Microbiological Research 171 (2015) 52–64

Contents lists available at ScienceDirect

Microbiological Research

journal homepage: www.elsevier.com/locate/micres

Multi -omics and metabolic modelling pipelines: Challenges and tools for systems microbiology

Marco Fondi[a,b,*], Pietro Liò[c]

[a] Florence Computational Biology Group (ComBo), University of Florence, Via Madonna del Piano 6, Sesto Fiorentino, Florence 50019, Italy
[b] Laboratory of Microbial and Molecular Evolution, Department of Biology, University of Florence, Via Madonna del Piano 6, Sesto Fiorentino, Florence 50019, Italy
[c] University of Cambridge, Computer Laboratory, 15 JJ Thomson Avenue, CB3 0FD Cambridge, UK

ARTICLE INFO

Article history:
Received 3 October 2014
Received in revised form 2 January 2015
Accepted 3 January 2015
Available online 7 January 2015

Keywords:
Multi -omics
Systems microbiology
Computational biology
Metabolic modelling
-omics integration

ABSTRACT

Integrated -omics approaches are quickly spreading across microbiology research labs, leading to (i) the possibility of detecting previously hidden features of microbial cells like multi-scale spatial organization and (ii) tracing molecular components across multiple cellular functional states. This promises to reduce the knowledge gap between genotype and phenotype and poses new challenges for computational microbiologists. We underline how the capability to unravel the complexity of microbial life will strongly depend on the integration of the huge and diverse amount of information that can be derived today from -omics experiments. In this work, we present opportunities and challenges of multi -omics data integration in current systems biology pipelines. We here discuss which layers of biological information are important for biotechnological and clinical purposes, with a special focus on bacterial metabolism and modelling procedures. A general review of the most recent computational tools for performing large-scale datasets integration is also presented, together with a possible framework to guide the design of systems biology experiments by microbiologists.

Contents

* Corresponding author at: Department of Biology, University of Florence, Via Madonna del Piano 6, S. Fiorentino, Florence, Italy. Tel.: +39 055 4574736.
E-mail address: marco.fondi@unifi.it (M. Fondi).

http://dx.doi.org/10.1016/j.micres.2015.01.003
0944-5013/

Introduction

The ease at which genomes are currently sequenced has assigned to genomics one of the first steps in microbial systems biology. Regardless of the technique used, assembly and annotation typically follow genome sequencing and return an almost complete picture of the genetic reservoir of a given microorganism. On the other hand, genome sequence only represents a snapshot of the real phenotypic capabilities of an organism, providing very few indications on other crucial aspects of the underlying life cycle such as response to environmental and genetic perturbations, fluctuations in time, gene essentiality and so on. To gain a systemic and exhaustive description of living entities, static information deriving from genome sequence is not enough and other levels of knowledge must be taken into consideration. Nowadays, technologies do exist for measuring, in a large-scale fashion, other crucial aspects of cellular life, including the level of RNA within the cell (transcriptomics), the nature of metabolites present within the cell (metabolomics), the interaction among different proteins (protein–protein interaction) and many others (detailed below). Also, metabolic biodiversity of microbial communities can be today evaluated through metagenomics and metatranscriptomics approaches. However, no single -omics analysis can fully unravel the complexities of fundamental microbiology (Zhang et al. 2010). Multi- and integrated -omics approaches have thus started spreading among several research areas, from bio-based fuel production (Zhu et al. 2013) to biopharmaceuticals processes (Schaub et al. 2012), from medical research (Wiench et al. 2013) to host–pathogen interactions (Ansong et al. 2013b). The integration of such diverse data types may be considered one of the key challenges of present-day bioinformatics, due to different data formats, high data dimensionality and need for data normalization.

