Published in final edited form as:

IEEE Rev Biomed Eng. 2011 ; 4: 3–5. doi:10.1109/RBME.2011.2177935.

In the Spotlight: Bioinformatics, Computational Biology and Systems Biology

May D. Wang

SECTION I. INTRODUCTION

During the last decade, the Human Genome Project [1], Human Proteome Project [2], Human Physiome Project [3], The International HapMap Project, Genome-Wide Associate Study (GWAS), The 1000 Genomes Project, The Human Microbiome Project, Epigenome, Metabolome, LIPID MAPS, and Next Generation Sequencing have generated large volumes of molecular data about human genes, proteins, physiology, SNPs (single nucleotide polymorphism), metabolites, and lipids etc. This has led to the fast growth of several research sub-areas: 1) high-volume molecular data mining and interpretation; 2) systems biology based on molecular features (i.e., biomarkers) extracted from high-throughput data. Examples include molecular interaction network reconstruction and pathway modeling; 3) synthetic biology, where researchers use engineering approach to make artificial cell and organism; and 4) translational bioinformatics that identifies genomic and proteomic biomarkers of disease to understand molecular mechanisms and guide clinical decision making. This field has grown so rapidly that there are a dozen bioinformatics conferences each year organized by researchers with different backgrounds such as biological or medical sciences, computing or computer science, and engineering.

In 2004, the IEEE Engineering in Medicine and Biology Society (EMBS) started the theme “Bioinformatics and Computational Biology, Systems Biology, Modeling Methodologies.” It started with high throughput omic data analysis, structural bioinformatics, physiological modeling, then moved to systems biology, and recently included synthetic biology, with future movement towards translational bioinformatics in biomedical applications such as cancer prevention, therapeutics, clinical data mining and decision making. In this spotlight, we summarize key topics of importance for the EMBS community.

SECTION II. HIGH-THROUGHPUT DATA ANALYSIS

The landmark publication of microarray gene expression has marked the new era of modern biology and medicine, where advanced high-throughput technologies have been invented and developed to study biology and medicine. Since then, the bio-molecular data volume has exploded. There are several common challenges existing in high-throughput bio-molecular data analysis: data quality control, the “curse-of-dimensionality” (i.e., the number of samples is significantly lower than the number of feature dimensions) and false discoveries

*Corresponding Author: Contact information for the corresponding author: maywang@bme.gatech.edu, Phone: 404-385-2954, Fax: 404-894-4243, Suite 4106, UA Whitaker Building, 313 Ferst Drive, Atlanta, GA 30332, USA.

in molecular feature extractions, the validation and predictability of computational classifiers. Thus, using Microarray as the baseline platform, the FDA-led International MicroArray Quality Control (MAQC) Consortium, which consists of 100+ experts from several dozen institutions around the world, conducted comprehensive investigations on issues such as batch effect, cross-platform difference, biomarker classifier selection, classifier performance metric, prevalence, etc. The FDA MAQC Phase-I has published a series of articles showing that differential gene detection is highly reproducible across a number of microarray platforms [4]. The MAQC Phase-II investigation has led to a Nature Biotechnology Supplement, discussing practice guidelines to address these issues [5].

It is widely believed that a full understanding of human diseases from the genetic perspective depends on understanding variations in normal genetic phenotypes. Thus, another bioinformatics effort is to identify and quantify variations in the genome that explain human diversity (e.g., the HapMap Project and GWAS), which has clinical implications in personalized medicine. The cataloging of human genomic variations involves 1) quickly and efficiently sequencing the genomes of a large number of individuals and 2) mining the data to identify variations such as single nucleotide polymorphisms (SNPs) and insertions and deletions (INDELs), then linking these variations to phenotypic traits. Thus, there is major push for the 1000 Genomes Project to reduce costs. Because protein functions often have a clear effect on phenotype, proteomic acquisition and data analysis methods using mass spectrometry are active bioinformatics research areas for cancer and cardiovascular disease diagnosis and therapy [5]. In recent years, tissue imaging mass spectrometry (TIMS), which acquires multiple mass spectra at different spatial locations of the same tissue specimen, has emerged. Researchers identify and quantify biomarkers that are spatially associated with normal or diseased tissue regions to better understand disease progression and treatment.

