Chinese Journal of Cancer

Editorial

Cancer bioinformatics: detection of chromatin states, SNP-containing motifs, and functional enrichment modules

Xiaobo Zhou

Abstract

In this editorial preface, I briefly review cancer bioinformatics and introduce the four articles in this special issue highlighting important applications of the field: detection of chromatin states; detection of SNP-containing motifs and association with transcription factor-binding sites; improvements in functional enrichment modules; and gene association studies on aging and cancer. We expect this issue to provide bioinformatics scientists, cancer biologists, and clinical doctors with a better understanding of how cancer bioinformatics can be used to identify candidate biomarkers and targets and to conduct functional analysis.

Key words Chromatin sates, SNP-containing motifs, functional enrichment analysis, gene association

Cancer bioinformatics is a field that has recently emerged from traditional bioinformatics and clinical informatics. The general purpose of cancer bioinformatics is to discover biomarkers, therapeutic targets, novel pathways, and relationships among diseases or individuals, and to identify candidate causal factors of cancer initiation or progression. Topics in cancer bioinformatics include genomics, proteomics and metabolic data analysis, image data analysis, survival data analysis, data integration, and database building. Genomics data can be collected from arrays at the DNA and RNA levels, including single nucleotide polymorphism (SNP) array, gene array, microRNA array, axon array, tile array, and many others, and from association studies between any two types of array data. Genomics data can also include DNA and transcriptome sequencing data for the human genome. Proteomics data include measures at total protein expression level, such as mass spectrometry data, and measures of protein activity, such as phosphoproteomic data. Imaging data include data from microscopic cellular imaging, three-dimensional (3D) confocal microscopy, flow cytometry, 3D computed tomography (CT), magnetic resonance imaging (MRI), diffusion tensor imaging (DTI), and ultrasound. Survival data include information on patient survival, risk analysis, and stratification. Each type of data requires distinct bioinformatic technologies, for example, hierarchical clustering, K-means, and Fuzzy c-means for disease or individual clustering and correlation analyses, Bayesian network inference for signaling network studies, classification of patients and diseases, eQTL mapping, and others.

Biomedical research is becoming an information science. However, there are many challenges to developing advanced bioinformatics approaches. First, biological questions and events must be understood in depth, a challenging task for bioinformatics scientists who were not trained in biology. Second, bioinformatics scientists develop sophisticated algorithms to solve a biological question, and biologists must understand how and why the algorithms can answer the question. Thus, bioinformatics scientists and biologists must "speak each other's language" and communicate effectively. Third, one major obstacle in modeling biological data is the limited sample size in almost all applications, forcing the use of power analysis in each project. However, power analysis can only be performed based on preliminary data. This again highlights the need for seamless collaboration between bioinformatics scientists and biologists.

In this April issue of *Chinese Journal of Cancer*, four articles from the international leaders in cancer bioinformatics are published to describe their works with state-of-the-art technologies and methodologies and to

Author's Affiliation: Center for Bioinformatics and Systems Biology, Wake Forest University-School of Medicine, Winston-Salem, NC 27103, USA.

Corresponding Author: Xiaobo Zhou, Center for Bioinformatics and Systems Biology, Wake Forest University-School of Medicine, Winston-Salem, NC 27103, USA. Email: xzhou@tmhs.org.

doi: 10.5732/cjc.013.10045

share their insights on the trends and outlook for the field. Wang *et al.*[1] reports the detection and characterization of regulatory elements using probabilistic conditional random field and hidden Markov models. These models, which were developed to mark gene regions with different states based on recurrent and spatially coherent characters, revealed chromatin states that may correspond to enhancers and promoters, transcribed regions, transcriptional elongation, and low-signal regions. Fan *et al.*[2] identified SNP-containing regulatory motifs in the myelodysplastic syndromes model using data from SNP and gene expression arrays. More specifically, they developed an approach to associate SNP-containing motifs introduced by genetic variation and mutation with transcription factor-binding sites. A set of SNP-containing motifs were discovered as candidate transcription factor-binding sites when using regression model-based LARS-EN algorithm. The identified motifs and their location genes can be considered potential biomarkers. In the third article, Huang *et al.*[3] introduce GOMA, a functional enrichment tool based on GO modules. Unlike many pathway enrichment tools, which output a long list of significantly enriched terms that are often redundant, GOMA systematically reveals GO modules based on an optimization model. This method simplifies enrichment analysis results and provides more biologically meaningful results. Finally, Wang[4] reviews the association studies on aging and cancer, of which there is currently limited knowledge. Using gene expression profiling data for aging cells, stem cells, and cancer cells, Wang determined that genes related to aging or to human embryonic stem cells are often highly expressed in cancer. Notably, the product of these genes are involved in critical cancer processes, including cell cycle regulation, metabolism, DNA damage response, apoptosis, p53 signaling, immune/inflammatory response, and others.

Cancer bioinformatics is becoming well integrated into biomedical research. Indeed, the National Cancer Institute (NCI) issued a call for proposals for cancer bioinformatics technology[5]. Some common themes include automation in data collection, data processing and analysis, data quality assessment, data integration, data presentation and visualization, text mining and natural language processing, archiving, organization, retrieval, and dissemination of data and knowledge, and environment for interactive modeling and simulation. Data integration and modeling is an interesting topic to many cancer researchers. For example, there are some discussions[6] on the integration of array data (e.g., SNP array, gene expression array, microRNA array, axon array, and tile array). Although it is beyond the scope of this special issue, we also recommend reference [7], which presents the advantages and disadvantages of different platforms used for image analyses—another area of interest in the field. Ultimately, we expect that advancements in cancer bioinformatics will contribute significantly to our understanding of the complex mechanisms of cancer and lead to better therapeutic strategies.

Received: 2013-03-08; revised: 2013-03-16;
accepted: 2013-03-16.

References

[1] Wang HY, Zhou XB. Detection and characterization of regulatory elements using probabilistic conditional random field and hidden Markov models. Chin J Cancer, 2013,32:186–194.
[2] Fan J, Dy JG, Chang CC, et al. Identification of SNP-containing regulatory motifs in the myelodysplastic syndromes model using SNP arrays and gene expression arrays. Chin J Cancer, 2013,32:170–185.
[3] Huang Q, Wu LY, Wang Y, et al. GOMA: functional enrichment analysis tool based on GO modules. Chin J Cancer, 2013,32: 195–204..
[4] Wang X. Discovery of molecular associations among aging, stem cells, and cancer based on gene expression profiling. Chin J Cancer, 2013,32:155–161.
[5] http://itcr.nci.nih.gov/
[6] modENCODE Consortium. Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science, 2010,330:1787–1797.
[7] Zhou X, Wong STC. A primer on image informatics of high content screening. Haney SA, ed. High Content Screening, Wiley, 2007.