Bioinformatics and genomic medicine

Ju Han Kim, MD, PhD

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 science. Clinical informatics has long developed methodologies to improve biomedical research and clinical care by integrating experimental and clinical information systems. The informatics revolution in both bioinformatics and clinical informatics will eventually change the current practice of 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, in much the same way that biochemistry did a generation ago. This paper describes how these technologies will impact biomedical research and clinical care, emphasizing recent advances in biochip-based functional genomics and proteomics. Basic data preprocessing with normalization and filtering, primary pattern analysis, and machine-learning algorithms are discussed. Use of integrative biochip informatics technologies, including multivariate data projection, gene-metabolic pathway mapping, automated biomolecular annotation, text mining of factual and literature databases, and the integrated management of biomolecular databases, are also discussed. ***Genet Med* 2002:4(6, Supplement):62S–65S.**

Key Words: *bioinformatics, genomic medicine, functional genomics, proteomics, DNA microarray*

Clinical informatics and bioinformatics

The decade of the 1940s brought the first electronic digital computers, as well as the first antibiotic, penicillin. Motivated by these revolutionary innovations, by the late 1950s a few biomedical researchers had started to explore the possible utility of digital computers. By the 1960s, there was extensive use of computers in the medical sciences, which are fundamentally information-intensive. The English term *medical informatics* (a translation from the Russian *informatika*) first appeared in 1974 because of the need for a name for this area of new biomedical knowledge and because of the lack of a single English term that includes both *information* (what is processed) and *computers* (how it is processed). The name also needed to encompass 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. The emergence of modern bioinformatics obtained enormous insight from carefully constructed clinical genetics databases, such as disease-specific mutation databases and genotype-phenotype analyses. 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 science. Not only are many of the fundamental problems in genomics/proteomics, such as string sequence homology, pattern recognition, structure prediction, and network analysis, the problems of computational science, but so also are the structural, behavioral, and developmental features of living organisms fundamentally *informatical* phenomena.

Biomedical informatics, the convergence of bioinformatics and clinical informatics, is radically transforming our biomedical understanding much the same way that biochemistry did a generation ago. Some academic institutions have already integrated bioinformatics and clinical informatics programs that have shared areas of research,[2,3] core methodologies, challenges, goals, and impact.[4–6] As bioinformatics moves from constructing raw biomolecular data into their biological functions and clinical importance, quality clinical information will become the critical part of further progress. A 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 this could threaten the right of privacy and confidentiality. Comprehensive integration of bioinformatics and clinical informatics systems, then, will be one of the primary challenges in the next decades.

Accomplishments of bioinformatics and the clinical relevance of biochip informatics

The critical dependence of the success of the Human Genome Project on bioinformatics is just one example of the remarkable accomplishments of bioinformatics. Other areas where bioinformatics has been crucial include sequence alignment of DNA and protein, natural genetic variation, predic-

From the Children's Hospital Informatics Program, Harvard Medical School, Boston, Massachusetts.

Ju Han Kim, MD, PhD, Children's Hospital Informatics Program, Harvard Medical School, 300 Longwood Avenue, Boston, MA 02115.

Received: July 11, 2002.

Accepted: September 30, 2002.

DOI: 10.1097/01.GIM.0000041505.96252.86

tion of the 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, and rational 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). This has become even more serious with 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 abundance of transcripts (or mRNAs). An image scanner translates fluorescent intensities into a numerical matrix of expression profiles.

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

Integrative biochip informatics in functional genomics and proteomics

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 analysis. Normalization includes those transformations that control systematic variabilities within a chip or across multiple chips. The simplest way data normalization can be done is by dividing or subtracting all expression values by a representative value for the system or by a linear transformation to a fixed mean (i.e., 0.0) and unit variance (i.e., 1.0) (sometimes called "median polishing"). However, the linear response between the true expression level and measured fluorescent intensity may not be guaranteed,[9,10] especially when dye biases depend on array spot intensity or when multiple print tips are used in the microarray spotter.[11]

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.[12] Statistical significance testing, such as the analysis of variance and multiple comparisons, can also be used to filter data that show no significant change across conditions when a sufficient 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.[13,14] Standardized expression profiling, consistent database design, and streamlining the experimental process management are all crucial,[15,16] as are the supervised and unsupervised machine-learning algorithms that make sense of the mountains of genomic data. Here now is a brief description of the various machine-learning approaches to deciphering genomic data.

