 doi:10.20944/preprints202007.0123.v1

The status of causality in biological databases for logical modeling: data resources and data retrieval possibilities

Vasundra Touré[1], Åsmund Flobak[2,3], Anna Niarakis[4], Steven Vercruysse[5], Martin Kuiper[1]

[1] Department of Biology, Norwegian University of Science and Technology (NTNU), Trondheim, Norway.
[2] Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology (NTNU), Trondheim, Norway.
[3] The Cancer Clinic, St. Olav's Hospital, Trondheim University Hospital, Trondheim, Norway.
[4] Department of Biology, Univ. Évry, University of Paris-Saclay, Genopole, 91025, Évry, France.
[5] Independent Scientist, Trondheim, Norway.

Abstract

Causal molecular interactions represent key building blocks used in computational modeling, where they facilitate the assembly of regulatory networks. These regulatory networks can then be used to predict biological and cellular behavior by system perturbations and *in silico* simulations. Today, broad sets of these interactions are being made available in a variety of biological knowledge resources. Moreover, different visions, based on distinct biological interests, have led to the development of multiple ways to describe and annotate causal molecular interactions. Therefore, data users can find it challenging to efficiently explore resources of causal interaction and to be aware of recorded contextual information that ensures valid use of the data. This manuscript presents a review of public resources collecting causal interactions and the different views they convey, together with a thorough description of the export formats established to store and retrieve these interactions. Our goal is to raise awareness amongst the targeted audience, i.e., logical modelers, but also any scientist interested in molecular causal interactions, about existing data resources and how to get familiar with them.

Keywords: causal interactions, databases, interoperability, biological pathway, logical modeling, computational biology

 doi:10.20944/preprints202007.0123.v1

Background

Causality is a generic principle describing the effect that one event has on another, but causal associations between elements can be difficult to ascertain. Correlation could be a sign of causation, but "correlation does not imply causation" [1]. To truly demonstrate causality, observations that permit its validation are needed. Sometimes, causality can be determined from an intuitive and Humean understanding [2] (e.g., a ball that is rolling hits another stationary ball, which causes this second ball starting to roll), while in other cases, justifying that A causes B can be hard to demonstrate without thorough evidence (e.g., controlled intervention experiment, repeatability of the observation, and plausibility).

Systems biology aspires to understand cellular behavior emerging from molecular mechanisms, and causality constitutes a key concept to clarify how biological entities interact and affect each other, and how they are implicated in cellular activities. A causal interaction is defined as a biological event where a source entity (i.e., the regulator) affects and changes a target entity (i.e., the regulated entity), in a certain context and with defined biological states of the entities [3]. For example, a protein can affect and regulate the expression of a gene; or it can regulate the activity of another protein or even its own activity (e.g., when a version of a protein stimulates the modification of itself into a different form, or by autocatalysis); or a miRNA can affect its target. These interactions constitute fundamental pieces of information for network modeling, specifically for discrete computational modeling approaches such as logical (Boolean or multivalue) modeling [4,5]. Logical modeling portrays networks composed of nodes (i.e., biological entities), linked by directed edges that represent information about the effect (i.e., activation, inhibition) of an input node ('source') on an output node ('target'). For example in a Boolean Network (BN), every node is assigned a binarized value depending on its state (0 for absent or inactive, and 1 for present or active). In order to simulate the behavior of a BN over time, the logical modeling formalism defines that the state of each node in each time step is dictated by the state of its regulators, and is described by a logical formula using the operators AND, OR and NOT. The simulation's updating schemes can vary from synchronous (all nodes updated simultaneously) to asynchronous (only one node updated at a time) [6]. While the regulatory information embedded in these types of networks may seem basic, they form a reasonable approximation to study biological information processing [7]: Boolean models produce valuable biological insights that uncover new biological mechanistic properties [8–10] and can provide fairly accurate predictions that help in biomedical applications [11,12]. However, the model building process can be a time consuming task: each component and its connections need to be carefully derived from reported experimental observations. The use of existing, curated and well annotated causal interactions as building blocks of information can facilitate this process. Therefore, investing in the biocuration of causal statements is a valuable effort that provides support for logical modeling with a set of connected entities.