One of the most important drawbacks associated with the booming of genomics resides in the possibility to (almost) automatically derive the potential metabolic landscape of a strain, given its genome. Bacteria continuously provide industry with novel products/processes based on the use of their metabolism and numerous efforts are being undertaken to deliver new usable substances of microbial origin to the marketplace (Beloqui et al. 2008), including pharmaceuticals, biofuels and bioactive compounds in general (George et al. 1983; Garcia-Ochoa et al. 2000; Lee et al. 2005; Zou et al. 2012). In this context, computational modelling and in silico simulations are often adopted by metabolic engineers to quantitatively simulate chemical reactions fluxes within the whole microbial metabolism. To exploit computational approaches, genome annotation-derived metabolic networks are transformed into models by defining the boundaries of the system, a biomass assembly reaction, and exchange fluxes with the environment (Durot et al. 2009). Also needed are (i) structured (mathematical) representation of that network, (ii) possibly quantitative parameters enabling simulations or predictions on the joint operation of all network reactions in a given environment and, in particular, (iii) predictions on the values of metabolite fluxes and/or concentrations (Papin et al. 2003). A constraint-based modelling framework can then be used to automatically compute the resulting balance of all the chemical reactions predicted to be active in the cell and, in turn, to bridge the gap between knowledge of the metabolic network structure and observed metabolic processes (Varma and Palsson 1994).

Innovative high-throughput technologies (see Fig. 1) represent a valuable resource also in the context of metabolic modelling, since data integration can be performed to gain a clearer and more comprehensive picture of the metabolic traits of a given organism. Diverse data types can be mapped onto metabolic models in order to elucidate more thoroughly the metabolism of a cell and its response to environmental factors; this is usually done by including functional characterization and accurate quantification of all levels of gene products, mRNA, proteins and metabolites, as well as their interaction (Zhang et al. 2010).

Here we review and discuss possible experimental and computational pipelines for multiple data integration in microbial research, allowing the simultaneous analysis of different datatypes and their mapping onto a de novo genome annotation. We discuss which layers of biological information have been shown to be important for biotechnological/clinical purposes and whether these layers are independent or have to be considered as a single complex system. Computational insights will be reviewed, including data mining, pre-processing, assimilation and iterative integration in order to exploit all available information.

Furthermore, given the link existing between microbial phenotypes and underlying metabolism, we will discuss a general framework of the major steps and checkpoints encountered when reconstructing the metabolic network of a given organism and in its consequent exploitation for computational simulation and/or phenotype prediction.

We underline the importance of integrating different sources of information to gain a more comprehensive view of genome annotation and metabolic features in general.

Information layers overview

This section provides a schematization of the sources of information (layers) that are currently the most exploited in systems biology. These layers represent the basis of multi -omics integration discussed in the next section.

Microbial genomics pipelines

Since fast genome sequencing and preliminary data post-processing have been achieved, well-grounded experimental design and strains selection have (re)gained a key position when drafting genomics-oriented research plans (Fig. 1A). Complete genome sequence is increasingly more often the starting point for integrative pipelines (see below and Fig. 1). Sample preparation and sequencing can be performed in a few days while data post-processing (including quality check, de novo assembly, gene prediction) still represents the most demanding bottleneck of the genomics pipeline. For this reason, most of the leading genomics centres (including BGI, DOE-JGI, Craig Venter Institute, Sanger) couple genome sequencing to bioinformatics analysis and usually make their software open access to the research community. In this way, preliminary assembly, annotation and analysis of genomes are carried out, although many other popular tools exist for de novo genome annotation (see for example (Angiuoli et al. 2008; Aziz et al. 2008; Seemann 2014)).

Besides basic genomic knowledge (e.g. gene presence/absence patterns), many other additional information layers are today available to be merged and integrated when trying to fully elucidate complex biological patterns of living entities (Fig. 1). References to study cases and computational tools related to these informational layers are reported in Supporting Informations 1 and 2, respectively. These include:

- *Gene constraints* represent the additional information present in DNA sequences and not fully exploited by functional annotation pipelines. These may include the detection of gene fusions, the identification of operons and the computation of gene Codon Adaptation Index (CAI). Gene structure data (e.g. the presence of gene fusions) can guide the identification of potential protein–protein interactions (Enright et al. 1999). Further, the study of operonic organization exploits co-directional intergenic