Currents in Integrative Biochip Informatics

Ju Han Kim, M.D., Ph.D.
Biomedical Informatics,
Seoul National University School of Medicine
E-mail: juhan_kim@harvard.edu

Abstract

Bioinformatics is a rapidly emerging field of biomedical research. A flood of large-scale genomic and postgenomic data means that many of the challenges in biomedical research are now challenges in computational sciences and information technology. The informatics revolutions both in clinical informatics and bioinformatics will change the current paradigm of biomedical sciences and practice of clinical medicine, including diagnostics, therapeutics, and prognostics.

Postgenome informatics, powered by high throughput technologies and genomic-scale databases, is likely to transform our biomedical understanding forever much the same way that biochemistry did a generation ago. In this talk, I will describe how these technologies will impact biomedical research and clinical care, emphasizing recent advances in biochip-based functional genomics. Basic data preprocessing with normalization and filtering, primary pattern analysis, and machine learning algorithms will be presented. Issues of integrated biochip informatics technologies including multivariate data projection, gene-metabolic pathway mapping, automated biomolecular annotation, text mining of factual and literature databases, and integrated management of biomolecular databases will be discussed. Each step will be given with real examples from ongoing research activities in the context of clinical relevance. Issues of linking molecular genotype and clinical phenotype information will be discussed.

CV

현재: 서울의대 생명의료정보학(Biomedical Informatics) 조교수

2001년-2001년: 미국 하버드의대 Biomedical Informatics 조교수

2000년-2001년: 미국 하버드의대 Biomedical Informatics 전임강사

1998년-2001년: M.S. Medical Informatics, 미국 MIT대, Exploration of complex data by unsupervised learning

1997년-2000년: Douglas Porter Fellowship in Medical Informatics, Harvard Medical School

1998년: Ph.D., 서울의대, Brain Image analysis methodology

1995년: M.S., 서울의대, Brain Image analysis methodology

1992년-1996년: 신경정신과전문의, 서울대학교병원

1988년: M.D., 서울의대

1. Clinical Informatics and Bioinformatics

The decade of 1940s brought the first electronic digital computers as well as the first antibiotic, penicillin. Motivated by these revolutionary innovations, a few biomedical researchers started to explore the possible utility of digital computers by the late 1950s. Remarkable use of computers in medical sciences, that are fundamentally information-intensive, was made by the 1960s. The English term *medical informatics* (a translation from the Russian *informatika*) first appeared in 1974 by the need of a name for this domain of new biomedical knowledge and by the lack of a single English term that includes both *information* (what is processed) and *computers* (how it was processed) and encompasses all the fields of *science, engineering,* and *technology*[1].

Bioinformatics, a newly named and rapidly emerging field of biomedical research, has been recognized for about a decade. A flood of large-scale genomic and postgenomic data, powered by high throughput technologies and large-scale databases, means that many of the challenges in biomedical research are now challenges in computational sciences. Not only are many of the fundamental problems in genomics/proteomics (i.e., four-letter and 20-letter-alphabet texts) the classical problems of computer sciences such as sequence homology, pattern recognition, structure prediction, and network analysis, but also are structural, behavioral, and developmental features of living organisms fundamentally *informatical* phenomena.

Biomedical informatics, the convergence of clinical informatics and bioinformatics will radically transform our biomedical understanding forever much the same way that biochemistry did a generation ago. Some schools have already integrated bioinformatics and clinical informatics programs[2][3] that have shared areas of research, core methodologies, challenges, goals, and impact[4][5][6]. As bioinformatics moves from constructing raw biomolecular data to their biological functions and clinical importance, quality clinical information will become the critical part of further progress. Patient's biomolecular information such as personal and familial genetic code will soon be included in his/her electronic medical record as the most predictive clinical information for diagnostics, therapeutics, and prognostics, and threaten the right of patient's privacy and confidentiality. Comprehensive integration of clinical informatics and bioinformatics systems will be one of the primary challenges in the next decades.

2. Accomplishments of Bioinformatics and the Clinical Relevance of Biochip Informatics

The critical dependence of the success of Human Genome Project (HGP) on bioinformatics

constitutes just one example among the remarkable accomplishments of bioinformatics: sequence alignment of DNA and protein, natural genetic variation, prediction of structure and function of biological macromolecules, analysis of biomolecular interaction networks, integration of heterogeneous biological databases, biomolecular knowledge representation, simulation of biological processes, analysis of the data created by large-scale biological experiments, molecular and drug design.