In this manuscript we provide a thorough review of publicly available databases that contain explicit, implicit or integrated causal information, as well as tools and pipelines developed to infer

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 July 2020 doi:10.20944/preprints202007.0123.v1

causality. Additionally, an overview is provided of the different data formats used to store molecular causal interactions, in order to better assess the accessibility of causal interactions and the data interoperability between the resources. This should raise awareness amongst data users about the existing resources, variations in data description and formats for an improved use of causal molecular interactions.

Molecular causal interaction data resources

Molecular causal interactions can be found in a wide range of databases providing either 'explicit' or 'implicit' causality (see Figure 1 and Table 1), differentiated by the way the resources represent data. In the following subsections, we first provide the list of the most prominent databases with explicit causal interactions, i.e., data resources that specifically describe causal interaction with, at least, a source entity, a target entity, and the causal relationship. Next, we will define and describe databases with implicit causal interactions, usually pathway databases where information about causal interactions must be inferred to be usable for Boolean modeling. Together with these resources, we report on computational inference algorithms and tools developed to extract causal mechanisms from various biological datasets.

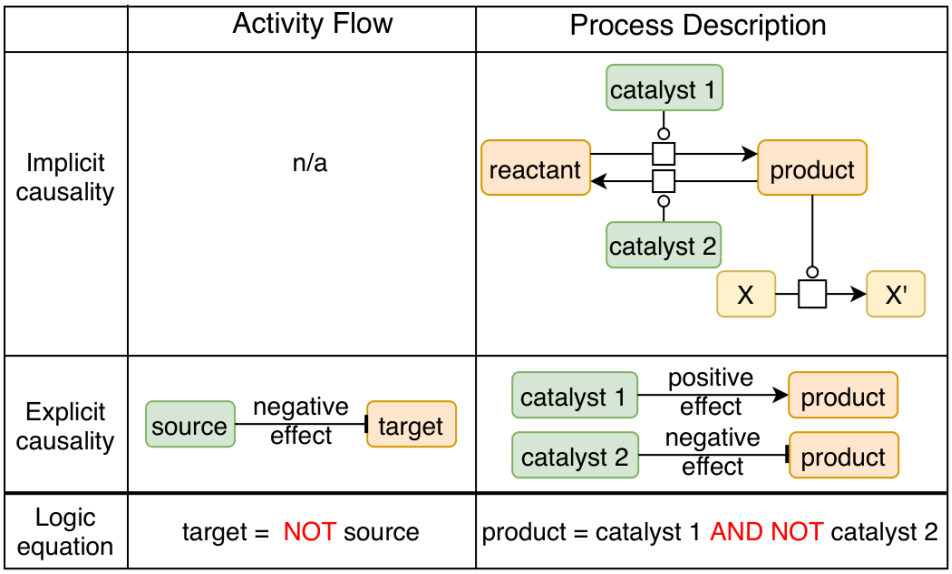

Figure 1. Representation example of explicit and implicit causal interactions in Activity Flow and Process Description. A conversion scheme from SBGN Process Description to SBGN Activity Flow has been defined in [13]. In Activity Flow, causal interactions are visible from the network's structure: the source entity has an effect on the target entity. Here, the effect is 'negative' (represented with a directed and inhibitory edge). In Process Description, the 'implicit causality' shows an example of a first metabolic reaction (state change of an entity) where a reactant entity is consumed to produce a product entity. This reaction is catalyzed by a catalyst (catalyst 1). The reverse reaction is also shown, catalyzed by catalyst 2. The product is furthermore involved, in return, as a catalyst of another reaction. Therefore, catalyst 1 has a positive effect on the product since it enables the product to perform its biological function as catalyst. Alternatively, catalyst 2 has a negative effect on the product since it prevents the product to perform that function. The causality can be inferred as: both catalyst 1 and catalyst 2 have a possible effect on an activity

