Editorial: Annotation and curation of uncharacterized proteins: systems biology approaches

Prashanth Suravajhala [1, 2, 3]*, Alfredo Benso [4] and Jayaraman K. Valadi [1, 5]

[1] Bioclues.org, Hyderabad, India, [2] Bioinformatics.org, Hudson, MA, USA, [3] Faculty of Health and Medical Sciences, University of Copenhagen, Frederiksberg, Denmark, [4] Politecnico di Turino, Turin, Italy, [5] Department of Informatics Center, Shiv Nadar University, Noida, India

Genomes encode thousands of sequences that play significant roles in diverse biological processes. Understanding the biological functions of these sequences is a challenging, yet promising task. Besides a large set of well-known genes, genomes also contain un-annotated regions whose functions are not known, aptly called "known unknowns" (KU) (Logan, 2009). Research on characterizing the KUs has not only been used to analyze the regulatory effects of genome variation, but also to improve next generation sequencing algorithms. In categorizing functional moieties from hypothetical proteins (HPs), functional genomics has opened new opportunities in identifying the cause for many diseases. For example, the machine learning based approach would allow us to predict sequences for larger sets of the KUs. Although systems biology approaches are applied in classifying the probable functions of these KU products, a few challenges remain.

This research topic gives a synopsis of the current state-of-the art methods to classify and functionally annotate uncharacterized proteins. Ijaq et al. (2015) provide one such framework. The authors discuss the need for functional characterization of HPs with next generation sequencing methods to accelerate multiple areas of genomics, and suggest the use of mass spectrometry as a promising analytical technique in validating protein characterization methods. Discovery and classification of HPs is covered in two papers. Barnkob et al. (2014) present a project designed to collect the necessary data in characterizing the expression of all membrane proteins within the scheme on hematopoietic cells. Another work by Micale et al. (2014) deliberates a way to functionally annotate uncharacterized proteins based on local sequence similarities.

To show how the annotation of HPs may be useful, Ravooru et al. (2014) demonstrate how to annotate uncharacterized proteins with the help of metabolic pathways involved in a known disease. There are two papers that focus on practices to optimize the annotation process. In the first one, Mazandu and Mulder (2014) discuss the problem of comparing genomes annotated using Gene Ontology (GO) terms by proposing a genome-scale approach for integrating annotations from different pipelines using semantic similarity measures. In the second paper, Anton et al. (2014) push for the scientific community to accelerate the rate of gene function validation as a necessary paradigm shift in assigning gene function from the gush of new genome sequences.

Keeping in view the need for rapid identification and characterization of un-annotated proteins, we argue that noncoding RNAs (ncRNAs) may play a large role in understanding the genomic repertoire encoding the un-annotated regions of the genome. We have earlier proposed a six-point classification scoring schema for annotating HPs (Suravajhala and Sundararajan, 2012) and further, the work was projected on the lines of predicting functions using similactors, which are similar proteins and yet not interacting (Benso et al., 2013). We may reason that such approaches may be applicable for these protein interactions. This system-wide omics' approach, we believe would considerably improve the translation of bioinformatics data generated into wet-lab experiments for predicting better drug targets, which in turn may serve as prognostic and diagnostic markers

for various diseases (Prensner and Chinnaiyan, 2011). Creating mechanistic archetypes of all uncharacterized regions associated with coding and noncoding genomic repertoire could bring hope for characterizing the uncharacterized.

References

Anton, B. P., Kasif, S., Roberts, R. J., and Steffen, M. (2014). Objective: biochemical function. *Front. Genet.* 5:210. doi: 10.3389/fgene.2014.00210

Barnkob, M. S., Simon, C., and Olsen, L. R. (2014). Characterizing the human hematopoietic CDome. *Front. Genet.* 5:331. doi: 10.3389/fgene.2014.00331

Benso, A., Carlo, S. D., Rehman, H. U., Politano, G., Savino, A., and Suravajhala, P. (2013). A combined approach for genome wide protein function annotation/prediction. *Proteome Sci.* 11(Suppl. 1):S1. doi: 10.1186/1477-5956-11-S1-S1

Ijaq, J., Chandrasekharan, M., Poddar, R., Bethi, N., and Sundararajan, V. S. (2015). Annotation and curation of uncharacterized proteins-challenges. *Front Genet.* 6:119. doi: 10.3389/fgene.2015.00119

Logan, D. C. (2009). Known knowns, known unknowns, unknown unknowns and the propagation of scientific enquiry. *J. Exp. Bot.* 60, 712–714. doi: 10.1093/jxb/erp043

Mazandu, G. K., and Mulder, N. J. (2014). The use of semantic similarity measures for optimally integrating heterogeneous Gene Ontology data from large scale annotation pipelines. *Front. Genet.* 5:264. doi: 10.3389/fgene.2014.00264

Micale, G., Pulvirenti, A., Giugno, R., and Ferro, A. (2014). Proteins comparison through probabilistic optimal structure local alignment. *Front. Genet.* 5:302. doi: 10.3389/fgene.2014.00302

Prensner, J. R., and Chinnaiyan, A. M. (2011). The emergence of lncRNAs in cancer biology. *Cancer Discov.* 1, 391–407. doi: 10.1158/2159-8290.CD-11-0209

Ravooru, N., Ganji, S., Sathyanarayanan, N., and Nagendra, H. G. (2014). *Insilico* analysis of hypothetical proteins unveils putative metabolic pathways and essential genes in *Leishmania donovani. Front. Genet.* 5:291. doi: 10.3389/fgene.2014. 00291

Suravajhala, P., and Sundararajan, V. S. (2012). A classification scoring schema to validate protein interactors. *Bioinformation* 8, 34–39. doi: 10.6026/97320630008034