Most researchers agree that the challenge now is to understand all the data. The speed of data generation now exceeds that of interpretation (i.e. more sequences than related publications in GenBank). It becomes even more serious by the introduction of biochips that measure the functional activities of genes and proteins. DNA microarrays are microscopic slides containing a large number of cDNA (or oligonucleotide) samples as fluorescently labeled probes to quantitatively monitor the abundances of transcripts (or mRNA's). Image scanner translates fluorescent intensities into a numerical matrix of expression profiles.

Now that we have comprehensive maps of human genome and transcriptome and that biochip technology can be applied to cells or tissue samples without pulling genes or proteins from them, it is such a fascinating technique to address the comprehensive spatial and temporal genomic complexity in living organisms under different experimental conditions. Biochip informatics with comprehensive expression profiling clearly constitute one of the most straightforward bridge from biomolecular informatics to clinical medicine to improve diagnostics, therapeutics and prognostics.

3. Integrated Biochip Informatics in Functional Genomics/Proteomics

3-1. Biochip informatics: basic data analysis

Because there are many sources of noise and systematic variability in microarray experiments[7][8], data normalization and preprocessing are crucial in the analysis. Normalization includes those transformations that control systematic variabilities within a chip or across multiple chips. The simplest way of data normalization can be done by dividing or subtracting all expression values by a representative value for the system or by a linear transformation to a fixed mean and unit variance. However, the linear response between true expression level and measured fluorescent intensity may not be guaranteed[9], especially when dye biases depend on array spot intensity or when multiple print tips are used in microarray spotter[10].

Data preprocessing includes those transformations that prepare the data for the subsequent analysis. Scaling and filtering are the major steps of data preprocessing. A low-variation

filter to exclude genes that did not change significantly across experiments has been successfully applied in many studies[11]. Statistical significance testing such as analysis of variance and multiple comparisons can also be used to filter the data when enough number of repeated observations are available.

The importance of data visualization cannot be overemphasized. It is highly recommended to scatter plot the data, whenever possible. The most straightforward approach to microarray data analysis is to find differentially expressed genes across different experimental conditions[12][13]. Standardized expression profiling, consistent database design, and streamling experimental process management are all crucial as well as all the following supervised and unsupervised machine learning algorithms to make sense of mountains of genomic data.

3-2. Biochip informatics: functional clustering and machine learning approaches

A general question in many research areas is how to organize observed data into meaningful structures. One very common difficulty in biochip data analysis is the very high dimensionality of the data. Data projection method reduces high dimensionality and projects complex data structure on a lower dimensional space. Cluster analysis, by reducing dimensionality, creates hypothesized clusters and helps researchers to infer unknown functions of genes based on the assumption that a group of genes with similar expression profiles may be functionally associated.

Principal component analysis, a statistical approach to reduce dimensionality without losing significant information by paying attention only to those dimensions that account for large variance in the data, has been applied to microarray data analysis[14]. Mutidimensional scaling, a data projection method originally developed in mathematical psychology, has also been shown to be a powerful tool in functional genomics research[15].

Cluster analysis is currently the most frequently used multivariate technique to analyze microarray data. Clusters can be developed using a variety of similarity or distance metrics: Euclidean distance, correlation coefficients, or mutual information. Hierarchical tree clustering joins similar objects together into successively larger clusters in a bottom-up manner (i.e., from the leaves to the root of the tree), by successively relaxing the threshold of joining objects or sets[16][17]. The relevance networks take the opposite strategy[18]. It starts with a completely connected graph with the vertices representing each object and the edges representing a measure of association and then links are increasingly deleted to reveal 'naturally emerging' clusters at a certain threshold.

Partitional clustering algorithms, such as *K*-means analysis and self-organizing maps (SOM)[19],

which minimize within-cluster scatter or maximize between-cluster scatter were shown to be capable of finding meaningful clusters from functional genomic data[20)21)]. Creation of hierarchical-tree structure in a top-down fashion (i.e., from the root to the leaves of the tree) by successive 'optimal' binary partitioning based on graph theory[22)] and geometric space-partitioning principle[23)] has also been also introduced.

The 'optimal' paritioning problem (i.e., the best clustering) is fundamentally NP-hard and can be viewed as an optimization problem. Most of the meta-heuristic algorithms such as simulated annealing, genetic algorithms[24)], and Tabu search can all be applied to attain better understanding of the complex data structure of genomic-scale expression profiles. Reliability of clusters as well as cluster quality measures for evaluation of clustering solutions should be addressed.