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 July 2020 doi:10.20944/preprints202007.0123.v1

performed by the product entity. The product is in an active state (i.e., the state in which it catalyzes another reaction), and therefore catalyst 1 activates this particular activity of the product, while catalyst 2 inhibits it. In the case of the AF, the logic equation describes that the target is present or active in the absence of the inhibitor (specified with the operator 'NOT'). In the case of PD description, the logic equation describes that the product is present or active in the presence of catalyst 1 and in the absence of catalyst 2 (specified with the operators 'AND' and 'NOT'). A conversion from Process Description-like graphs directly to Boolean models has been described in [14], where researchers propose a four-step process for graph conversion and automatic inference of logical equations based on topology and semantics. The structure of the obtained Boolean models is an AF-like diagram.

Data resources with explicit causal interactions

Resources focused on the explicit representation of causal interactions gather their data mainly by manual curation from experimental outcomes or by computationally integrating data from other resources to assess causality between biological entities. Explicit causal interactions can be compared to the representation provided in System Biology Graphical Notation (SBGN [15], standard for visualization of biological networks) Activity Flow (AF) language. In these AF interactions, databases simply describe causal interactions in which biological entities or the activities of biological entities are central components that are linked together with an interaction that represents a specific biological effect (e.g., A regulates B, see Figure 1). In addition, some resources may provide additional contextual or defining details about the causal interaction. We next present five resources providing explicitly causal interactions.

SIGNOR

The SIGnaling Network Open Resource (SIGNOR [16,17]), https://signor.uniroma2.it/, is currently the most comprehensive biological data resource of manually assessed causal interactions, with more than 23,000 interactions annotated from experiments described in literature. SIGNOR's web interface enables the search for causal interactions based on entities of interest, which provides a global vision on the range of connections between the entity of interest with its regulators and targets. This resource mainly focuses on causal interactions between proteins, simple chemicals, complexes, families and phenotypes, where complexes and families are annotated by SIGNOR curators. The annotation of posttranslational modifications (i.e., target entities' modified residues) is recorded, as well as locational information (e.g., cellular component, cell line, tissue). The curated statements can be either computationally accessed through a RESTful service or exported in the PSI-MITAB2.8 [18] format. Note that DISNOR [19], an extension of SIGNOR, uses knowledge from DisGeNET (i.e., a collection of genes and variants associated with human diseases, [20]) to infer disease pathways by linking disease-genes to causal interactions.

SignaLink

SignaLink2.0 [21], http://signalink.org/, constitutes a manually curated resource augmented with automatically inferred causal statements, spanning multiple levels of types of biological interactions (i.e., transcription factor - target gene interactions, miRNA-mRNA interactions,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 July 2020 doi:10.20944/preprints202007.0123.v1

protein-protein interactions). These interactions are assembled to define pathways of specific biological signaling mechanisms. By integrating different sources, SignaLink associates causal effects between biological components; however, the type of effect (i.e., increase or decrease) of the source upon the target is not always known. When exporting the whole data in a CSV format, the resource currently consists of more than 89000 interactions, of which approximately 74000 have an "unknown" effect, 14627 have an 'inhibition' effect and 575 have a 'stimulation' effect. SignaLink provides exports to several formats among which PSI-MITAB2.7 [22], BioPAX level 3 (i.e., standard format for exchanging pathway information [23]) and SBML (i.e., standard format for storing biological models [24]).

Causal Biological Networks (CBN)

The Causal Biological Networks database (CBN [25]), http://causalbionet.com/, is a manually curated signaling pathways database, currently focused on biological processes in pulmonary and vascular systems [26–31]. This database contains both causal and non-causal interactions for human, mouse and rat; and combines these interactions into 46 modular networks leading to specific phenotypes for each taxon. CBN explicitly not only defines interactions between physical entities such as genes, RNAs, proteins, protein fusions, protein complexes, and small molecules; but also causal interactions between activities including biological processes, phenotypes, and pathologies; as well as interactions between entities and activities. The data can be exported in a JSON Graph File (JGF, https://jsongraphformat.info/) and the Simple Interaction Format (SIF). In addition, the CBN-BEL converter (https://github.com/pybel/cbn-bel) enables the export of these networks into the Biological Expression Language (BEL, [32]).

