Editorial

Bioinformatics Algorithms and Genomics

Feng Cheng[1,2,*]

[1]Department of Pharmaceutical Sciences, College of Pharmacy, [2]Department of Epidemiology and Biostatistics, College of Public Health, University of South Florida, 12901 Bruce B. Downs Blvd., MDC 30 Room 3102B, Tampa, FL, 33612-4749

The Human Genome Project was completed in April 2003 [1]. All of the 3.2 billion base pairs in the entire human genome have been sequenced in this project. The Human Genome Project is a revolutionizing biological and medical research [2]. However, it was just the starting point for us to understand the function of our genome. Extraordinary amount of information generated from the Human Genome Project needed to be analyzed accurately and efficiently. Also several powerful genome-wide detection tools such as genotyping micro arrays and tiling microarrays have been developed based on genetic information from the Human Genome Project. Nowadays, the analysis of huge amount of data from these high throughput tools is still very challenging [3].

Bioinformatics, an interdisciplinary field that involves biology, computer science mathematics and statistics, provides various algorithms or programs to collect, analyze, store and integrate genetic information from the Human Genome Project and other genome-wide studies. Development of algorithms that enable efficient analysis of different types of information is one of the main goals of bioinformatics. Many algorithms were widely used for genome research including prediction of gene function, identification of genomic regions involved in gene regulation, dissection of molecular evolution, and modeling of alternative splicing. This special issue provides a snapshot of the applications of some bioinformatics algorithms in genome research and some genome-wide studies including the genome-wide association studies (GWAS) [4] and the next generation sequencing (NGS).

This special issue starts with a review by Drs. Gengxin Li and Hongjian Zhu, which gives an introduction to the linear mixed models dealing with sample structure in GWAS. GWAS could identify the genetic variants such as single nucleotide polymorphism (SNP) contributing to complex traits or diseases in a genome-wide scale. GWAS has the potential to find novel disease-related genes because it investigates all human genes in the whole genome instead of specific genes or chromosome regions [4]. GWAS have discovered numerous genetic variants contributing to many human diseases successfully. However, there are some false positives caused by the sample structure which is the presence of related individual within study samples [5]. Sample structure includes population structure, family structure, and cryptic relatedness. The authors reviewed current models dealing with sample structure such as principal components analysis (PCA), multidimensional scaling (MDS) and structured association (SA), and especially focused on the linear mixed model (LMM) based approaches in genome-wide association studies and their performance. They also discussed some unsolved issues and future work of these LMM-based approaches.

In the second paper in this special issue, Dr. Hui Zhang from the St. Jude Children's Hospital and his colleagues reviewed statistical challenge for over-dispersion in mRNA sequencing (mRNA-seq) count data. mRNA-seq is a high throughput technique that could rapidly and economically sequence the entire transcriptome [6]. mRNA-seq has been widely applied to exam gene expression, discover novel gene and identify gene variants. mRNA-seq generate tens of millions of short segments (reads) that could be mapped back to the human genome. The counts of reads mapped to interested gene can be used for digital gene expression measurement [7]. Log-linear models based on Poisson distributions are widely used for analyzing count data. However, theses models often results in a bias and misleading conclusions because the variance of the mRNA-seq count data is often much larger than its mean (referred as over-dispersion) [8]. When over-dispersion occurs, the Poisson distribution does not hold. The authors review some approaches for detecting and modeling over-dispersion, including Negative Binomial, Quasi-likelihood Poisson and the newly developed two-stage adaptive methods. This review will draw to the researchers some more attention on over-dispersed mRNA-seq count data and help researchers to analyze these data more appropriately.

