Transforming Healthcare through Translational Bioinformatics

Ju Han Kim

Abstract

Bioinformatics is a rapidly emerging field of biomedical research. A flood of multiple high throughput genomic and post-genomic data means that many of the challenges in biomedical research and healthcare are now challenges in integrative and computational sciences. Postgenome informatics, powered by multiple high throughput technologies and genomic-scale databases, is likely to transform our biomedical understanding forever much the same way that biochemistry and molecular biology did a generation ago.

In this talk, I will describe how these technologies and translational bioinformatics will impact biomedical research and clinical practice, introducing recent advances in integrating many different but complementary genomic profiles with the emphasis on the necessity of tight integration of private and public databases and integrative informatics technologies.

I will introduce some of our research efforts for translational bioinformatics approaches. Xperanto (Expressionist's Esperanto in XML) as a part of Bio-EMR system semantically integrates major data models for DNA microarray, tissue microarray and arrayCGH data with extended clinical and histo-pathological information models and analysis tools in an effort to establish a comprehensive knowledge and trial base for translational bioinformatics research. BioEMR is designed to support clinical trial informatics infrastructure and believed to realize the vision of 'clinical trial as a byproduct of clinical practice'. Each step will be given with real examples from ongoing research activities in the context of clinical relevance.

Key words BioEMR, clinical information, comparative genomic hybridization, DNA microarray, Health Watch, tissue microarray, translational bioinformatics.

Introduction

Bioinformatics is a rapidly emerging field of biomedical research, and as genomic and biomedical data accumulate, many of the challenges in biomedical research and healthcare are becoming challenges in integrative genomics and computational sciences. Post-genome informatics, powered by multi-modal high-throughput technologies and genomic-scale databases, is transforming our biomedical understanding in much the same way that biochemistry did a generation ago.

In this talk, I will describe how these technologies and translational bioinformatics affect both biomedical research and clinical practice, using examples from ongoing research activities with clinical relevance. I will focus in particular on how genomic profiles can be integrated into bioinformatics databases.

Genomics, Bioinformatics, and Clinical Care

I will present our vision regarding the ongoing convergence of genomics with bioinformatics and clinical medicine. Among scientific advancements in this area, of particular note is the use of gene expression profiles to guide drug therapies. For example, how does a physician decide whether to administer the toxic anti-cancer drug Tamoxifen to a patient after a mastectomy in the absence of any lymph node involvement, given that the rate of cancer recurrence in lymph node-negative disease is very low, but is not zero? A physician must

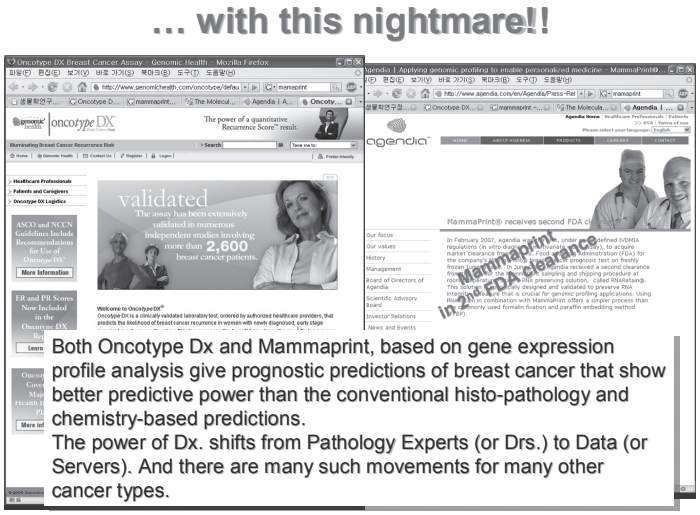

Figure 1

balance the possibility of administering a toxic drug to a patient who may never have a recurrence against the possibility of failing to administer a drug with proven efficacy to a patient who may suffer a recurrence. Molecular diagnostic test kits have been developed to help make such decisions. Based on the expression profiles of about 70 genes, MammaPrint applies prognostic predictors for breast cancer to suggest an optimal course of treatment for a specific patient (Figure 1). OncoType Dx is another approach with similar strategy.