Gene Ontology Causal Activity Models (GO-CAMs)

The Gene Ontology Causal Activity Models [33], https://geneontology.cloud/, consists of models annotated based on the Gene Ontology (GO) terminology [34] (i.e., with GO terms describing biological processes, molecular functions and cellular components). Causality between entities in a GO-CAM model is defined via building blocks of the form: "a molecular function, enabled by a gene product, regulates a molecular function, enabled by the same or another gene product". Therefore, causality is defined in an activity-centric way in these models, and gene products are the considered entities. GO-CAMs also enable the depiction of causal regulation between biological processes, and between a molecular function and a biological process. The models are created via the web-based collaborative editor Noctua, http://noctua.berkeleybop.org, and are available in RDF Turtle (TTL, https://www.w3.org/TR/turtle/), BlazeGraph Journal (JNL, https://blazegraph.com/) and SIF formats.

Cell Collective

Cell Collective [35], https://cellcollective.org/, is a collaborative platform for building and sharing models, thereby providing a data resource of causal interactions that are connected together in simulatable Boolean models. Thus, the Cell Collective is aiming to go one step further by

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 July 2020 doi:10.20944/preprints202007.0123.v1

providing already contextualized and executable models. The resource currently contains approximately 80 models that can be exported in the SBML qual format [36], which is an extension of SBML to support qualitative modeling.

Pathway databases with implicit causality

Pathway databases are commonly built to provide a comprehensive representation of known biological mechanisms portrayed as biochemical reactions. The reactions usually represent the transformation of molecules (i.e., reactants) that are consumed to produce other molecules (i.e., products), catalyzed by enzymes. This type of representation is defined in the SBGN Process Description (PD) language [37]. SBGN PD networks commonly refer to metabolic and signaling representations of biological events (i.e., networks and interactions) where the mechanistic details are preserved: entities go through physical or locational state changes, sometimes because of the action of a catalyst or regulator (e.g., A catalyzes the reaction transforming B1 in B2, where B2 is a phosphorylated form of B1). While these networks can be used for ODE modeling, the presence of implicit causality makes the data interesting for Boolean modeling as well. However, the causal aspect between the entities of these types of networks is buried in the process description representation, which is why we name them “implicit causalities” in this manuscript. In this case, causality is not explicitly expressed and needs to be inferred based on the knowledge provided by the reaction as described in Figure 1.

Some of the popular pathway databases with signaling and gene regulatory information consist of Kyoto Encyclopedia of Genes and Genomes (KEGG) [38], Reactome [39], Atlas of Cancer Signaling Network (ACSN) [40], WikiPathways [41,42], and Disease Maps [43,44]. These resources are being used by a wide range of scientists for interpreting the outcome of experimental results. KEGG is a set of databases focusing on the understanding of functions of biological systems (i.e., the cell, the organism and the ecosystem), among others from information at the molecular level obtained through genome sequencing and high-throughput experimental technologies. The Reactome pathway database is a manually curated resource of biological mechanisms, analysis algorithms and predictive computational models. ACSN is a multi-scale repository of manually curated biological networks focused on the depiction of disease mechanisms (i.e., molecular processes found in cancer cells and tumor microenvironment). WikiPathways is a growing, collaborative platform for the curation, dissemination, visualization and analysis of biological pathways, focusing on genes, proteins, and metabolites. Disease Maps is a growing community effort that extensively builds a collection of biological networks representing mechanisms affecting different types of diseases (e.g., COVID-19 Disease Map [45], Parkinson’s Disease Map [46], AsthmaMap [47], RA-map [48,49]). The Disease Maps are manually curated and involve clinicians and domain experts to validate the annotated pathways. In addition to these resources, several individual, but valuable knowledge bases have been published and made available as pathway maps describing specific biological conditions [50–56].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 July 2020 doi:10.20944/preprints202007.0123.v1

