Chapter 2

sbv IMPROVER: Modern Approach to Systems Biology

Svetlana Guryanova and Anna Guryanova

Abstract

The increasing amount and variety of data in biosciences call for innovative methods of visualization, scientific verification, and pathway analysis. Novel approaches to biological networks and research quality control are important because of their role in development of new products, improvement, and acceleration of existing health policies and research for novel ways of solving scientific challenges. One such approach is sbv IMPROVER. It is a platform that uses crowdsourcing and verification to create biological networks with easy public access. It contains 120 networks built in Biological Expression Language (BEL) to interpret data from PubMed articles with high-quality verification available for free on the CBN database. Computable, human-readable biological networks with a structured syntax are a powerful way of representing biological information generated from high-density data. This article presents sbv IMPROVER, a crowd-verification approach for the visualization and expansion of biological networks.

Key words Systems Biology, Network Model, Signaling Pathway, Crowdsourcing, Crowd Verification, sbv IMPROVER, Biological Expression Language (BEL)

1 Introduction

Over the last few decades, there was a surge in biomedical sciences that has resulted in increasing amount of diversified data. For instance, in 2014, MEDLINE counted over 21 million citations from 5647 indexed journals [1]. That is more than a 5% increase in the amount of citations from academic journals from the previous year, and more than a 100% increase from 2000. This increasing number of peer-reviewed publications in biomedical sciences creates several challenges.

First, visualization, which helps scientists in understanding biological pathways and uncovering important properties of the underlying processes. There are different pathway databases, the most popular being: KEGG, Reactome, PID, BioCyc, Cyclone, RegulonDB, WikiPathways, Pathway Commons, Pathway Assist, and NetPath. The KEGG database, initiated in 1995 by Minoru Kanehisa, is one of the first databases of biological signaling pathways freely available to the general scientific community [2].

Tatiana V. Tatarinova and Yuri Nikolsky (eds.), *Biological Networks and Pathway Analysis*, Methods in Molecular Biology, vol. 1613, DOI 10.1007/978-1-4939-7027-8_2, © Springer Science+Business Media LLC 2017

Next, verification, which aims to extract the maximum value out of verified data, is another challenge. In the wake of high-profile controversies, scientists are facing up problems with replication [3, 4]. There is growing alarm about results that cannot be reproduced. Strict guidelines to improve the reproducibility of experiments are a welcome move [5]. Verification helps to avoid inaccurate conclusions and determine the right algorithms and models. Advancements in research process verification can contribute to challenges made to existing theories. More evidence can either prove existing theories or reveal different interpretations and flaws within it. The quality of scientific predictions has become more dependent on the samples of systems that are modeled, measured, and analyzed. When only a small minority of results is tested, it raises concerns over the legitimacy of the results and the entire set of predictions. Therefore, scientific developments and diversification of data now require a community approach for its scientific verification. Community feedback is the basis of crowdsourcing, which highlights a new trend in science and technology: people working together to innovate and create extraordinary data and to find new solutions for extant challenges. Community approaches are seen as an attempt to reach consensus in the sciences. Some see progress in science as a social process dominated by the scientific community at a particular moment in time. Therefore, it can be a reflection of the paradigm of "what is right," as adopted by scientific society.

Among the most exemplary projects that utilize crowdsourcing as a data analysis tool in biosciences is the sbv IMPROVER Challenge, also known as the System Biology Verification project. IMPROVER is an abbreviation for **I**ndustrial **M**ethodology for **PRO**cess **VE**rification in **R**esearch [6]. The sbv IMPROVER project is a collaborative effort that includes scientists from IBM Research (Yorktown Heights, NY) and Philip Morris International (PMI), Research & Development (Neuchâtel, Switzerland). The goal of the project is to develop a more transparent and robust process for assessing complex scientific data in systems biology (the study of biological organisms, viewed holistically as integrated and interacting networks of genes, proteins, and biochemical reactions). This approach has implications for a wide variety of industries including pharmaceuticals, biotechnology, nutrition, and environmental safety—essentially any area that requires a more meaningful scientific analysis of Big Data [7]. Systems biology verification and industrial methodology for process verification in research are the basis of sbv IMPROVER. Researchers at IBM and Philip Morris International R&D (PMI; Neuchâtel, Switzerland) have been collaborating on a vision for quality assurance in systems biology research. The goal of collaboration is to assure the validity of complex scientific results in the area of systems biology, and recognize the power of communities to assess methodological

aspects of scientific research. Although industry shares many of the same needs for validation as academia, a methodology for verifying research is needed in the industrial setting that recognizes both speed and protection of proprietary data constraints, as well as the importance of market considerations and consumer protection. sbv IMPROVER has further advanced crowdsourcing and implemented crowd-verification; a strategy scientists use to verify networks [8]. It shows what is possible to create by combining science, technology, and organized human and social capital. Researchers who are participants in the challenge compete for grants and opportunities to present their data at the sbv IMPROVER Symposium, an international symposium that features the work of scientists from Belgium, France, Germany, India, Italy, Japan, Luxemburg, Malaysia, Poland, Russia, Spain, Switzerland, UK, and the US [9].

