Genome Biology

MEETING REPORT

Conquering computational challenges of omics data and post-ENCODE paradigms

Yves A Lussier[1,2,3,*], Haiquan Li[1,3] and Mark Maienschein-Cline[3]

Abstract

A report on the 21st Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) and 12th European Conference on Computational Biology (ECCB), held in Berlin, Germany, July 21-23, 2013.

Keywords: Epigenetic network, machine learning, next-generation sequencing, post-transcriptional modification, post-translational modification, regulation, statistic modeling, translational bioinformatics.

The past few years have witnessed an unprecedented development of high-throughput technologies, particularly in next-generation sequencing, that measure nearly every molecule of life and its modifications. As observed by the keynote speaker Lior Pachter (University of California, Berkeley, USA), these include SNPs, DNA methylation, open chromatin, all forms of RNAs, proteins, protein-DNA interactions and microRNA-mRNA interactions. Furthermore, high-throughput measurements have recently reached a new zenith with the publication of the ENCODE and TCGA projects, which have significantly increased the capacity of biology and computation to interrogate interactions across different types of molecules of life. In expanding our understanding of mechanisms in biological systems, especially in previously unstudied non-coding regions of DNA, these developments have also inspired new challenges for accurate and effective analysis as we model across multiple scales and ever more data. The 2013 ISMB-ECCB joint conference provided a platform where pioneering solutions to these problems were presented. Here, we highlight representative works in several areas of computational biology: new algorithms for analyzing high-throughput data, regulatory network modeling, translational bioinformatics and methodologies for post-ENCODE studies.

High-throughput methodology for big data

Novel foundational methods address theoretical or empirical challenges associated with deep-sequencing biotechnologies. For example, the annual trend for omics data storage and analysis follows a geometric curve that far outpaces that of the famous Moore's Law for computations, paradoxically yielding a net reduction of omics analytical capacity. Michael Baym (Harvard Medical School, USA) described improvements in sequencing capacity achieved by pioneering 'compressive genomics', which leverages a meta-alignment approach that does not require decompressing redundant consensus sequences and has accelerated search efficiency by an order of magnitude (100% positive predictive value, 99% recall, compared with BLAST). Andrew Smith (University of Southern California, USA) provided an accurate estimation of the maximum number of distinct reads that can be obtained from a DNA library, given the read frequency distribution in limited preliminary sequencing. The estimation matched well with deep-read sequencing distributions from human and chimpanzee samples and can be extended to related biotechnologies such as ChIP-seq. Yaron Orenstein (Tel-Aviv University, Israel) transformed the problem of designing double-stranded DNA probes for protein-binding arrays into a problem of sequence coverage that affords unbiased measurements and more-comprehensive protein binding assessments.

Tools for efficient and accurate analysis of newly generated high-throughput data are continuing to be developed. Henry CM Leung (The University of Hong Kong, China) described a *de novo* RNA-seq assembler, IDBA-tran, designed to remove assembly paths of de Bruijn graphs associated with sequencing errors and for merging paths caused by polymorphisms. The algorithm achieved both better sensitivity (more than 10%) and specificity (more than 5%) in read-data associated with poorly expressed isoforms. Wei Wang (University of California, Los Angeles, USA) presented a post-assembler tool using support vector machines, GeneScissors, to correct misaligned fragments in RNA-seq caused by

*Correspondence: ylussier@uic.edu
[1]Department of Medicine, University of Illinois at Chicago, Chicago, IL 60612, USA
Full list of author information is available at the end of the article

sequence repeats and errors (more than 88% precision and recall; 54% fewer pseudogenes than from the TopHat/Cufflinks pipeline). Bo Xie (Georgia Institute of Technology, USA) reported on an approach integrating a hidden Markov model and a support-vector machine (SVM) for poly(A)-motif discovery that improved the average error rate, false-negative rate and false-positive rate by 26%, 15% and 35%, respectively, compared with state-of-the-art approaches. Jinbo Xu (Toyota Technological Institute at Chicago, USA) presented PhyCamp, an algorithm for improving the accuracy of protein-structure self-contact map predictions from primary sequences; the algorithm exploits various physical constraints in protein structure in evolutionary biology using integer programming after Random Forest modeling. Wyatt Clark (Indiana University, USA) improved our understanding of 'misinformation' and 'remaining uncertainty' information-theoretic measures for protein function prediction methods using gene ontology.

Regulation and molecular networks

Scaling from individual molecular interactions to the underlying regulatory network is a key step in understanding and predicting cellular behavior. In several presentations, integration of multiple sources of data for a single biological story yielded a richer biological story about the regulatory connections involved. Christopher Ng (MIT, USA) demonstrated the mis-regulation of a number of genes in Huntington's disease in relation to DNA methylation. He identified transcriptional regulators, through motif analysis and ChIP-seq, whose binding sites are associated with methylation changes, suggesting new disease mechanisms. Anthony Gitter (Carnegie Mellon University, USA) modeled signaling and regulatory networks using molecular pathways with transcriptional networks to predict the response of interacting genes to invading pathogens and disease progression. Following the same theme of improving inference by combining multiple data types, Andrew Sedgewick (University of Pittsburgh, USA) merged genomic and mRNA expression data from TCGA to augment literature-curated gene and protein interactions.

