Biochimica et Biophysica Acta 1810 (2011) 967–977

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

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

Review

Connecting genotype to phenotype in the era of high-throughput sequencing ☆

Christopher S. Henry [a,b,*], Ross Overbeek [c], Fangfang Xia [a,b], Aaron A. Best [d], Elizabeth Glass [a], Jack Gilbert [a,e], Peter Larsen [e], Rob Edwards [a,i], Terry Disz [a,b], Folker Meyer [a,b], Veronika Vonstein [c], Matthew DeJongh [f], Daniela Bartels [a], Narayan Desai [a], Mark D'Souza [a], Scott Devoid [a,b], Kevin P. Keegan [a], Robert Olson [a], Andreas Wilke [a], Jared Wilkening [a], Rick L. Stevens [g,h]

[a] Mathematics and Computer Science Division, Argonne National Laboratory, 9700 South Cass Avenue, Argonne, IL 60439, USA
[b] Computation Institute, The University of Chicago, 5640 S. Ellis Avenue, Chicago, IL 60637, USA
[c] Fellowship for Interpretation of Genomes, 15W155 81st Street, Burr Ridge, IL 60527, USA
[d] Department of Biology, Hope College, 35 E. 12th Street, Holland, MI 49423, USA
[e] Biological Science Division, Argonne National Laboratory, 9700 South Cass Avenue, Argonne, IL 60439, USA
[f] Department of Computer Science, Hope College, 27 Graves Place, Holland, MI 49423, USA
[g] Computer Science Department and Computation Institute, University of Chicago, 5640 South Ellis Avenue, Chicago, IL 60637, USA
[h] Computing, Environment, and Life Sciences Directorate, Argonne National Laboratory, 9700 South Cass Avenue, Argonne, IL 60439, USA
[i] Department of Computer Science, San Diego State University, 5500 Campanile Drive, San Diego, CA 92182, USA

ARTICLE INFO

Article history:
Received 10 October 2010
Received in revised form 17 February 2011
Accepted 13 March 2011
Available online 21 March 2011

Keywords:
SEED
RAST
MG-RAST
Metagenomics
Genome-scale metabolic models
Assembly

ABSTRACT

Background: The development of next generation sequencing technology is rapidly changing the face of the genome annotation and analysis field. One of the primary uses for genome sequence data is to improve our understanding and prediction of phenotypes for microbes and microbial communities, but the technologies for predicting phenotypes must keep pace with the new sequences emerging.
Scope of review: This review presents an integrated view of the methods and technologies used in the inference of phenotypes for microbes and microbial communities based on genomic and metagenomic data. Given the breadth of this topic, we place special focus on the resources available within the SEED Project. We discuss the two steps involved in connecting genotype to phenotype: sequence annotation, and phenotype inference, and we highlight the challenges in each of these steps when dealing with both single genome and metagenome data.
Major conclusions: This integrated view of the genotype-to-phenotype problem highlights the importance of a controlled ontology in the annotation of genomic data, as this benefits subsequent phenotype inference and metagenome annotation. We also note the importance of expanding the set of reference genomes to improve the annotation of all sequence data, and we highlight metagenome assembly as a potential new source for complete genomes. Finally, we find that phenotype inference, particularly from metabolic models, generates predictions that can be validated and reconciled to improve annotations.
General significance: This review presents the first look at the challenges and opportunities associated with the inference of phenotype from genotype during the next generation sequencing revolution. This article is part of a Special Issue entitled: Systems Biology of Microorganisms.

Published by Elsevier B.V.

1. Introduction

The field of biology is currently in the midst of a sequencing revolution [1,2]. The commercially viable sequencing platforms available today can produce $>10^{11}$ base pairs of data every week. At an average of 10^8 base pairs needed to sequence and assemble a small bacterial or archaeal genome that translates into 10^3 microbial genomes a week; nearly equivalent to all the closed bacterial and archaeal genomes that are publically available at this time (www.genomesonline.org). We have also seen a dramatic expansion at the opposite end of the sequencing hardware market, i.e., cheap platforms which can produce comparatively small sequence datasets (10^6–10^8 bp) in a matter of days, enabling the democratization of sequencing. Previously, it was necessary to prioritize the genomes and biological samples that were sequenced given limited sequencing capacity. This paradigm is now being reversed, as we divert efforts to finding new samples to sequence and exploring new ways of applying sequencing technology (e.g., RNAseq [3]).

We focus this review on technologies and methodologies for efficiently utilizing the genomic data generated by the sequencing revolution to improve our ability to understand and predict microbial phenotypes (Fig. 1). In particular we focus on technologies that

☆ This article is part of a Special Issue entitled: Systems Biology of Microorganisms.

* Corresponding author at: Building 240, Argonne National Lab, 9700 S. Cass Avenue, Argonne, IL 60439, USA. Fax: +1 847 637 1932.
E-mail address: chenry@mcs.anl.gov (C.S. Henry).