Although this promising technology can be invaluable, its use can also create new problems. In fact, there exists the potential for a nightmare. Traditionally, a surgeon will extract a tissue sample from a patient, submit it for analysis, and receive an informational report back from the pathologist and clinical laboratory. The physician then makes a treatment decision based on this information. However, with the use of tests based on gene expression profiles, a physician receives a treatment decision, rather than information. These tests use an analytical algorithm to weigh many factors in a single kit, and the physician is probably not even aware of how a treatment decision is made. The test kit is in a black box. This is actually happening in the world today. Furthermore, companies who are selling products in Korea and Japan at a price of US$3,000 are not required to sell through hospitals, but are marketing products directly to patients.

Thus far, the US Food and Drug Administration has approved four commercial molecular diagnostic test products for clinical trials. Two were shown in Figure 1. A third is AlloMap (Figure 2), a predictor of the acceptance or rejection ratio of a cardiac transplant based on gene expression profiles. The fourth product, which was recently approved by the FDA, is the Tissue of Origin. This test predicts the tissue of origin of a metastatic cancer and may be useful in cases that fail to yield the primary cancer site, despite a thorough checkup. Tissue of Origin analyzes the gene expression profile of a metastatic cancer to predict the differentiation lineage of the cancer, thereby helping the physician to choose the best treatment.

The development of molecular diagnostic technologies is a booming business, and more commercial products are in the pipeline. A few

Figure 2

years ago, a friend at Stanford and I counted the different emerging genomic technologies, and our list included more than 40 different microarrays, consisting of gene expression microarrays, ArrayCGH, micro-RNA microarrays, SNP arrays, tiling arrays, and tissue microarrays and so on. I exaggerate little in saying that a new technology is introduced every two to three months. All of these technologies can be used to transform the way medicine is practiced, especially in integration.

One of the most important technologies in this field is the next generation of nucleotide sequencing, making it possible to sequence personal genomes. The first personal genome sequenced was that of a Caucasian, followed by that of an African, and then that of a South Asian. Finally, a colleague recently completed sequencing the entire personal genomes of a Korean male and a Korean female, representing one of the largest ethnicities in the world, one that covers an area from Turkey to Mongolia, Manchuria, Korea, and Japan. The Korean sequences have been accepted for publication in Nature[1].

In my hospital, Seoul National University Hospital, we decided to follow these successes by developing an institution-wide research information infrastructure that would combine all clinical information and genomic data, correctly integrating the wide variety of clinical phenotypes with all the clinical genotypes. We are truly moving into an era of personal genomics. Our vision and hope are to use all available genomic and proteomic data to make hyper-personalized medicine a reality.

Information Systems at Seoul National University and the Hospital

I will present a brief overview of our information system at Seoul National University Hospital and College of Medicine (Figure 3). Our Legacy system has been in existence since the early 1980s, but we did not complete our electronic medical records system until 2004. This system is for transaction control and has allowed us to dispense with all paper records; however, the system is not linked to clinical or genomic research data. Therefore, we have developed a small clinical record pilot system based on data from the annual checkups of healthy Koreans. As we perform a lot of annual checkups in Korea, I decided to develop a data warehouse

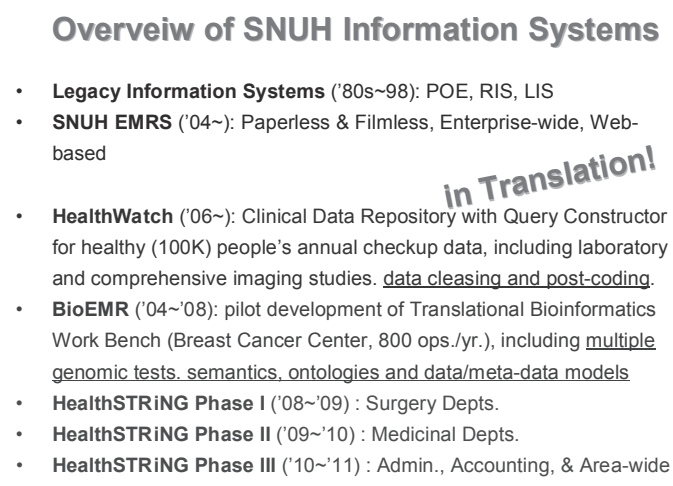

Figure 3

and named it HealthWatch. To date, this data warehouse for patient information gathered at the Gangnam Center in Seoul National University Hospital includes full checkup data sets for about 100,000 very healthy people. HealthWatch, piloted in the year 2006, was developed in about six months after years of planning.

