Computational Biology and Chemistry 59 (2015) 123–138

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

Review article

Reprint of "Abstraction for data integration: Fusing mammalian molecular, cellular and phenotype big datasets for better knowledge extraction"☆

Andrew D. Rouillard[a,b,c], Zichen Wang[a,b,c], Avi Ma'ayan[a,b,c,*]

[a] Department of Pharmacology and Systems Therapeutics, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place Box 1215, New York, NY 10029, United States
[b] BD2K-LINCS Data Coordination and Integration Center, United States
[c] Illuminating the Druggable Genome Knowledge Management Center, United States

ARTICLE INFO

Article history:
Received 9 March 2015
Received in revised form 4 June 2015
Accepted 5 June 2015
Available online 18 August 2015

Keywords:
Data integration
Bioinformatics
Systems biology
Systems pharmacology
Network biology

ABSTRACT

With advances in genomics, transcriptomics, metabolomics and proteomics, and more expansive electronic clinical record monitoring, as well as advances in computation, we have entered the Big Data era in biomedical research. Data gathering is growing rapidly while only a small fraction of this data is converted to useful knowledge or reused in future studies. To improve this, an important concept that is often overlooked is data abstraction. To fuse and reuse biomedical datasets from diverse resources, data abstraction is frequently required. Here we summarize some of the major Big Data biomedical research resources for genomics, proteomics and phenotype data, collected from mammalian cells, tissues and organisms. We then suggest simple data abstraction methods for fusing this diverse but related data. Finally, we demonstrate examples of the potential utility of such data integration efforts, while warning about the inherit biases that exist within such data.

Contents

1. Introduction 124
2. High-content datasets and resources 124
 2.1. Organizing and abstracting phenotype–genotype associations 124
  2.1.1. Mouse Genome Informatics Mammalian Phenotype Ontology (MGI-MPO) 124
  2.1.2. Online Mendelian Inheritance in Man (OMIM) 124
  2.1.3. Genome wide association studies (GWAS) 125
 2.2. Signatures of differentially expressed genes 125
  2.2.1. The Gene Expression Omnibus (GEO) 125
  2.2.2. The Connectivity Map and LINCS 125
 2.3. Genome mapping 125
  2.3.1. Encyclopedia of DNA Elements (ENCODE) 125
  2.3.2. The Roadmap Epigenomics Project 126
  2.3.3. Genotype-Tissue Expression (GTEx) Project 126
 2.4. Drug induced cellular phenotypes 126
  2.4.1. Cancer Target Discovery and Development (CTD^2) Network 126
  2.4.2. Cancer Cell Line Encyclopedia (CCLE) and Genomics of Drug Sensitivity in Cancer (GDSC) 126

DOI of original article: http://dx.doi.org/10.1016/j.compbiolchem.2015.06.003.
☆ A publishers' error resulted in this article appearing in the wrong issue. The article is reprinted here for the reader's convenience and for the continuity of the special issue. For citation purposes, please use the original publication details "*Computational Biology and Chemistry*" 58 (2015) 104–119.
* Corresponding author at: Department of Pharmacology and Systems Therapeutics, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place Box 1215, New York, NY 10029, United States.
E-mail address: avi.maayan@mssm.edu (A. Ma'ayan).

http://dx.doi.org/10.1016/j.compbiolchem.2015.08.005
1476-9271/

2.5. Properties of drugs 127
2.5.1. DrugBank, Pubchem and PharmGKB 127
2.5.2. Side effect resource (SIDER), FDA Adverse Event Reporting System (FAERS) and Offsides 127
2.6. Pathway databases 127
2.7. Physical interaction databases 128
2.7.1. Protein–protein interaction databases 128
2.7.2. The human complexome 128
2.8. Molecular patient data 128
2.8.1. The Cancer Genome Atlas (TCGA) 128
2.9. Resources overview 128
3. Data structures 129
3.1. Attribute tables and bi-partite graphs 129
3.2. Adjacency matrices to represent single entity networks 130
3.3. Set libraries 130
4. Data analysis and data integration 130
4.1. Supervised and unsupervised learning 130
4.1.1. Unsupervised clustering 130
4.1.2. Supervised classification and learn-to-rank algorithms 132
4.1.3. Predicting PPI links 132
4.1.4. Benchmarking computational methods 134
4.1.5. Predicting transcription-factor/target–gene and drug/target interactions 135
4.1.6. Graphical models 135
4.2. Biases within big biomedical datasets 135
5. Conclusions 136
Acknowledgements 136
References 136