3-3. Integrative biochip informatics

Exploratory data analysis like clustering is appropriate when there is no *a priori* knowledge about the area of research. Such technique is known as unsupervised machine learning in artificial intelligence community. With increasing knowledge of complex biological systems, supervised machine learning techniques (or classification algorithms) are also increasingly introduced to functional genomics resulting significant success[25)26)].

In addition to clustering and classifying (or unsupervised and supervised machine learning) expression profiles, systematic integration and streamlining of appropriate informatics technologies can magnificently enhance the productivity of funcitonal genomics research. For example, PubGene[27)] links gene expression profiles to biomedical literature by combining gene ontology and text mining techniques applied to MEDLINE database. A variety of meta-databases[28)] and natural language processing techniques[29)] are being applied to extract biomolecular interaction networks from biomedical literature and factual databases. Linking these information to genetic regulatory network and metabolic pathway information like KEGG is under vigorous research. Structural sequence information can be used to greatly enhance functional understanding[30)].

We have also developed automatic annotation machines for each microarray probes by integrating many of the publicly available bioinformatics databases. An automated inference engine to predict the functional annotation of genes is working together with all the streamlined biochip informatics technologies including basic data analysis, functional clustering, and supervised classification algorithms. Mangement of integrated database as well as intelligent modules are getting more and more important and challenging. We are heading to integrate these biochip informatics technologies to the advanced clinical

information systems at Seoul National University Hospital.

4. Biomedical Informatics: the New Paradigm for Biomedical Research.

Large areas of medical research and biotechnological development will be permanently transformed by the evolution of high throughput techniques and informatics. Biochip technology is one of the most readily applicable bioinformatics innovations to biomedical research and clinical medicine. It was demonstrated that certain form of cancer can be classified by large-scale gene expression profiling[31]. The capability of new disease class discovery as well as prognostic prediction were also demonstrated[32]. Drug discovery is being transformed by new developments in molecular cell biology and bioinformatics.

This spectacular capability of biochip technology for clinical medicine is no wonder considering that what it essentially does is simultaneously performing tens of thousands of molecular marker studies with comprehensive sets of biologically the most informative molecules, genes and proteins, in a very systematic and quantitative fashion. It uncovers the molecular basis of histopathological processes, the fundamentals of modern diagnostics.

Bioinformatics won't replace experiments, but miniaturization and automation of the laboratory processes can magnificently streamline and enable the discovery process. Integrating quality clinical information is crucial to achieve real improvements in clinical diagnostics, therapeutics, and prognostics. It will, in turn, permanently transform the structure and function of our biomedical knowledge bases.

Weaving the horizontally integrated 'omic' revolution (i.e., genome, transcriptome, proteome, metabolome, and biome) in biomedical sciences with the vertical integration of biomedical informatics (i.e., bio-molecular informatics, computational cell biology[33], computational physiology[34] (ex., neuroinformatics[35]), digital anatomy[36] (i.e., structural informatics), chemoinformatics[37][38], clinical informatics[39], and public health informatics[40]) has now come of age. The new biomedical science will be both molecularly-informed and informatically-empowered.

References

1) Collen M.F. A history of medical informatics in the United States 1950 to 1990. 1995, American Medical Informatics Association

2) Altman RB. The Interactions Between Clinical Informatics and Bioinformatics: A Case Study. J Am Med Inform Assoc 2000;7(5):439-443.

3) Miller PL. Opportunities at the intersection of bioinformatics and health informatics: a case study. J Am Med Inform Assoc. 2000 Sep-Oct;7(5):431-8.

4) Rinfleisch T.C., Brutlag D.L. Directions for clinical research and genomic research into the next decade: implications for informatics. J Am Med Inform Assoc. 1998;5(5):404-411

5) Altman R.B. Bioinformatics in support of molecular medicine. Proc AMIA Symp. 1998;:53-61.

6) Kohane I.S. Bioinformatics and clinical informatics: the imperative to collaborate. Editorial comment. J Am Med Inform Assoc 2000;7(5):512-515

7) Schuchhardt J, Beule D, Malik A, Wolski E, Eickhoff H, Lehrach H, Herzel H. Normalization strategies for cDNA microarrays. Nucleic Acids Res 2000 May 15;28(10):E47

8) Wildsmith SE, Archer GE, Winkley AJ, Lane PW, Bugelski PJ. Maximization of signal derived from cDNA microarrays. Biotechniques 2001 Jan;30(1):202-6, 208

9) Kepler TB, Crosby L, Morgan KT. Normalization and analysis of DNA microarray data by self-consistency and local regression. Functional Genomics IPAM Fall 2000

10) Yang YH, Dudoit S, Luu P, Speed TP. Normalization for cDNA Microarray Data. Tech.report, University of Berkeley, December 2000.