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 common difficulty in biochip data analysis is the very high dimensionality of the data. The data projection method reduces high dimensionality and projects complex data structure onto a lower dimensional space. Cluster analysis, by reducing dimensionality, creates hypothesized clusters and helps researchers 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.[17,18] Mutidimensional scaling, a data projection method originally developed in mathematical psychology,[19] has also been shown to be a powerful tool in functional genomics research.[20]

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 (Fig. 1).[21,22] The relevance-networks approach takes the opposite strategy.[23] 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,[24] which minimize within-cluster scatter or maximize between-cluster scatter, were shown to be capable of finding meaningful clusters from functional genomic data (Fig. 1).[25,26] Creation of a 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[27] and geometric space-partitioning principle[28] has also been introduced.

The "optimal" partitioning 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 and genetic algorithm[29] and mod-

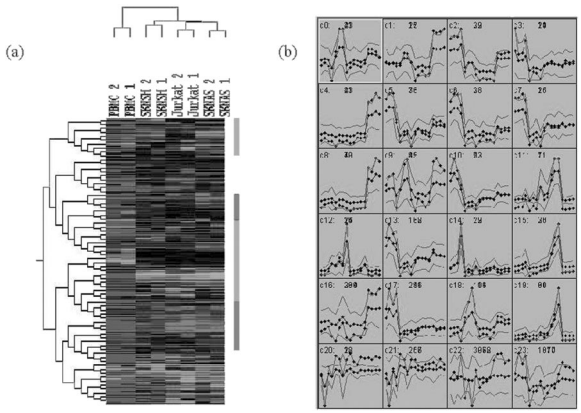

Fig. 1 Cluster analysis and graphical display of genome-wide expression patterns (Jurkat T cells under gamma irradiation). (a) Hierarchical clustering creates functional clusters with color-coded expression patterns. (b) Partitional clusters with geometric grid structure are created by self-organizing maps.

el-based search,[30] can all be applied to attain better understanding of the complex data structure of genomic-scale expression profiles. The reliability and quality measures of clusters, as well as multilevel visualization for the evaluation of clustering solutions, should be addressed.[31,32]

Integrative biochip informatics

Exploratory data analysis, such as clustering, is appropriate when there is no a priori knowledge about the area of research. Such a technique is known as unsupervised machine learning in the artificial intelligence community. With increasing knowledge of complex biological systems, supervised machine-learning techniques (or classification algorithms) are also being increasingly introduced into functional genomics with significant success.[33,34]

In addition to clustering and classifying expression profiles (or unsupervised and supervised machine learning), systematic integration and streamlining of appropriate informatics technologies can greatly enhance the productivity of functional genomics research. For example, PubGene[35] links gene expression profiles to biomedical literature by combining gene ontology and text mining techniques applied to the PubMed database (http:/www.pubgene.org). A variety of meta-databases[36] and natural language processing techniques[37] are being applied to extract biomolecular interaction networks from biomedical literature and factual databases. Linking this information to genetic regulatory network and metabolic pathway information like KEGG is undergoing vigorous research. Structural sequence information can be used to greatly enhance functional understanding.[38,39]

At the Harvard Medical School–affiliated Children's Hospital in Boston, we have also developed automatic annotation machines for each microarray probe by integrating many of the publicly available bioinformatics databases. An automated inference engine to predict the functional annotation of genes works together with all the streamlined biochip informatics technologies, including basic data analysis, functional clustering, and supervised classification algorithms. The management of integrated databases, as well as intelligent modules, is becoming more important and challenging. We are currently integrating these biochip informatics technologies into the advanced clinical information systems at Children's Hospital.