The collection of networks that resulted is freely available to the scientific community in a centralized web-based repository: The Causal Biological Network database. It is composed of over 120 manually controlled and well-annotated biological network models. It can be accessed at http://causalbionet.com. The website uses a MongoDB tool that allows users to search for genes, proteins, biological processes, small molecules, and keywords in the network descriptions. This systematic approach allows users to retrieve biological networks of interest. The content of networks can be searched and visualized. Nodes and edges can be filtered with all supporting evidence. The information on the resource is linked to the original articles in PubMed. Moreover, networks can be downloaded for further visualization and evaluation [10].

2 Materials

Peer-reviewed scientific articles from PubMed constituted the majority of the project's resources. They were used to analyze investigations on the topic, to combine the data, and to determine the most effective methods for their visualization and verification.

3 Methods

There are different tools and methods for pathway analysis that help determine the pathways in comprehensive biological networks. Among the most important tools for pathway analysis are GEPAT, PAGE, CPath, and EASE, as well as Cytoscape, ONDEX HTML, and Pathview. Some of these tools and methods also require the use of the biological pathway exchange languages, such as SBML, Kappa, BioPAX, and BEL.

3.1 Network Language

The networks at the sbv IMPROVER project were built using the Biological Expression Language (BEL), which is an open-source language (http://www.openbel.org/) that can represent scientific findings from life sciences in a computable form. BEL was designed to represent research by capturing causal and correlative relationships in context, where the context can include information about the biological and experimental system in which the relationships were observed, as well as the supporting publication citations. The structure of a BEL node, which includes the biological entity, the namespace, or database to standardize the nomenclature of the entity, and the function that describes the type of entity (protein, chemical, biological process, family, complex, etc.), shows the definition of the prefixes for BEL namespaces and functions that appear in the networks.

BEL statements contain three components: a subject, a predicate, and an object, representing discrete scientific findings and their relevant contextual information as qualitative causal relationships. Subjects and objects are visualized as nodes in the biological networks. Predicates are statements that connect two nodes (i.e., network edges), maintain the computability of networks, and are supported by evidence from the scientific literature. All semantic triples are in a defined ontology, for example, HGNC (www.genenames.org), SwissProt (www.uniprot.org), EntrezGene (www.ncbi.nlm.nih.gov/gene), Rat Genome Database (www.rgd.mcw.edu), or ChEBI (www.ebi.ac.uk/chebi). BEL provides the means to describe biological interactions qualitatively, but not to quantify the magnitude or rate of these interactions. This limitation is by design, as quantitative information has significant variability and is not consistently reported in the literature. BEL-based models not only represent all molecular species but also preserve the directionality of interactions [11].

3.2 IMPROVER Methodology

sbv IMPROVER is an open database for the scientific community: https://bionet.sbvimprover.com/

The crowd-verification of biological network models is performed through the following steps [12]:

1. Develop a high-performance platform for the crowd-verification of biological network models and import created biological network models onto the platform.
2. Start the crowd-verification phase by making the platform accessible to the research community, with associated incentives to stimulate online verification of nodes and edges supported by scientific findings.
3. Interpret the results after a predetermined period to identify questionable edges (e.g., edges that did not obtain a consensus from the community).

4. Organize a "jamboree" session where community members that contributed significantly to the online verification can meet recognized experts and analyze scientific evidence for the questionable edges identified in the previous step. Publish the verified and extended networks.
5. Assess the resulting networks and determine to what extent the biological mechanisms were further expanded, revised, or invalidated. Disseminate the networks for public use.

sbv IMPROVER is a robust methodology that verifies systems biology approaches using double-blind performance assessments and applies the wisdom of crowds to solve scientific challenges. The sbv IMPROVER Network Verification Challenge (NVC) asks participants to verify, modify, or create edges in selected biological network models. Its aim is to build consensus around which parts of the networks are accurate, incorrect, or incomplete.

IMPROVER building blocks need to accommodate a priori unknown input–output functions. The development of appropriate scoring metrics is a key element for the verification methodology that helps identify the strength or weakness of a building block when precise knowledge of an input–output relationship is not possible. The verification can be done internally by members of a research group, or externally by crowdsourcing to interested community members. IMPROVER is, therefore, a mix of internal/non-public and external/public assessment tests or challenges.