As more and more types of high throughput data acquisition techniques become available, there has been a push for integrative bioinformatics that extracts and integrates useful information from heterogeneous types of data and knowledge. Perhaps one of the most ambitious integrative bioinformatics projects is the Virtual Physiological Human (VPH)/Physiome Project [6]. The goal of the VPH Physiome Project is to integrate data from a wide variety of sources to understand the physiological functions of a human body.

SECTION III. SYSTEMS BIOLOGY AND SYNTHETIC BIOLOGY

Integrating data from various technological and biological sources in order to understand a complete biological story has led to the development of systems biology. This is a fast growing interdisciplinary field that focuses on modeling the interactions among components of biological systems [7]. It attempts to uncover the mechanisms under gene, protein, metabolite, and physiological pathways using high-throughput bio-molecular data mining and control/optimization theory-based systems modeling. For instance, to model static network topology, correlation networks, Boolean networks, and Bayesian networks have been used to describe functional interactions among features and patterns extracted from high-throughput gene, protein, and other types of biomolecular data. Correlation networks use correlation coefficients to predict interaction of two genes. The Boolean networks model quantifies expression with ON and OFF states and uses Boolean functions to determine

whether one gene's Boolean state is functionally related to other genes. Bayesian networks capture uncertainty from any contributing factors, such as model parameters, hidden variables, and observations using probabilistic methods.

To model dynamic systems, because of the nature of bio-molecular systems, nonlinear ordinary differential equations (ODEs) are commonly used in bottom-up modeling platform. However, in practice, there are a few issues in using ODEs approaches to model bio-molecular system dynamics: 1) First, nonlinear ODE models require significant *a priori* knowledge about higher-order reaction rate functions that are usually not obvious or intuitive; 2) the limited number of samples makes it difficult for reaction rate parameter estimation. This in turn can lead to the problem of over-fitting (i.e., the estimated parameters are sensitive to data noise and differ from measured values). Thus, based on the observation that robust behavior in biological systems resembles stabilized dynamics in engineering systems, researchers have recently suggested top-down control-theory-based approaches to model bio-molecular systems dynamics in order to achieve stability and robustness. Because in general, the conditions for stability and robustness established using control theory are *sufficient* but not *necessary*, experimental validation is required to prove the uniqueness of top-down models.

True understanding of biology can be demonstrated by creating an artificial bio-system. Thus, synthetic biology was proposed to use “unnatural molecules to reproduce emergent behaviors from nature biology, with the goal of creating artificial life”, or to seek “interchangeable parts from natural biology to assemble into systems that functional unnaturally” [8]. Interesting topics include using analog and digital logic circuitry to program cells, engineering synthetic gene circuits, engineering complex dynamics using synthetic gene networks, and constructing synthetic organisms using large genome engineering, and so on. Specific examples include: creating synthetic metabolic enzyme pathway *in vivo* that mimics natural metabolism [9]; modulating the expression of specific proteins with synthetic ribosome binding sites and selectively introducing mutations to reduce unwanted network interactions (noise); applying synthetic biology at the level of cellular systems or artificial organisms, such as developing multiplex automated genome engineering (MAGE) for large-scale programming and cell evolution by Wang *et al.* [10]. In practice, how to characterize, predict, and reduce crosstalk between host and synthetic circuitry, remains an engineering challenge for creating new biological components (macromolecules), devices (networks), and systems (organisms).