1. Introduction

Big Data does not have to be defined by sheer size, i.e., gigabytes, tera-bytes, or peta-bytes of data, but by the fact that almost all the variables of a complex system can be measured over time and under different conditions (Mayer-Schönberger and Cukier, 2013). Computational biology tools and databases rapidly emerge with an attempt to organize and integrate molecular and phenotype data for the ultimate goal of making predictions by performing virtual experiments. Data integration enables imputing missing values given the already existing data, identifying unexpected relationships between variables, mostly through correlation analyses such as unsupervised clustering, learn-to-rank methods such as enrichment analyses, network reconstruction methods, and supervised machine learning algorithms which are used to make predictions for unseen instances. Integrating x-omics data, a.k.a. the integrome is not as difficult as it may seem because most diverse datasets and resources represent their data in a relatively structured format with common fields such as cells, genes, proteins, drugs, diseases, and assays. Such diverse but structured data can be converted into attribute tables, bi-partite graphs, single-node-type networks, hierarchies and set libraries. Such data structures provide different views of the same data and are useful for different data integration purposes. Combining two or more datasets, if they share common entities such as: genes/proteins, cells, small-molecules/drugs, tissues/tumors/patients, or diseases/phenotypes/side-effects, can lead to new insights. Here we summarize some of the most relevant resources for x-omics data integration for better extracting knowledge from Big Data. We then define the data structures that can be used to combine such resources, and briefly review the primary methods that can be used to operate on the combined data for knowledge discovery, while providing a few examples applied to real data. While we recognize that typically system level data and the methods to integrate and analyze such data were initially developed for model organisms such as yeast, worm, fly and zebra fish, the focus of this review is on data collected from the mammalian system, as well as databases and computation tools applied to the data from mammalian cells, tissues and organisms. Finally, we discuss the concept and implications of the different biases that may exist across the diverse datasets we describe. In this next section we enlist major relevant emergent Big Data resources in computational systems biology.

2. High-content datasets and resources

2.1. Organizing and abstracting phenotype–genotype associations

2.1.1. Mouse Genome Informatics Mammalian Phenotype Ontology (MGI-MPO)

The Mammalian Phenotype Ontology (Smith et al., 2004) initially developed by the Mouse Genome Informatics group at the Jackson Labs (Blake et al., 2014) and expanded to an international initiative called KOMP (Austin et al., 2004) is a useful resource for connecting gene knockouts in mice to phenotypes. The MGI-MPO ontology is a controlled vocabulary of mouse phenotype terms that are related to each other in a hierarchical network, where at each branch-point a term is linked to a set of more specific sub-terms. Each phenotype is annotated with the genotypes of the mice that display the phenotype. Some of the annotated genotypes are from transgenic mice that mimic human diseases. Gene knockout annotations can be pulled from MPO to create an un-weighted attribute table connecting phenotypes to the gene knockouts known to cause the phenotypes. Similarity matrices connecting phenotypes based on shared gene knockouts or connecting gene knockouts based on shared phenotypes can be derived from the attribute table to create single-node-type networks. Similarly, a gene set library can be created by "cutting" the phenotype tree at a specific appropriate and useful level. We previously "cut" the MPO tree at level 3 and 4 to create gene set libraries for Enrichr (Chen et al., 2013), Lists2Networks (Lachmann et al., 2010), Network2Canvas (Tan et al., 2013), and Expression2Kinases (Chen et al., 2012).

2.1.2. Online Mendelian Inheritance in Man (OMIM)

The Online Mendelian Inheritance in Man (OMIM) is a database of human diseases with known genetic basis (Amberger et al., 2011, 2009). Each entry in OMIM summarizes the current state of knowledge about gene-phenotype relationships in humans. The content