Biomedical informatics: The emergence of new medicine

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 has been demonstrated that certain types of cancer can be classified by large-scale gene expression profiling.[40] The capability of new disease class discovery, as well as prognostic prediction, has also been demonstrated.[41] Drug discovery is being transformed by developments in molecular cell biology and bioinformatics.[42]

The spectacular capability of biochip technology to aid clinical medicine is no wonder considering that, essentially, the technology simultaneously performs tens of thousands of molecular marker studies with comprehensive sets of the biologically most informative molecules, genes, and proteins, in a very systematic and quantitative fashion. By doing so, biochip technology uncovers the molecular basis of histopathological processes, the fundamentals of modern diagnostics.

Bioinformatics will not replace experiments, but miniaturization and automation of laboratory processes can streamline and enable the discovery process to an extraordinary degree. Integrating quality clinical information is crucial to achieve real improvements in clinical diagnostics, therapeutics, and prognostics. Thus bioinformatics is not merely a tool to assist the discovery process; it becomes an integral part of discovery and in this way will permanently transform the structure of our biomedical knowledge bases.

The weaving of the horizontally integrated "omic" revolution of all biological building blocks (genome, transcriptome, proteome, metabolome, and biome) with the vertical integration of biomedical informatics [molecular bioinformatics, computational cell biology,[43] computational physiology[44] (neuroinformatics),[45] digital anatomy[46] (structural informatics), chemoinformatics,[47,48] clinical informatics,[49] and public health informatics[50]] has come of age. The new medicine will be both molecularly informed and informatically empowered.

Acknowledgment

This study was supported by a grant from the Korea Health 21 R&D Project, Ministry of Health & Welfare, Republic of Korea (01-PJ10-PG6-01GM01-0004).

References

1. Collen MF. A history of medical informatics in the United States 1950 to 1990. American Medical Informatics Association, 1995.
2. Altman RB. The interactions between clinical informatics and bioinformatics: a case study. *J Am Med Inform Assoc* 2000;7:439–443.
3. Miller PL. Opportunities at the intersection of bioinformatics and health informatics: a case study. *J Am Med Inform Assoc* 2000;7:431–438.