The extraction of causal interactions from these pathway databases involves a conversion of the data structure in order to explicitly represent causality for Boolean modeling. Several studies have been performed to identify patterns of various biological interactions that would translate into causal interactions between biological entities. For instance, Vogt et al., [13] proposed rules to translate maps from the SBGN PD language to SBGN AF to obtain smaller and more manageable maps. Recently, Aghamiri et al., developed CaSQ, [14] a tool to automatically infer executable Boolean models from static molecular interaction maps represented in the CellDesigner [57] format. The researchers proposed a framework of graph conversion for PD representations including SBGN schemes, concerning various biological scenarios like complex formation, phosphorylation, ligand-receptor interaction etc., with simultaneous inference of logical formulae describing regulation. The logical rules are inferred based on topology and semantics already encoded in the original maps, resulting in Boolean models with an AF-like layout. The tool is able to process large and complex maps and produce Boolean models in a standard output format, SBML qual, directly executable using popular modeling tools. References, annotations and layout of the CellDesigner molecular maps are retained in the obtained model, facilitating interoperability and reusability of the content. Both approaches base their graph conversion rules on graph structure and glyph semantics rather than shared ontology. This also holds true for the latter approach that goes one step further as to infer logical equations based on topology and semantics encoded in the original maps. As such, they are not directly interoperable, with the generally applicable framework of OWL-based inference [58], used in the Gene Ontology community for validity checking of GO-CAMs and for automatic inferencing of knowledge that can be derived from curated knowledge [59].

Table 1. Summary of the listed 'explicit' and 'implicit' data resources with causal information. The 'Curation' column indicates whether the data is the result of 'manual' or a 'mix' of manual curation and automatic inference from either large datasets (i.e., SignaLink) or other curated knowledge (i.e., GO-CAM). The 'PTMs' (post-translational modifications) column informs whether PTMs are described in the database. The 'Exports' column indicates the formats in which causal interactions are accessible or in which formats the pathways can be extracted. The '+' symbol indicates that the database does support a specific characteristic.

	Data resource	Curation	Reference to primary literature	API	PTMs	Causal view (see Figure 2)	Exports
Explicit causality	SIGNOR	manual	+	+	+	protein-centric	PSI-MITAB2.8
	SignaLink	mix	+			entity-centric	SIF
	CBN	manual	+		+	entity-centric	BEL, SIF
	GO-CAM	mix	+	+	+	activity-centric,	OWL, SIF, JNL

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 July 2020 doi:10.20944/preprints202007.0123.v1

						protein-centric	
	Cell Collective	manual	+	+		entity-centric	SBMLqual, other format such as GPL
Implicit causality	KEGG Pathways	manual		+	+	n/a	KGML
	Reactome	manual	+	+	+	n/a	PSI-MITAB2.7, SBML, SBGN, BioPAX
	ACSN	manual	+		+	n/a	SBGN, CellDesigner XML
	Disease Maps	manual	+		+	n/a	SBML, SBGN, BioPAX