Automated network reconstruction allows for rapid analysis of multiple species data, and verification that inferences are not incomplete is key to obtaining high-confidence predictions. John Pinney (Imperial College London, UK) used network topologies to predict the validity of inferred enzyme annotations, demonstrating that topological descriptors in metabolic networks can provide a useful test of such automated methods. Masaaki Kotera (Kyoto University, Japan) trained a support vector machine (SVM) classifier on known pairs of compounds linked by enzymatic reactions to predict additional potential reactions, allowing him to construct metabolic pathways *de novo*. Timothy Tickle (Harvard School of Public Heath, USA) inferred metagenomic phylogenies by shotgun sequencing of 16S rRNA to characterize the differential populations of digestive-tract microbiota and relate these changes to inflammatory bowel disease, treatments and the environment (such as smoking and antibiotics). Analyzing many omics data sources coincidently will be a valuable strategy in future research as this approach mitigates the (often significant) noise present in individual experiments and platforms and allows true signals to reinforce each other.

Translational bioinformatics

The progression from molecular interactions to organism-level phenotypes remains an area of exciting development, with many open avenues of research. New insights into stem cell differentiation were presented by Nicola Bonzanni (VU University Amsterdam, The Netherlands), who used a Boolean *cis*-regulatory network model to study the differentiation of hematopoietic stem cells. His model illustrated the emergence of a number of intermediate and heterogeneous cell states before termination of the stem cell state and identified the need for external triggers to complete the exit from the stem cell state. Importantly, a key prediction from the model - the negative regulation of the transcription factor Fli1 by Gata1 - was confirmed experimentally, demonstrating the power of such models to generate experimentally testable predictions about transcriptional control. In another approach, the ATARiS method by Aviad Tsherniak (Broad Institute, USA) examines patterns of gene expression across many RNA interference data sets to link genes quantitatively to suppression phenotypes. This systematic approach alleviates common issues of differential and off-target gene suppression in RNA interference studies.

As we gain an understanding of the molecular interactions involved in a phenotype, understanding how to manipulate the biology becomes an important next step. For example, Paula Petrone (Hoffmann-La Roche, Switzerland) presented an alternative approach for screening interactions between compounds and the proteome to generate candidate pharmaceuticals automatically. Her method bypasses typical homology approaches in favor of biological functional similarity, allowing the discovery of compounds with different structural properties yet similar biological effects. Philippe Sanseau (GlaxoSmithKline, UK) used genome-wide association studies, together with drug disease indications, to identify potentially novel uses for drugs. Additionally, determining the appropriate treatments for disease extends beyond predicting compound-protein interactions, as the variation of molecular and genetic features between individuals continues to confound

efforts to delineate clearly relationships with diseases across large cohorts. We (Yves Lussier and Haiquan Li, University of Illinois at Chicago, USA) presented a novel method - the Functional Analysis of Individual Microarray Expression (FAIME) - for translating the gene expression of each array into pathways and functions that can be used for predicting clinical outcome or response to therapy. The method was able to discriminate clinical samples and predict patient survival with high accuracy, providing human-interpretable classifiers composed of biologic mechanisms. Further refining the granularity at which diseases are considered, Chen-Hsiang Yeang (Academia Sinica, Taiwan) used genetic evolutionary dynamics to study heterogeneity in cancer at the single-cell level. His methods allow computational screening of multiple compounds, facilitating inclusion of the genetic dynamics among subpopulations into treatment plans.

Noncoding DNA and post-ENCODE analysis

The ENCODE project has produced abundant information for biological discovery, which might have as much of an impact on biology and medicine as the original human genome project. Misook Ha (Samsung Advanced Institute of Technology, Korea) described a probabilistic model for chromatin histone H3K4me3 occupancy based on primary sequence at the single-base-pair level. The method unveiled the genomic DNA preference for H3K4me3 modifications and was consistent with results from ENCODE in enriched H3K4me3 binding regions. Junwen Wang (The University of Hong Kong, China) demonstrated an integrative system for annotation, prioritization and visualization of SNP function. His approach integrated data from a comprehensive number of biological scales - expression quantitative trait loci, microRNA and post-translational modification - into reprioritized scores to detect buried SNPs in genome-wide association study data.

Taken together, the ISMB-ECCB research community engages in highly collaborative science and comprises computer scientists, statisticians, mathematicians and biologists who have significantly advanced the field of biology with computational approaches in the post-genome era. With the launch of the first comprehensive ENCODE release, the community is now poised to solve more-challenging modeling problems, discovering the mechanisms within and across multiple biological scales and to translate these solutions to study the variant or aberrant molecular mechanisms underpinning human diseases.

Abbreviations

ENCODE, Encyclopedia of DNA Elements; SNP, single-nucleotide polymorphism; TCGA, The Cancer Genome Atlas.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This work has been supported in part by the University of Illinois Cancer Center, the University of Illinois CCTS (UL1TR000050) and The Institute for Interventional Health Informatics.

Author details

[1]Department of Medicine, University of Illinois at Chicago, Chicago, IL 60612, USA. [2]Department of Bioengineering, University of Illinois at Chicago, Chicago, IL 60612, USA. [3]Institute for Interventional Health Informatics, University of Illinois at Chicago, Chicago, IL 60612, USA.

Published: 23 August 2013

doi:10.1186/gb-2013-14-8-310
Cite this article as: Lussier YA, *et al*.: Conquering computational challenges of omics data and post-ENCODE paradigms. *Genome Biology* 2013, **14**:310.