4. Rinfleisch TC, Brutlag DL. Directions for clinical research and genomic research into the next decade: implications for informatics. *J Am Med Inform Assoc* 1998;5: 404–411.
5. Altman RB. Bioinformatics in support of molecular medicine. *Proc AMIA Symp* 1998:53–61.
6. Kohane IS. Bioinformatics and clinical informatics: the imperative to collaborate [editorial comment]. *J Am Med Inform Assoc* 2000;7:512–515.
7. Schuchhardt J, Beule D, Malik A, Wolski E, Eickhoff H, Lehrach H, Herzel H. Normalization strategies for cDNA microarrays. *Nucleic Acids Res* 2000;28:E47.
8. Wildsmith SE, Archer GE, Winkley AJ, Lane PW, Bugelski PJ. Maximization of signal derived from cDNA microarrays. *Biotechniques* 2001;30:202–206, 208.
9. Kepler TB, Crosby L, Morgan KT. Normalization and analysis of DNA microarray data by self-consistency and local regression. *Genome Biol* 2002;3:RESEARCH0037.
10. Tseng GC, Oh M, Rohlin L, Liao JC, and Wong WH. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variation and assessment of gene effects. *Nucleic Acids Res* 2001;29:2549–2557.
11. Yang YH, Dudoit S, Luu P, Speed TP. Normalization for cDNA microarray data. SPIE BiOS 2001, San Jose, CA, January 2001.
12. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, 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;96:2907–2912.
13. 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;14:457–460.
14. 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;94:2150–2155.
15. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M. Minimum information about a microarray experiment (MIAME): toward standards for microarray data. *Nat Genet* 2001;29: 365–371.
16. Perou CM. Show me the data! *Nat Genet* 2001;29:373.
17. 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. *J Natl Cancer Inst* 1999;91:453–459.
18. Yeung KY, Ruzzo WL. Principal component analysis for clustering gene expression data. *Bioinformatics* 2001;17:763–774.
19. Shepard RN. Multidimensional scaling, tree-fitting, and clustering. *Science* 1980; 210:390–397.
20. Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben-Dor A, Sampas N, Dougherty E, Wang E, Marincola F, Gooden C, Lueders J, Glatfelter A, Pollock P, Carpten J, Gillanders E, Leja D, Dietrich K, Beaudry C, Berens M, Alberts D, Sondak V. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* 2000;406:536–540.
21. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 1998;95:14863–14868.
22. 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; 283:83–87.
23. Butte AJ, Kohane IS. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomput* 2000:418–429.
24. Kohonen T. Self-organized formation of topologically correct feature maps. *Biol Cybern* 1982;43:59–69.
25. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. *Nat Genet* 1999;22:281–285.
26. 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;96:2907–2712.
27. Sharan R, Shamir R. CLICK: a clustering algorithm with applications to gene expression analysis. *Proc Int Conf Intell Syst Mol Biol* 2000;8:307–316.
28. Kim JH, Ohno-Machado L, Kohane IS. Unsupervised learning from complex data: the matrix incision tree algorithm. *Pac Symp Biocomput* 2001:30–41.
29. 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.
30. Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL. Model-based clustering and data transformations for gene expression data. *Bioinformatics* 2001;17:977–987.
31. Yeung KY, Haynor DR, Ruzzo WL. Validating clustering for gene expression data. *Bioinformatics* 2001;17:309–318.
32. Kim JH, Kohane IS, Ohno-Machado L. Visualization and evaluation of clusters for exploratory analysis of gene expression data. J Biomed Inform. In press.
33. 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. *Proc Natl Acad Sci U S A* 2000;97:262–267.
34. Heyer LJ, Kruglyak S, Yooseph S. Exploring expression data: identification and analysis of coexpressed genes. *Genome Res* 1999;9:1106–1115.
35. Jenssen TK, Laegreid A, Komorowski J, Hovig E. A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet* 2001;28:21–28.
36. 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.
37. Park JC, Kim HS, Kim JJ. Bidirectional incremental parsing for automatic pathway identification with combinatory categorial grammar. *Pac Symp Biocomput* 2001: 396–407.
38. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. *Nat Genet* 1999;22:281–285.
39. Zhu Z, Pilpel Y, Church GM. Computational identification of transcription factor binding sites via a transcription-factor-centric clustering (TFCC) algorithm. *J Mol Biol* 2002;318:71–81.
40. 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;286:531–537.
41. 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;403:503–511.
42. Weinstein JN, Myers TG, O'Connor PM, Friend SH, Fornace AJ Jr, Kohn KW, Fojo T, Bates SE, Rubinstein LV, Anderson NL, Buolamwini JK, van Osdol WW, Monks AP, Scudiero DA, Sausville EA, Zaharevitz DW, Bunow B, Viswanadhan VN, Johnson GS, Wittes RE, Paull KD. An information-intensive approach to the molecular pharmacology of cancer. *Science* 1997;275:343–349.
43. Tomita M. Whole-cell simulation: a grand challenge of the 21st century. *Trends Biotechnol* 2001;19:205–210.
44. 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;292:929–934.
45. Chicurel M. Databasing the brain. *Nature* 2000;406:822–825.
46. Brinkley JF. Structural informatics and its applications in medicine and biology. *Acad Med* 1991;66:589–591.
47. Brown FK. Chemoinformatics: what is it and how does it impact drug discovery? *Annu Rep Med Chem* 1998;33:375–384.
48. Hann M, Green R. Chemoinformatics: a new name for an old problem. *Curr Opin Chem Biol* 1999:379–383.
49. Degoulet P, Fischi M. Introduction to clinical informatics. New York: Springer, 1997.
50. Friede A, Blum HL, McDonald M. Public health informatics: how information-age technology can strengthen public health. *Annu Rev Public Health* 1995;16: 239–252.