Biochemical and Biophysical Research Communications 445 (2014) 757–773

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

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

Review

Integration, visualization and analysis of human interactome

Chiara Pastrello [a,1], Elisa Pasini [b,1], Max Kotlyar [a], David Otasek [a], Serene Wong [a,c], Waheed Sangrar [a], Sara Rahmati [a,d], Igor Jurisica [a,d,e,*]

[a] Princess Margaret Cancer Centre, University Health Network and TECHNA Institute for the Advancement of Technology for Health, TMDT, Room 11–314, 101 College Street, Toronto, ON M5G 1L7, Canada
[b] CRO Aviano National Cancer Institute, Cancer Bio-Immunotherapy Unit, Via Franco Gallini 2, 33081 Aviano, PN, Italy
[c] York University, Department of Computer Science and Engineering, Toronto, ON, Canada
[d] University of Toronto, Department of Medical Biophysics, Toronto, ON, Canada
[e] University of Toronto, Department of Computer Science, Toronto, ON, Canada

ARTICLE INFO

Article history:
Available online 1 February 2014

Keywords:
Omics
Interactome
Network analysis
Gastric cancer
Visual data mining

ABSTRACT

Data integration and visualization are crucial to obtain meaningful hypotheses from the diversity of 'omics' fields and the large volume of heterogeneous and distributed data sets. In this review we focus on network analysis as a key technique to integrate, visualize and extrapolate relevant information from diverse data. We first describe challenges in integrating different types of data and then focus on systematically exploring network properties to gain insight into network function. We also describe the relationship between network structures and function of elements that form it. Next, we highlight the role of the interactome in connecting data derived from different experiments, and we stress the importance of network analysis to recognize interaction context-specific features. Finally, we present an example integration to demonstrate the value of the network approach in cancer research, and highlight the importance of dynamic data in the specific context of signaling pathways.

Contents

* Corresponding author at: Princess Margaret Cancer Centre, University Health Network, TMDT, Room 11–314, 101 College Street, Toronto, ON M5G 1L7, Canada.
E-mail address: juris@ai.utoronto.ca (I. Jurisica).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.bbrc.2014.01.151
0006-291X/

1. Omics

The suffix "omics" is appended to words describing a field of study and usually involves large scale, comprehensive and systematic techniques. The first use of omics in this manner was genomics [1], and the first international project to collect the most complete data set, a building block for genomics was the Human Genome Project, launched in 1990 and completed in 2003 (http://www.genome.gov/10001772). The list of completed genomes includes 170 Eukaryota organisms, almost 3500 Virus and more than 2500 Bacteria (as listed in http://www.ebi.ac.uk/genomes/; last accessed 13th Dec. 2013). Comparably, there are 277 completed proteomes for Eukaryota organisms, more than 1600 Bacteria and more than 1100 Virus (as listed in http://www.uniprot.org/taxonomy/complete-proteomes; last accessed 13th Dec. 2013). Moreover, the structure of more than 94,000 proteins across different organisms has been described so far (as listed in http://www.rcsb.org/pdb/home/home.do; last accessed 13th Dec. 2013; the organism with the highest number of structures being *Homo Sapiens*).

Increasing data collections enable more comprehensive analyses, but coping with this data deluge is not trivial. For example, one mass spectrometry experiment can result in thousands to hundreds of thousands of spectra for one sample [2]. Likewise, next-generation sequencing produces millions of reads per sample [3]. The amount of data being generated calls for uniformity, standardization and optimized workflows [4]. Even if these are very basic concepts, they are not as widespread as one would expect. For example, there are 338 protein–protein interaction (PPI) databases, 243 metabolic pathway databases and 202 signaling pathway databases (as listed in http://www.pathguide.org/; last accessed 13th Dec. 2013) of which only some are in a format that supports data interchange across databases. Navigating through these vast resources can be challenging but integrating such data is both beneficial and increasingly necessary.

2. Integration

The emergence of high-throughput (HT) assays shifted research from hypothesis-driven exploration to data-driven hypothesis generation. However, generating substantially more data, HT methods in turn led to shifting from predominantly using statistical tools to depending on computational biology approaches, especially data mining and machine learning algorithms, to aid data analysis and interpretation [5,6]. As the number of omics disciplines grows, and with them the amount of data, the combination of an increasing number of different perspectives can give the scientist a more complete (and more realistic) view of the system they are studying. Now, the challenge is data integration, and in turn integrative data analysis [7]. For example, integrating gene expression with copy number variation data, mutation status, methylation profile and microRNA targeting can highlight the key players in a specific disease. Complex, multifactorial diseases can only be fully investigated with this type of approach. Relationships between these data and entities can effectively be represented as graphs. Thus, network visualization and analysis is becoming one of the key tools for integrative analysis.

3. Accurate representation of omics data

Data integration requires immense attention to information representation, annotation and support for accurate data exchange. Integrative computational biology supports modeling biological processes using data integrated across many omics fields. To address this, numerous data architectures have been established to effectively and efficiently collect, store, annotate and exchange data. These architectures vary in scope, intent, and standards they use. They are continuously being updated to represent the most current knowledge, and as such will contain inconsistencies and incompletions. Often, researchers rely on one or many such architectures to integrate pre-existing research with their own, or to share their own results. Understanding the nature of the architectures available as well as being able to accurately specify which have been used is critical to reducing ambiguities in this process, and improves the quality and utility of published results.

3.1. Data collection and storage

Information in omics data changes frequently, arriving in the form of peer-reviewed studies. These can be small-scale, hypothesis-based studies with a small number of results, or wider scale HT studies with thousands of results. Collecting and storing these results has necessitated the use of many diverse omics databases. These databases vary greatly in scope; for our purposes, they address entities, relationships between them, and annotations. Entity databases cover proteins, genes, small molecules, or other biologically relevant objects, and include for example UniProt [8], GeneCards [9], RefSeq [10], and DrugBank [11,12]. Relationship databases describe how these entities relate to each other. These include, but are not limited to, protein or gene interactions, drug targeting, and biological pathways. Some examples include IntAct [13], I2D [14–16], BioGRID [17], and KEGG [18]. IMEx consortium [19] and PSICQUIC registry [20] are notable collections of interaction databases that share representation and curation workflow standards [19]. Annotation databases attempt to create indices of terms and definitions. These terms are intended to unambiguously describe entities or relationships between them, often structured as an ontology, which further describes how terms relate to each other in a standard way. Some examples include Gene Ontology [21] and the numerous controlled vocabularies included in the PSI-MI standard [22]. Some databases may rely on each other's standards: for example, UniProt and other sequence databases may contain references to relevant Gene Ontology terms in their records. Resources such as GeneCards [23] integrate heterogeneous