIT-J, SSJ, and MSJ.