JOURNAL OF PROTEOMICS 105 (2014) 5–18

Available online at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/jprot

Review

Non-model organisms, a species endangered by proteogenomics☆

Jean Armengaud[a,*], *Judith Trapp*[a,b], *Olivier Pible*[a], *Olivier Geffard*[b], *Arnaud Chaumot*[b], *Erica M. Hartmann*[a]

[a]*CEA, DSV, IBEB, Lab Biochim System Perturb, Bagnols-sur-Cèze F-30207, France*
[b]*Irstea, UR MALY, F-69626 Villeurbanne, France*

ARTICLE INFO

Available online 16 January 2014

Keywords:
Non-model organisms
High-throughput proteomics
Next-generation sequencing
Draft genome
RNA-seq
Proteogenomics

ABSTRACT

Previously, large-scale proteomics was possible only for organisms whose genomes were sequenced, meaning the most common model organisms. The use of next-generation sequencers is now changing the deal. With "proteogenomics", the use of experimental proteomics data to refine genome annotations, a higher integration of omics data is gaining ground. By extension, combining genomic and proteomic data is becoming routine in many research projects. "Proteogenomic"-flavored approaches are currently expanding, enabling the molecular studies of non-model organisms at an unprecedented depth. Today draft genomes can be obtained using next-generation sequencers in a rather straightforward way and at a reasonable cost for any organism. Unfinished genome sequences can be used to interpret tandem mass spectrometry proteomics data without the need for time-consuming genome annotation, and the use of RNA-seq to establish nucleotide sequences that are directly translated into protein sequences appears promising. There are, however, certain drawbacks that deserve further attention for RNA-seq to become more efficient. Here, we discuss the opportunities of working with non-model organisms, the proteomic methods that have been used until now, and the dramatic improvements proffered by proteogenomics. These put the distinction between model and non-model organisms in great danger, at least in terms of proteomics!

Biological significance

Model organisms have been crucial for in-depth analysis of cellular and molecular processes of life. Focusing the efforts of thousands of researchers on the *Escherichia coli* bacterium, *Saccharomyces cerevisiae* yeast, *Arabidopsis thaliana* plant, *Danio rerio* fish and other models for which genetic manipulation was possible was certainly worthwhile in terms of fundamental and invaluable biological insights. Until recently, proteomics of non-model organisms was limited to tedious, homology-based techniques, but today draft genomes or RNA-seq data can be straightforwardly obtained using next-generation sequencers, allowing the establishment of a draft protein database for any organism. Thus, proteogenomics opens new perspectives

☆ This article is part of a Special Issue entitled: Proteomics of non-model organisms.

* *Corresponding author at:* Laboratoire de Biochimie des Systèmes Perturbés, CEA Marcoule, DSV, iBEB, SBTN, LBSP, F-30207 Bagnols-sur-Cèze, France. Tel.: +33 4 66 79 68 02; fax: +33 4 66 79 19 05.
E-mail address: jean.armengaud@cea.fr (J. Armengaud).

1874-3919/$ – see front matter
http://dx.doi.org/10.1016/j.jprot.2014.01.007

for molecular studies of non-model organisms, although they are still difficult experimental organisms.

This article is part of a Special Issue entitled: Proteomics of non-model organisms.

Contents

1. Introduction . . . 6
2. Model *versus* non-model organisms . . . 8
   2.1. Defining model, and conversely non-model, organisms . . . 8
   2.2. Key biological questions require the investigation of non-model organisms . . . 9
3. Traditional tools and strategies for proteomics of non-model organisms . . . 9
   3.1. Sample preparation . . . 9
   3.2. Data analysis . . . 10
4. A survey of some recent achievements with traditional approaches . . . 11
   4.1. Non-model plants . . . 11
   4.2. Non-model animals . . . 11
   4.3. Pathogens, parasites, and exotic organisms take the stage . . . 12
   4.4. Megaviruses masquerading as normal cellular organisms . . . 12
5. Proteogenomics, *sensu stricto* and *sensu lato* . . . 12
   5.1. Proteogenomics, *sensu stricto* . . . 12
   5.2. Recent trends in proteogenomics . . . 12
   5.3. Proteogenomics, *sensu lato* . . . 13
6. The winning alliance of genomics and proteomics for non-model organisms . . . 14
   6.1. Unfinished genome sequences bypass time-consuming sequence assembly and annotation . . . 14
   6.2. RNA-seq generates directly informative mature protein sequences . . . 14
7. Perspectives and concluding remarks . . . 15
Acknowledgments . . . 15
References . . . 15

1. Introduction

Biological systems have adapted to a large variety of habitat conditions found previously and currently on Earth. Taxonomists have delineated these specialized life forms within species, the basic unit of biological classification, defined by significantly different traits. For multicellular eukaryotes, primarily animals and plants, the frontier between two species is the ability to interbreed and produce fertile offspring; however, the definition is more ambiguous for unicellular organisms [1]. The current estimate of diversity is still controversial and probably far from reality because the tools to assess this diversity are still under development and unexplored biomes are awaiting characterization. The recent expansion of molecular techniques highlighted important genetic heterogeneities surpassing the richness of species diversity previously predicted from morphological similarities [2]. Today, the panoply of life on Earth, *i.e.*, the entire biosphere, is immensely rich in species with between 2 and 50 million non-microbial species [3–5] and between 10 million and 1 billion prokaryotic species [6,7]. The complete census of a soil sample, which can contain upwards of 1 billion bacterial cells per gram, is currently impossible, especially given that estimates of its diversity vary over three logs of magnitude up to 100,000 species, with most of these species having never been observed or otherwise detected (Fig. 1). A glimpse of the huge diversity of microorganisms was revealed in metagenome sequencing surveys that resulted in the definition of more than 60 novel bacterial phyla, half of which have no cultivable representatives and are currently considered "microbial dark matter" because of the difficulties their characterization poses [8]. An initial assessment of the diversity of microorganisms can be obtained using shotgun DNA sequencing, and the recent development of whole-cell matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry allows the rapid identification of isolates and so could foster the discovery of new cultivable organisms from the poorest characterized branches of the tree of life [9]. Metaproteomics is also of interest to assess microbial diversity, in addition to its primary goal of characterizing microbial community functions [10]. Metaorganisms appear well characterized in terms of taxonomy, but at the molecular level much work remains. The existence of 0.4 million plants, 1.5 million fungi and 8.7 million animals has been estimated [5]. Among animals, the largest diversity arises from arthropods (90,000 myriapods, 150,000 crustaceans, 600,000 arachnids, and 5 million insects), while chordata (5500 mammals, 10,000 birds, 10,000 reptiles, 15,000 amphibians and 40,000 fishes) are less diverse. However, the number of sequenced genomes falls far short of the number of known organisms and poorly reflects the true diversity of life on Earth (Fig. 1).

The legacy of Charles Darwin, his formulation of the central principles of evolutionary biology in *On the Origin of Species* in 1859, is based on the obvious existence of shared traits between organisms. Molecular biology approaches applied to a few model organisms from the tree of life confirmed the common