Exploiting nucleosome-positioning data for gene-finding; a feasibility study

Christopher N. Wilks* and Dietlind L. Gerloff

Biomolecular Engineering, University of California Santa Cruz,
Santa Cruz, CA , USA
**Email: cwilks@ucsc.edu*

Abstract: One consequence of the "omics era" is that genome-wide data sets from varied biological laboratory experiments are made available publicly. Based on previously observed correlations with, e.g. protein-encoding gene locations, such data might be exploitable to improve annotation. Viewing the malarial parasite *Plasmodium falciparum* as a model organism among AT-rich unicellular eukaryotic genomes, we have carried out a feasibility study to investigate whether data emerging from digestion-based nucleosome-positioning assays (MAINE and FAIRE) might be exploitable for gene finding in this organism. To this end we (i) characterized the experimental data from Ponts et al.[2] with reference to confidently labeled exon data from the malarial cDNA database, and (ii) investigated a baseline random forest (RF) classifier method. Our reference data consisted of several hundred thousand exonic and intergenic data points and we utilized only the raw experimental data (normalized scores from the digestion methods), so as to establish a lower-bound for future research.

The results of both parts of our study confirm the potential of the MAINE/FAIRE data to aid in gene finding that was previously noted by the authors of the experimental paper. In the case of the widely-studied malarial parasite, RNAseq transcriptome data are also available[1] to validate our classification. Our study indicates that nucleosome positioning data, where available, may be valuable to validate, or improve, gene location annotation in less studied organisms for which RNAseq-data will not emerge.

References

1. Otto D, Wilinski D, Assefa S, Keane T, Sarry L, Böhme U, Lemieux J, Barrell B, Pain A, Berriman M, Newbold C, and Llinás M. New insights into the blood-stage transcriptome of Plasmodium falciparum using RNA-Seq. Molecular Microbiology 2010; 76:12-24.

2. Ponts N, Harris EY, Prudhomme J, Wick I, Eckhardt-Ludka C, Hicks GR, Hardiman G, Lonardi S, Le Roch KG. Nucleosome landscape and control of transcription in the human malaria parasite. Genome Research 2010; 20:228-238.

* Corresponding author.