Biological network models are a representation of known biology within defined contextual boundaries (e.g., species, tissue, and disease). Networks consist of nodes (e.g., DNA, RNA, proteins, etc.) and edges, where edges are causal or correlative relationships between the nodes. For instance, the protein MDM2 negatively regulates the activity of the protein p53. MDM2 and p53 are the nodes, and "negatively regulates" is the edge. NVC participants are requested to verify this kind of relationship on the basis of peer-reviewed scientific literature.

The NVC website visualizes available networks, enabling participants to scrutinize relationships, and make submissions that will either extend the network or verify existing parts of the network.

Each new edge that is created must respect the network's contextual boundary conditions and be submitted with a supporting peer-reviewed academic article. New nodes can only be created as part of creating an edge. Participants can capture new edges using the Biological Expression Language (BEL).

Verification of the network includes the following:

- Supplementation of existing evidence to provide further support for an existing edge.
- Confirmation or rejection of evidence for edges, based on whether the provided reference supports the edge and whether an evidence form has been filled accurately.

When submitting additional evidence or voting on edges, participants ought to fill in the evidence and complete the vote form as completely and accurately as possible. This helps others to understand the rationale for submissions in the network and helps in the creation of the consensus-building process.

The outcome of the online verification process is the combination of submissions by different participants. Based on this, each edge can have four possible states by the end of the challenge:

- Verified: there is at least one verified piece of evidence associated with the edge. A piece of evidence is verified if the overwhelming majority of participants approved rather than rejected the evidence.
- Ambiguous: participants are divided on whether a piece of evidence supports the edge (less than 80% of participants approve or reject the edge).
- Rejected: all evidence that has been suggested in favor of an edge has been rejected by the overwhelming majority of participants during the course of the challenge.
- Not verified: the evidence for an edge did not receive sufficient submissions from participants to be considered verified.

Selection of edges that attracted a lot of attention and controversy from challenge participants is reviewed and discussed at the "jamboree." This face-to-face meeting takes place after the online verification process is completed.

4 Notes

Worldwide explosions of data generation in biomedical sciences have confronted a scientific community with a necessity for creating innovation in data visualization and high-throughput data verification.

BEL was adopted as the structured language to represent the network models in the sbv IMPROVER Network Verification Challenge (NVC). It enables the visualization of causal and correlative relationships between biological nodes and edges in computable and human-readable statements.

Biological Network Models in the sbv IMPROVER Network Verification Challenge (NVC) are verified by participants. The networks are split into five tracks: cell stress, cell fate, cell proliferation, immune response, and tissue response. The evidence is primarily based on human biology non-diseased respiratory tissue biology augmented with chronic obstructive respiratory disease biology.

Structure of the network models in the Network Verification Challenge includes nodes, edges, and context.

Nodes that are a wide range of biological entities are represented as nodes in the network models. They include proteins, DNA variants, noncoding RNA, phenotypic or clinical observations, chemicals, lipids, methylation states, and other modifications (e.g., phosphorylation). Existing nodes were identified using biological databases, such as SwissProt (www.uniprot.org), EntrezGene (www.ncbi.nlm.nih.gov/gene), Rat Genome Database (www.rgd.mcw.edu), and ChEBI (www.ebi.ac.uk/chebi).

Edge: the causal or correlative nature of relationships between nodes is represented as an edge. This allows the biological intent of the network model to be easily digested by a scientist. An example of a relationship, or edge, is TGF Beta 1 *increases* SMAD1.

Context: each edge is constructed within precisely defined contextual boundaries and based on a literature reference to justify the edge's existence. The context of an edge may include species, tissue, cell, and disease.

The nodes and edges in a network model are captured in BEL, a computable language designed for network biology.

The networks, as implemented on the NVC website, are dynamic. They can be modified as new knowledge becomes available and current edges and pieces of evidence are verified by the community.

The network models selected for the NVC were derived from CausalBioNet network models and represent important biological processes implicated in human lung physiology and specific processes related to COPD.

Non-disease networks include the following: cell proliferation, cellular stress, cell fate, pulmonary inflammation, tissue repair, and angiogenesis.

Chronic obstructive pulmonary disease (COPD) networks are: B-cell Activation and T-cell Recruitment and Activation subnetworks to represent immune processes and their role in COPD, Extracellular matrix (ECM) Degradation and Efferocytosis subnetworks were constructed by heavily modifying healthy models to specifically represent COPD-relevant mechanisms.

Networks are available for download upon registration on the sbv IMPROVER website (https://bionet.sbvimprover.com/) and are of great use to both academic and industry users in promoting future research in this area of great therapeutic importance.

