

# IEEE 7<sup>th</sup> BIBE Invited Plenary Keynote: Statistical Analysis of nucleosome occupancy and histone modification data\*

Jun S Liu

Harvard University  
Boston, MA 02115

[jliu@stat.harvard.edu](mailto:jliu@stat.harvard.edu)

\*This work was jointly with Guocheng Yuan

October 15, 2007

Harvard Medical School Conference Center Amphitheater, Boston, MA

**Abstract**— In eukaryotic cells, genomic DNAs wrap around bead-like molecules, called nucleosomes, so as to pack more compactly in the nucleus of the cell. The nucleosome is made up of four pairs of histone proteins (H2A, H2B, H3, and H4) who share a very similar structural motif. The positioning of nucleosomes as well as the modifications of various positions of the histone proteins (such as acetylation, methylation, etc.) play important but incompletely understood roles in gene regulation. Here we describe some statistical models we developed for predicting nucleosome positioning and histone modification patterns using only genomic sequence information.

The regulation of DNA accessibility through nucleosome positioning is important for transcription control. Computation models have been developed to predict genome-wide nucleosome positions from DNA sequences, but these models consider only nucleosome sequences, which may have limited their power. We developed a statistical multi-resolution approach to identify a sequence signature, called the N-score, that distinguishes nucleosome binding DNA from non-nucleosome DNA. The N-score is not sensitive to deletion of short DNA elements and can also be estimated reasonably accurately from coarse nucleosome positioning data. We found that the sequence information is highly predictive for local nucleosome enrichment or depletion, whereas the exact positions may be further fine-tuned by other regulatory factors. We observed that many characteristics of nucleosome positioning, such as the enrichment of transcription binding sites in nucleosome-depleted regions and nucleosome depletion in the promoter regions, can be predicted accurately by the sequence information through N-scores. More surprisingly, the N-scores estimated from the yeast data can also predict the nucleosome binding patterns in human.

In addition to nucleosome positioning, histone modifications, particularly acetylations, are also important in directing gene regulation. A comprehensive understanding of the regulatory role of histone acetylation is difficult because many different histone acetylation patterns exist and their effects are confounded by other factors, such as the transcription factor binding sequence motif information and nucleosome occupancy. We analyzed recent genomewide histone acetylation data using a few

complementary statistical models and tested the validity of a cumulative model in approximating the global regulatory effect of histone acetylation. Confounding effects due to transcription factor binding sequence information were estimated by using two independent motif-based algorithms followed by a variable selection method.

Our analysis confirms that histone acetylation has a significant effect on gene transcription rates in addition to that attributable to upstream sequence motifs, although the upstream sequence information is similarly significant. Our model fits well with observed genome-wide data. Strikingly, including more complicated combinatorial effects does not improve the model's performance, suggesting that a cumulative effect model for global histone acetylation is appropriate. Through a statistical analysis of conditional independence, we found that H4 acetylation may not have significant direct impact on global gene expression.



Jun Liu received a BS degree in mathematics in 1985 from Peking University, Beijing, China, and a Ph.D. in statistics in 1991 from the University of Chicago, USA. He is currently Professor of Statistics at Harvard University, with a courtesy Professor appointment at Harvard Biostatistics Department. He is also Guest Professor of Statistics and Mathematics at Peking and Tsinghua Universities of China. Before that, he held Assistant, Associate, and full professor positions at Stanford

University from 1994 to 2003. Dr. Liu was the recipient of the 2002 COPSS Presidents' Award (given annually by five leading statistical associations to one individual under age 40), the recipient of the Mitchell Prize from the International Society of Bayesian Analysis in 2000, and one of the recipients of the CAREER Award from the National Science Foundation in 1995. He was selected as a Terman Fellow by Stanford University in 1995, as a Medallion Lecturer by the Institute of Mathematical Statistics (IMS) in 2002, and as a Bernoulli Lecturer by the International Bernoulli Society in 2004. He was elected to Fellow of the Institute of Mathematical Statistics (IMS) in 2004 and Fellow of the American Statistical Association in 2005. He served as a council member of the IMS from 2005-2008. Dr. Liu's research interests include bioinformatics and computational biology, statistical genetics and genetic epidemiology, stochastic dynamic systems, signal processing and wireless communication, Bayesian modeling, and Monte Carlo methods. His current publication list includes more than 100 research articles appeared in peer-reviewed journals, a research monograph, and more than 20 book chapters, discussions, and conference proceedings. Dr. Liu is well-known for his theoretical and methodological work in Monte Carlo computational and statistical methods. He is also responsible for developing several popular algorithms for biological sequence analysis, regulatory motif discovery, and population genetics data analysis. Dr. Liu has served on numerous editorial boards of leading statistical journals and has frequented grant review panels of the NSF, and is the member of an NIH study section.