11) Tamayo P, Slonim D et al.: Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. Proc Natl Acad Sci USA, 96:2907-2912, 1999.

12) DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, Chen Y, Su YA, Trent JM. Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat Genet 1996;14(4):457-60

13) Heller RA, Schena M, Chai A, Shalon D, Bedilion T, Gilmore J, Woolley DE, Davis RW. Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. Proc Natl Acad Sci U S A 1997;94(6):2150-5

14) Hilsenbeck S, Friedrichs W, Schiff R, O'Connell P, Hansen R, Osborne C, Fuqua SW. Statistical analysis of array expression data as applied to the problem of Tamoxifen resistance. Journal of the National Cancer Institute. 1999;91(5):453-459

15) Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M et al. Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature 2000;406(3);536-540

16) M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, "Cluster analysis and display of genome-wide expression patterns" Proc Natl Acad Sci U S A 998;95(25):14863-8

17) Iyer VR, Eisen MB, Ross DT, Schuler G, Moore T, Lee JCF, Trent JM, Staudt LM, Hudson J Jr, Boguski MS, Lashkari D, Shalon D, Botstein D, Brown PO. The transcriptional program in the response of human fibroblasts to serum. Science 1999;283(5398):83-7

18) Butte AJ, Kohane IS. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. Pac Symp Biocomput 2000, 418-429.

19) Kohonen T. Self-organized formation of topologically correct feature maps. Biological Cybernetics 1982;43:59-69

20) Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. Nature Genetics 1999;22: 281-285

21) Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander E, Golub TR. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation Proc Natl Acad Sci U S A 1999;96(6):2907-12.

22) Sharan R, Shamir R. CLICK: A clustering algorithm with applications to gene expression analysis. Proc ISMB 2000:307-316.

23) Kim JH, Ohno-Machado L, Kohane IS. Unsupervised learning from complex data: the matrix incision tree algorithm. Pac Symp Biocomput 2001, 30-41.

24) Lee K, Kim JH, Chung TS, Moon BS, Lee H, Kohane IS. Evolution Strategy Applied to Global Optimization of Clusters in Gene Expression Data of DNA Microarrays. Proceedings of IEEE Congress on Evolutionary Computation, Seoul, Korea. May 27-30, 2001:845-850

25) Brown MPS, Grundy WB, Lin D, Christianini N, Sugnet CW, Furgey TS, Haussler D. Knowledge-based analysis of microarray gene expression data by using support vector machines. PNAS 2000;97(1):262-267

26) Heyer LJ, Kruglyak S, Yooseph S. Exploring expression data: Identification and analysis of coexpressed genes.

Genome Research 1999;9:1106-1115
27) Jenssen TK, Laegreid A, Komorowski J, Hovig E. A literature network of human genes for high-throughput analysis of gene expression. Nat Genet. 2001 May;28(1):21-8.
28) Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. Bioinformatics 1998 14: 656-664
29) Park JC, Kim HS, Kim JJ. Bidirectional Incremental Parsing for Automatic Pathway Identification with Combinatory Categorial Grammar. Pac Symp Biocomput 2001;6:396-407.
30) Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. Nature Genetics 22: 281-285
31) Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999 Oct 15;286(5439):531-7.
32) Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J Jr, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Staudt LM, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature. 2000 Feb 3;403(6769):503-11.
33) Tomita M. Whole-cell simulation: a grand challenge of the 21st century. Trends Biotechnol 2001 Jun;19(6):205-10
34) Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science 2001 May 4;292(5518):929-34
35) Chicurel M. Databasing the brain. Nature 2000;406:822-825.
36) Brinkley JF. Structural informatics and its applications in medicine and biology. Academic Medicine 1991;66:589-591
37) Brown FK. Chemoinformatics: What is it and How does it Impact Drug Discovery. Annual Reports in Medicinal Chemistry 1998;33:375-384.
38) Hann M, Green R, Chemoinformatics - A new name for an old problem. Current Opinion in Chemical Biology, 1999;379-383.
39) Degoulet P, Fischi M. Introduction to clinical inforamatics. 1997, Springer, New York.
40) Friede A, Blum HL, McDonald M. Public health informatics: how information-age technology can strengthen public health. Annu Rev Public Health. 1995;16:239-52.