We then moved from healthy people to real hospitalized patients. My aim was to pilot a small clinical genomic integration or translational bioinformatics project in collaboration with the Breast Cancer Center, where 800 breast cancer surgeries are performed annually. Eventually, we decided to expand this project to include the entire enterprise, which comprises a 2,000-inpatient bed main hospital; a 1,000-inpatient bed new hospital; a community-based hospital; and a check-up center that serves as an initial point of contact. We are creating a new data warehouse named HealthSTRiNG. We have completed connections to all the data from all surgical departments, including ophthalmology and otolaryngology, and are beginning to include all medical departments such as internal medicine and pediatrics. We are planning to also include all administrative accounting data in the system as of year 2011.

Among patients treated according to the same therapeutic strategy in the clinical setting, some will show complete remission, while others will experience bad outcomes. This reflects the current limitations to our knowledge of medicine. To begin to move beyond this point, we have piloted a bio-Electronic Medical Recod (BioEMR) system, into which we deposit three more types of data: genomic data, biological data from samples, and clinical data from observations. Although this is not a real clinical trial, we hope to eventually transform the whole system and increase the quality of data, such that a clinical trial would be a byproduct of clinical practice. After we develop and apply intelligent algorithms to these data, we plan to introduce recommendations for further treatment and then monitor outcomes.

Breast Cancer Center

Our Breast Cancer Center is an independent unit that performs 800 surgeries a year, with associated genomic and tissue bank data (Figure 4). After conferring with the director, I piloted a new data system designed to integrate at least three different

Breast Cancer Center, SNUH

- Full-scale EMR at SNUH
- Breast Center Research Information System
- Clinical cases: 800 ~ operations per year
- Administrative independence
- International Clinical Trials
- Stanford DNA microarray Collaboratory
- Clinico-pathologic database : 5,606 patient's data since 1981
- SNP database : 1994 – 2003, 1,114 patiens and 1004 controls
- Tissue bank : 1995. 4- 2004. 1, 582 speicmens
- Cell bank : 1194 cancer and 130 benign breast tumor cell lines

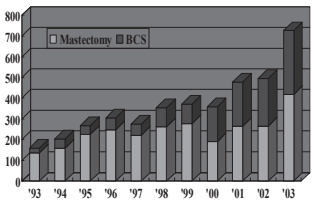

Figure 4

types of genomic technologies: functional genomic data, gene expression arrays, and structural genomic data such as sequences. One interesting central technology to be incorporated was a tissue microarray that includes histopathologies and clinical information. We also needed to include clinical information. After this system is operational, we plan to expand it to encompass all the different types of technologies previously listed.

When we look at gene expression microarrays, we can easily see that they simultaneously involve both a single tissue and tens of thousands of genes. By measuring gene expression in this way, we are essentially parallelizing the measurement of gene expression for a single tissue, which is functional genomics. Similarly, by measuring the single-nucleotide and copy-number polymorphisms throughout the whole genome, we are essentially parallelizing the measure of structural variation throughout the genome for a single tissue.

A tissue microarray (TMA) is a collection of tissues on a slide. To create the tissue collection, cylindrical samples of different tumors or tissues are formed and placed into a block in a lattice pattern. For example, for a 20 × 30 metric, 600 patient samples are placed inside the block. The tissue block can be sliced using a microtome to create thousands of slides containing the same histopathology. All type of in-sit molecular technology can be applied all of the samples and measured under the same conditions. The ability to analyze gene expression in multiple tissue samples with different histopathologies at the same time is a great advance for molecular biologists, who are accustomed to making single measurements, one at a time. Having thousands of slides for the same population allows us to parallelize the measurement and parallelize the pathology per specimen. The gene structures provided by structural genomic technologies and the gene functions provided by functional genomics technologies can be used to develop biomarkers. These high throughput hypothesis-generation technologies require TMA high throughput validation technology

There is an international data standard that encompasses the functional genomics domain like MAGE and MIAPE, but it does not integrate histopathologies, clinical information, and population-wide datasets. Interestingly, TMA is the genomic data that encompasses all these data types.

