OPEN ACCESS Freely available online

ISSN: 2153-0602
Journal of
Data Mining in
Genomics & Proteomics

Commentary

Distensive Framework for Characterizing Genome-Scale Metabolic Models

Lan Coffman[*]

Department of Genetics, Medical University of Bialystok, Bialystok, Poland

ABOUT THE STUDY

Large collections of Genome-Scale Metabolic Models (GSMM) can now be generated automatically to algorithmic advances and the availability of experimental datasets. Nonetheless, few tools are available to efficiently analyze such large sets of models, such as those used to investigate the relationship between genetic and metabolic heterogeneity. Machine Learning (ML) algorithms find patterns in large datasets by measuring the distance between data points. To apply ML to large model sets, a method for determining distance between genome-scale metabolic models was required. This problem is addressed by defining a different distance metric for each level of model representation: The Jaccard metric for metabolic reconstructions, graph kernels for network graph topology, and cosine similarity between flux distributions for constraint-based models. To compare the various metrics, we used two benchmark datasets, each of which contains hundreds of metabolic models: The first contains 100 human genome-scale models developed from proteomics data of four different cancer tissues, and the second contains more than 800 models of bacterial species inhabiting the human gut developed from metagenomic data.

Genetic and epigenetic variation, different micro-biome composition and function, and lifestyle factors such as physical activity all contribute to phenotypic variation in responses to the same stimuli, such as pharmaceutical treatments and diet. Genome-scale metabolic models represent an organism's or tissues metabolic network and simulate its response to specific environmental conditions such as nutrient availability. To produce context-specific models, these models can be parameterized with various types of molecular data, such as transcriptomics and proteomics. This feature makes them appealing platforms for the integration of multiomics datasets and the study of phenotypic heterogeneity, such as the complex genetics of metabolic diseases and conditions like frailty development with age. Genome scale metabolic models are knowledge-driven: Physicochemical laws are used to construct mechanistic models that can explain or approximate experimental data. Large libraries of genome-scale metabolic models, such as patient-derived models and human gut microbial communities; can now be built using automated mode generation algorithms. Despite the increase in the number of GSMM developed and published, very few methods exist to describe and study heterogeneity across the large number of different models included in such libraries. Machine Learning (ML) algorithms can automatically in patterns in large amounts of data, with no prior knowledge about the system. These methods scale well to large datasets, but it is difficult to incorporate any prior knowledge of biological systems into these algorithms. Unsupervised algorithms, also known as clustering algorithms, automatically detect underlying patterns and similarities in large amounts of un labeled data, whereas supervised methods, also known as classification algorithms, can learn known patterns from a set of annotated training data and use this information to create a predictive model capable of matching new un-seen data to a known profile. Many classification algorithms, including K-Nearest Neighbor (kNN), Support Vector Machine (SVM), and clustering methods like Hierarchical Clustering (HC) and K-means, rely on distance measures to discover similarities between different data points. We see great promise in combining genome-scale metabolic models with ML to address the issue of biological heterogeneity. As a first step toward integrating these two computational approaches, methods for measuring distance between metabolic models must be defined, allowing machine learning algorithms to find patterns in large sets of metabolic models.

The Jaccard similarity metric, defined as the size of the intersection divided by the size of the union of two sets has been used in several papers to assess metabolic similarity between GSMMs. This similarity metric can be applied to metabolic reconstructions to compute a similarity score between the sets of reactions or metabolites of two different models, but it does not take into account higher level model features such as metabolic network topology or constraint-based model constraints. As a result, we hypothesize that metrics based on different model properties could identify different patterns. To test this hypothesis, we identified three different representations of a

Correspondence to: Lan Coffman, Department of Genetics, Medical University of Bialystok, Bialystok, Poland, E-mail: coffmanlan@gmail.com

Received: 02-Sep-2022, Manuscript No. JDMGP-22-18526; **Editor assigned:** 06-Sep-2022, PreQC No. JDMGP -22-18526 (PQ); **Reviewed:** 20-Sep-2022, QC No JDMGP-22-18526; **Revised:** 27-Sep-2022, Manuscript No. JDMGP-22-18526 (R); **Published:** 04-Oct-2022. DOI: 10.4172/ 2153-0602.22.13. 265.

Citation: Coffman L (2022) Distensive Framework for Characterizing Genome-Scale Metabolic Models. J Data Mining Genomics Proteomics. 13: 265.

genome-scale metabolic model, namely a list of reactions, a graph topology, and flux constraints, and defined distance metrics based on each of these features. The distance metrics were then compared in a series of machine learning and phylogenetic analysis applications.

The concept of distance between metabolic models was expanded by developing distance metrics for three levels of model representation. We highlighted new properties of the Jaccard metric, such as its correlation with network similarity and function, and demonstrated how ML can be applied to large sets of genome-scale metabolic models, enabling efficient pattern recognition in large sets of models.