0304-4165/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.bbagen.2011.03.010

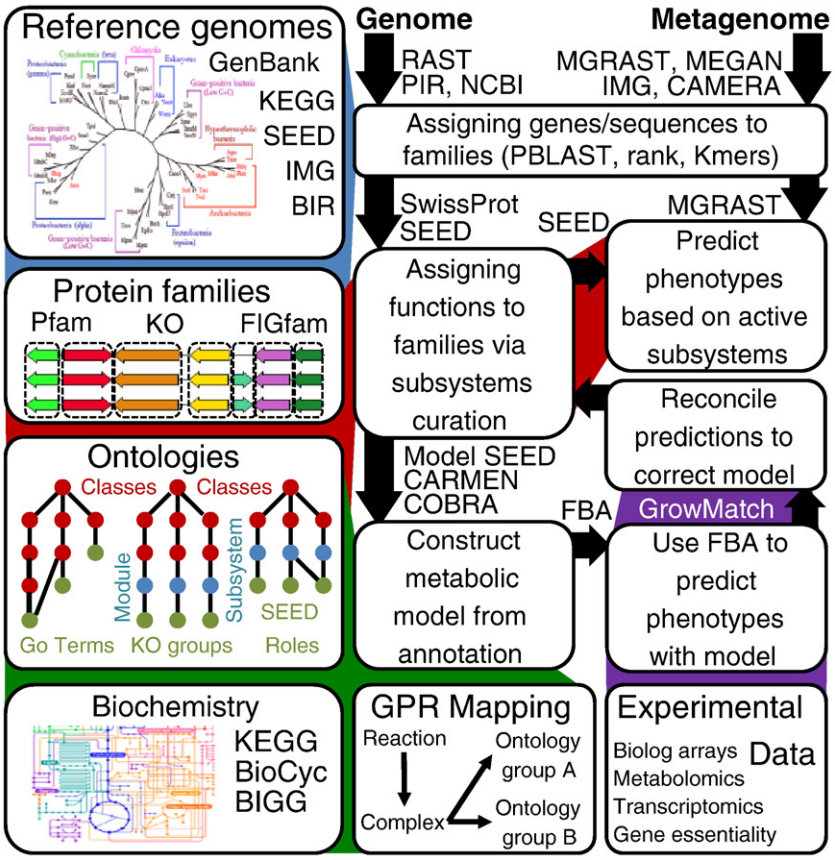

Fig. 1. Data sources and frameworks involved in predicting phenotype from genotype.

perform these tasks within the SEED framework for genome annotation and analysis, as this is the system we specifically use and develop [7,8].

We argue that there are two fundamental steps associated with the prediction of phenotypes from genotypes: (i) the process of sequence annotation that produces a list of the biological functions encoded within the genes of the sequence, and (ii) the inference of phenotypes based on an integrated analysis of the represented biological functions. A variety of frameworks and methods exist for the annotation of sequence data [4–12]. We briefly discuss the available frameworks along with their important strengths; then we focus on the Rapid Annotation using Subsystems Technology (RAST) [13] server developed within the SEED Project, with emphasis on how RAST makes high-quality, high-throughput annotation and scalable curation of annotations possible.

The list of biological functions that is output by the sequence annotation process serves as the input to the phenotype inference process. For these two processes to interface efficiently, this list of biological functions must follow a strictly controlled vocabulary. We explore how frameworks for the annotation of genome sequences accomplish this task with emphasis on how this is done within the SEED.

A variety of methods exist for translating lists of biological functions into phenotype predictions. In one approach, functions are organized into subsystems or pathways (e.g. DNA replication, alanine biosynthesis) for which the integrated overall function is understood. Phenotypes are then predicted based on the evidence for the presence or absence of a given subsystem or pathway from an organism. Another useful approach is genome-scale metabolic modeling [14]. Metabolic models enable the prediction of a wide variety of detailed metabolic phenotypes including gene essentiality, growth conditions, biosynthetic capabilities, and nutrient requirements [15]. We discuss the prediction of phenotypes based on subsystems and metabolic models, with a special focus on the Model SEED system for high-throughput model generation and analysis [16].

There are two fundamental types of sequencing taking place today: (i) sequencing of individual genomes from pure cell isolates (classical genomics), and (ii) sequencing of all the DNA content isolated from an environmental sample (metagenomics) [17]. The annotation and interpretation of these two types of sequence data involve unique challenges and distinct phenotype predictions, calling for the use of a range of methodologies and tools. The first portion of our review focuses on technologies for inferring phenotype from genotype based on single genome sequences. The final portion of our review examines the challenges and technologies associated with inferring phenotype from genotype based on metagenome sequences [18].

2. Decoding the genome from sequence to function

Genome sequence analysis and annotation has existed as a scientific field since the first large DNA molecules began to be sequenced over two decades ago. In that time, numerous frameworks for the annotation of genomes and genomic data have emerged, including the many notable frameworks that are in common use today: SwissProt [4], ExPasy [5], Entrez [6], KEGG [7,8], IMG [9], PIR [10], BioCyc [11], MicrobesOnline [19], and SEED [12]. Each of these frameworks has distinctive and significant strengths in genome annotation: Uniprot, SEED, and PIR include extensive manual curation efforts; Entrez integrates a database of publications and omics datasets; MicrobesOnline has extensive support for tree-based annotation and expression data; KEGG and Biocyc focus significantly on the collection and curation of metabolites, biochemical interactions, and metabolic pathways. Rather than conduct a detailed survey