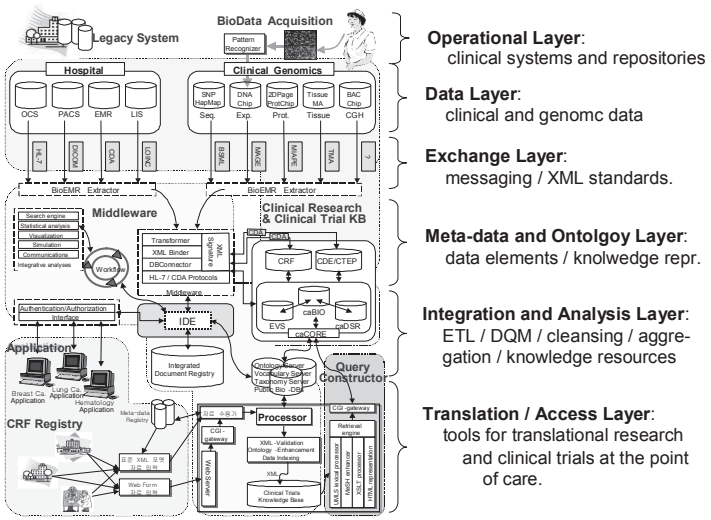

Figure 5

Thus, we chose to integrate all types of information based on international standards, in order to realize our vision of genome-phenome integration. A paper I published in 2006 in the Archives of Pathology and Laboratory Medicine[2] provides an example of a tissue microarray object model (TMA-OM) that encompasses the complex of histopathology, population, and clinical information and devises a genomic cancer subject model and a data model.

Clinical information systems, hospital information systems, and laboratory information systems all have some sort of data exchange format such as HL-7, DICOM, CDA, or LOINC, which are mainly encoded in XML format support for data exchange. In a clinical genomics laboratory, all data, including data for gene sequences, protein sequences, protein expression, and tissues, as well as their exchange centers use the same method (based on XML-based data models). We have developed an extractor and created a secondary repository or data warehouse for all of those elements. In particular, we have added a clinical research and trial knowledge base, an idea we borrowed primarily from the US National Cancer Institute's knowledge management system. We have integrated this knowledge base with the secondary data warehouses, the HealthSTRiNG, from the hospital and clinical genomics laboratories. Furthermore, we have developed a small pilot project for a couple of clinics to test our hypothesis regarding the facilitation of translational research. We have also extracted College of American Pathologist's cancer protocols for more than 46 cancers for entry into our registry, to create an integrated data model. We adhered to the International Standardization Organization (ISO) Standard 11179 when developing a metadata registry[3], and we have repeated the same work for our hospital information system. We have collected major clinical documents for evaluation, extracted all the data elements, and formulated them to adhere to the ISO 11179 standard. Finally, we have mapped all of these data elements with the registry, BMESH (Biomedical Meta-data Element Standard for Health) supported by BioEMR system for clinical trial and research.

One unique feature of this model is that it has not

only a conceptual domain but also a value domain, so one can completely model all forms in clinical research. Thus, we were able to semi-automatically map the terms to control the vocabularies as well as create a dataset by giving some structure to these data elements (Figure 5). We have extracted data elements from many standards and clinical trials and from our clinical documents and registry, all of which have permissible values and database schema levels. Moreover, we have combined them to create a template or research form. We have also created a clinical form, which can be used to manage the schedules of clinical research projects or trials.

Presently, we have an abundance of genomic data and a hospital information system, or electronic medical records system. The approach had been to extract the data into a secondary database; however, our approach now is to model these data elements into an international standard supported by a controlled vocabulary and to develop a form designer and template designer to manage all clinical research schedules, thereby allowing the integration of these two aspects. This provides a blueprint for a real multi-center clinical trial using our BioEMR system.

Conclusions

At the Breast Cancer Center, we have created about 5,000 data elements or clinical concepts. We envision that about 30,000 elements are needed to cover most of the concepts ever used in the Department of General Surgery. We may need about 100,000 data elements to cover most of the clinical practice concepts. Using our BioEMR system, we can also integrate information from the tissue bank and from genetic reporting systems to create a critique or recommendation system for patients. We are moving toward integrating our entire information system with the newly emerging genomic technologies.

References

1. Kim IJ, Ju YS, Park H, et al. A highly annotated whole-genome sequence of a Korean individual. Nature 2009; 460: 1011-1015.
2. Lee HW, Park YR, Sim J, et al. The tissue microarray object model: a data model for storage, analysis, and exchange of tissue microarray experimental data. Arch Pathol Lab Med 2006; 130: 1004-1013.
3. Park YR, Kim JH. Metadata registry and management system based on ISO 11179 for Cancer Clinical Trials Information System. AMIA Annu Symp Proc 2006: 1056.