An important topic in bioinformatics is to assign taxonomy and predict function based on genetic sequences. In this special issue, Dr. Georgiou and Dr. Karakasidis along with their collaborator Dr. A. Megaritis introduced Chou's pseudo amino acid composition approach and fuzzy set theory for genetic sequence analysis. The Chou's pseudo amino acid composition approach was initially introduced by

*Address correspondence to this author at the Department of Epidemiology and Biostatistics, College of Public Health, University of South Florida, 12901 Bruce B. Downs Blvd., MDC 30 Room 3102B, Tampa, FL, 33612-4749; Tel: 813-974-4288; Fax: 813-905-9890; E-mail: fcheng1@health.usf.edu

Prof. KuoChen Chou in 2001 [9] to predict attributes of proteins based on their amino acid sequences. The main idea of this method is to transform the protein sequence to a limited size numeric vector based on some important properties such as side chain mass, hydrophobicity and hydrophilicity. Recently, the Chou's pseudo amino acid composition approach has been extended to the analysis of genetic sequence analysis [10]. The authors reviewed methodologies developed based on this concept to transform long polynucleotides to representations in a lower dimension space. They showed the combination of the Chou's pseudo amino acid composition approach with elements of fuzzy set theory is promising for genetic sequence analysis in some studies.

Artificial Neural Networks (ANNs) technology could be also used for genetic sequence analysis. ANNs technology simulates the learning processes of neurons in the brain [11]. Artificial neuron unit in ANNs receives inputs from many external sources, processes them, and makes decisions (predictions). The approach is widely used to simulate various non-linear systems. In this special issue, Dr. Sutariya reviewed ANNs technology and its application in drug delivery and pharmaceutical research.

This special issue provides a "snapshot" of current activities in bioinformatics algorithm development and implement in genome studies. Currently, most of bioinformatics algorithms are developed by mathematicians, statisticians or computational biologists. However, a lot of algorithms are not easy enough for the majority of the biologists to accept and use, but the biologists are still using the most basic bioinformatical tools. It is one of the major limitations of bioinformatics today. Further progress in bioinformatics algorithms will be catalyzed by the development of online databases and analysis servers. These convenient online tools will help biologist to apply these algorithms to design experiments and analyze data accurately and efficiently.

REFERENCE

[1] C. International Human Genome Sequencing, "Finishing the euchromatic sequence of the human genome," *Nature,* vol. 431, no. 7011, pp. 931-45, 2004.

[2] E. S. Lander, "Initial impact of the sequencing of the human genome," *Nature,* vol. 470, no. 7333, pp. 187-97, 2011.

[3] F. Cheng, S. H. Cho, and J. K. Lee, "Multi-gene expression-based statistical approaches to predicting patients' clinical outcomes and responses," *Methods Mol. Biol.,* vol. 620, pp. 471-84, 2010.

[4] K. Ashmore, and F. Cheng, "Genome-Wide association studies on attention deficit hyperactivity disorder," *Clin. Exp. Pharmacol.,* vol. 3, pp. 119, 2013.

[5] A. L. Price, N. A. Zaitlen, D. Reich, and N. Patterson, "New approaches to population stratification in genome-wide association studies," *Nat. Rev. Genet.,* vol. 11, no. 7, pp. 459-63, 2010.

[6] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nat. Rev. Genet.,* vol. 10, no. 1, pp. 57-63, 2009.

[7] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq," *Nat. Methods,* vol. 5, no. 7, pp. 621-8, 2008.

[8] J. H. Bullard, E. Purdom, K. D. Hansen, and S. Dudoit, "Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments," *BMC. Bioinformatics,* vol. 11, pp. 94, 2010.

[9] K. C. Chou, "Prediction of protein cellular attributes using pseudo-amino acid composition," *Proteins,* vol. 43, no. 3, pp. 246-55, 2001.

[10] W. Chen, P. M. Feng, H. Lin, and K. C. Chou, "iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition," *Nucleic Acids Res.,* vol. 41, no. 6, p. e68, 2013.

[11] F. Cheng, and V. Sutariya, "Applications of artificial neural network modeling in drug discovery," *Clin. Exp. Pharmacol.,* vol. 2, p. e113, 2012.