SECTION IV. TRANSLATIONAL BIOINFORMATICS

Clinical applications demand technologies and biological information that can most effectively assist in patient care. Thus, there are a few challenges associated with translating bioinformatics discoveries to medical practice. First, the rate of genomic data generation such as NGS is expected to continue to outpace the rate of data analysis [9]. Because clinic genomic testing needs to avoid over-use (i.e., excessive and expensive medical procedures) and under-use (i.e., medical decisions based on inadequate information), it needs to be carefully filtered, and integrated with other biomedical data. Some opinions in the community have asserted that genomic biomarkers cannot be used independently of standard

clinical factors [11]. Second, genomic information alone may not provide all the information required to understand diseases due to post-transcriptional and post-translational modifications. Because proteins are responsible for biological function, direct applications of biomedical research require quantification and manipulation of proteins. However, at this moment, high-throughput mass spec based proteomics still have multiple unsolved analysis and experimental issues. Currently, some low throughput multiplex proteomic methods such as Quantum Dots (QD) in nanotechnology have proposed to do molecular imaging and perform targeted therapeutics to speed up the proteomics testing. Third, besides technical challenges, advances in genomic sequencing for clinical applications also have ethical and social issues.

Both next-generation sequencing and nanotechnology are at the forefront of personalized medicine. Thus, bioinformatics in the near future will see innovative methods for understanding and processing NGS data. It will also inspire new applications for new technologies in disease diagnosis and treatment.

In summary, the IEEE EMBS “Bioinformatics and Computational Biology, Systems Biology, Modeling Methodologies” is a fast growing theme, with huge potential for pursuing fundamental research in data mining, modeling centric systems biology, and design-centric synthetic biology. In addition, its translation to cancer prevention, therapeutics, and clinical decision making can lead to major societal impact. It is a corner stone for personalized health.

Acknowledgments

The author has been supported by grants from National Institutes of Health (NIH) (R01CA108468, 1RC2CA148265), Georgia Cancer Coalition (Distinguished Cancer Scholar Award to MDW).

M. D. Wang would like to thank J. H. Phan, G. Alterovitz, C. F. Quo, and C. Kaddi for contributions.

References

1. Kidd JM. Mapping and sequencing of structural variation from eight human genomes. *Nature*. 2008; 453(7191):56–64. [PubMed: 18451855]
2. Rabilloud T, Hochstrasser D, Simpson RJ. Is a gene-centric human proteome project the best way for proteomics to serve biology? *Proteomics*. 2010; 10(17):3067–3072. [PubMed: 20648483]
3. Hunter PJ, Borg TK. Integration from proteins to organs: The physiome project. *Nature Rev Molecular Cell Biol*. 2003; 4(3):237–243.
4. Shi L. The MicroArray Quality Control (MAQC) project shows inter-and intraplatform reproducibility of gene expression measurements. *Nature biotechnol*. 2006; 24(9):1151–1161. [PubMed: 16964229]
5. Moxon JV. Challenges, current status and future perspectives of proteomics in improving understanding, diagnosis and treatment of vascular disease. *Eur J Vascular and Endovascular Surg*. 2009; 38(3):346–355.
6. Hunter P, Viceconti M. The VPH-physiome project: Standards and tools for multiscale modeling in clinical applications. *IEEE Rev in Biomed Eng*. 2009; 2:40–53.
7. Macilwain C. Systems biology: Evolving into the mainstream. *Cell*. 2011; 144(6):839–841. [PubMed: 21414474]
8. Sismour SABAAM. Synthetic biology. *Nature Reviews Genetics*. 2005; 6:11.

9. Dueber JE. Synthetic protein scaffolds provide modular control over metabolic flux. *Nature Biotechnol.* 2009; 27(8):753–759. [PubMed: 19648908]
10. Wang HH. Programming cells by multiplex genome engineering and accelerated evolution. *Nature.* 2009; 460(7257):894–898. [PubMed: 19633652]
11. Roukos D. Twenty-one-gene assay: Challenges and promises in translating personal genomics and whole-genome scans into personalized treatment of breast cancer. *J Clinical Oncol.* 2009; 27(8): 1337. [PubMed: 19188673]