Therefore, crowdsourcing efforts that take advantage of new trends in social networking have flourished. These initiatives match discipline-specific problems with problem solvers who are motivated by different incentives to compete and show that their solution is the best.

Challenge-based approaches create metrics for the comparison of possible solutions to those challenges designed to verify building blocks. The effectiveness of one methodology can promote community acceptance of the best performing methodology and can

then be used as a reference standard. sbv IMPROVER offers a complement and enhancement to the peer-review process in which the results of a submitted paper are measured against benchmarks in a double-blind, challenge-assisted peer-review process. The sbv IMPROVER approach can be applied to a variety of fields where the output of research projects is fed as input into other projects, as is the case in industrial research and development, and where verification of the individual projects or building blocks is elusive, as it is in the case of systems biology.

This approach allows for the application of network pharmacology and systems biology beyond toxicological assessment and can be applied in areas such as drug development, consumer product testing, and environmental impact analysis [13, 14].

The sbv IMPROVER approach differs from other scientific crowdsourcing approaches in that it focuses on the verification of processes in industrial contexts in addition to basic scientific questions.

Web-based graphical interfaces allow for visualization of causal and correlative biological relationships represented using crowdsourcing principles. It enables participants to communally annotate these relationships based on evidence. Gamification principles are incorporated to further engage domain experts throughout the biological sciences to gather robust peer-reviewed information from which relationships can be identified and verified.

The resulting network models represent the current status of biological knowledge within the defined boundaries, in this case, for processes relating to human lung disease. These models are amenable to computational analysis. For some period following the conclusion of the challenge, the published models will remain available for continuous use and expansion by the scientific community.

Collaborative competition has the unique ability to facilitate analysis of high-throughput data and to become an elevator to solutions. Such approaches to research allow for the organization and processing of information in a trustworthy and effective way.

References

1. Medline/Pubmed resources Detailed Indexing Statistics: 1965–2014. http://www.nlm.nih.gov/bsd/index_stats_comp.html. Accessed 24 Feb 2016
2. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27–30
3. Yong E (2012) Replication studies: bad copy. Nature 485:298–300. doi:10.1038/485298a
4. Ioannidis JP, Allison DB, Ball CA et al (2009) Repeatability of published microarray gene expression analyses. Nat Genet 41:149–155
5. Repetitive flaws (2016) Nature 529:256. http://www.nature.com/news/repetitive-flaws-1.19192
6. Meyer P, Alexopoulos LG, Bonk T et al (2011) Verification of systems biology research in the age of collaborative competition. Nat Biotechnol 29(9):811–815. doi:10.1038/nbt.1968
7. Peitsch M C (2013) sbv IMPROVER: species translation challenge open to the scientific community for submissions. American Laboratory, http://www.americanlaboratory.com/913-Technical-Articles/138841-sbv-IMPROVER-Species-Translation-Challenge-

Open-to-the-Scientific-Community-for-Submissions/. Accessed 24 Feb 2016

8. Meyer P, Hoeng J, Rice JJ et al (2012) Industrial methodology for process verification in research (IMPROVER): toward systems biology verification. Bioinformatics 28(9):1193–1201. doi:10.1093/bioinformatics/bts116
9. Boue S, Fields B, Hoeng J et al (2015) Enhancement of COPD biological networks using a web-based collaboration interface. F1000Res 4:32. doi:10.12688/f1000research.5984.1
10. Boue S, Talikka M, Westra JW et al (2015) Causal biological network database: a comprehensive platform of causal biological network models focused on the pulmonary and vascular systems. Database (Oxford) 2015:bav030. doi:10.1093/database/bav030
11. Younesia E, Hofmann-Apitius M (2013) Biomarker-guided translation of brain imaging into disease pathway models. Sci Rep 3:3375. doi:10.1038/srep03375
12. Ansari S, Binder J, Boue S et al (2013) On crowd-verification of biological networks. Bioinform Biol Insights 7:307–325. doi:10.4137/BBI.S12932
13. Hoeng J, Deehan R, Pratt D et al (2012) A network-based approach to quantifying the impact of biologically active substances. Drug Discov Today 17(9–10):413–418
14. Sewer A, Hoeng J, Deehan R et al (2014) Systems biology approaches for compound testing. In: Hoffmann RD, Gohier A, Pospisil P (eds) Data mining in drug discovery, 1st edn. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany



http://www.springer.com/978-1-4939-7025-4

Biological Networks and Pathway Analysis

Tatarinova, T.V.; Nikolsky, Y. (Eds.)
2017, XIV, 509 p. 110 illus., 98 illus. in color. With online files/update., Hardcover
ISBN: 978-1-4939-7025-4
A product of Humana Press