Data resources with integrated causal interactions

Several databases integrate information from other resources to incorporate and combine knowledge describing causal interactions. These resources usually provide powerful querying interfaces, via command line, web services, graphical interfaces or scripting, to retrieve data of interest for a particular study (see Supplementary File 1). For instance, OmniPath, today the most comprehensive integrated resources for causal interactions, combines a vast range of curated and computed resources, to offer a single endpoint for querying data [60,61]. These resources include among others, SIGNOR, SignaLink, and SPIKE [62] that contain directed interactions (full list available at http://omnipathdb.org/). They can be accessed and analyzed through pypath (http://saezlab.github.io/pypath), a Python module enabling the construction of models from the multiple supported resources. BioGateway [63] is a semantic knowledge base with curated information from a variety of Elixir core data resources [64] and recently augmented with a knowledge graph containing regulatory interactions between transcription factors and target genes, which can be queried using the SPARQL query language. Its use in network building has been facilitated by the BioGateway app [65], presenting a user-friendly interface to query the BioGateway triple store from the Cytoscape network editor, making data querying more broadly accessible to biologists, and allowing the export of data in all the formats supported by Cytoscape, including the SIF file format. NDEx [66] is an online exchange platform building a commons of biological networks. The resource aims to be a collaborative platform where scientists can deposit and share their networks, as well as use resources from others with multiple application possibilities such as visualization and analysis of the networks in Cytoscape. Likewise, the Pathway Commons incorporates knowledge from 22 pathway and interaction databases [67] and enables to export data in BioPAX and SIF, among other formats. PathMe aligns and integrates

pathways from KEGG, Reactome, and WikiPathways in BEL [68] and the greater Bio2BEL [69] ecosystem of which it is a part integrates more than 50 biological databases that can be exported to the different formats supported by PyBEL [32].

Causal inference in software and pipelines

In addition to resources providing causal interaction information, a broad spectrum of pipelines have emerged to infer causality from various types of biological datasets. This inferred knowledge is generally coupled with prior knowledge evidence extracted from literature. We provide here a brief overview of these tools. The INDRA-IPM modeling web interface enables the assessment of causality by translating biological mechanisms from natural language processing [70] and enabling the export of the models in various standard formats (e.g., SBML [24], SBGN [15]). The CAusal Reasoning for Network identification using Integer VALue programming (CARNIVAL [71]) is a pipeline that integrates gene expression data to identify upstream regulatory signaling pathways. CARNIVAL implements a reverse engineering process and uses a prior knowledge network obtained from OmniPath [60,61]. Similarly, CausalR [72] is an R package that extracts causality from genome expression datasets. CoRegNet [73] is an R package that infers co-regulatory networks of transcription factors and target genes by analyzing transcriptomics datasets and estimating activities of transcription factors. Whistle [74] implements the Reverse Causal Reasoning algorithm to discover upstream causal regulators from transcription profiles. It uses prior knowledge described in the BEL format to identify possible molecular mechanisms that explain the gene expression data. ARACNE [75] enables the reconstruction of mammalian transcriptional regulatory networks using both biological data (i.e., microarray datasets) and mathematical methods (i.e., Relevance Networks and Bayesian Networks algorithms). Martin et al., [76] developed a scoring method applied in their study to the biomedical field (e.g., impact of disease, drug treatment, and environmental agents on humans), the Network Perturbation Amplitude (NPA). When combined with high-throughput data and prior knowledge, the NPA algorithm identifies change of activities in targeted biological processes and thereby helps to better understand biological mechanisms leading to diseases. CausalPath [77] infers causal interactions from prior-knowledge resources (i.e., the pathway databases) combined with proteomics cell lines profiles. This technique enables the automatic contextualization of causal interactions for diseases in which they are experimentally observed. CausalPath uses mainly the Pathway Commons data to assess causality [67]. Finally, Causaly (www.causaly.com) is a commercial interface that uses artificial intelligence, machine learning and text mining to infer evidence and causality from the biomedical data. A variety of data sources, including Biomedical Literature, Clinicaltrials.gov and several side effect databases are machine-read and integrated in the Causaly Knowledge Graph. All evidence is represented as a cause-effect network operating in a Graph database, and can be queried and explored through a defined REST API.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 July 2020 doi:10.20944/preprints202007.0123.v1

Data exchange formats

Formats have been developed from multiple perspectives, each expressing causal interactions with different aspects that answer specific use cases. These formats can range from the most simple form with two entities and the regulation sign (e.g, .sif file) and more complex representations with the storage of additional metadata (e.g., the Biological Expression Language (BEL), GO-CAM using the Web Ontology Language (OWL), PSI-MITAB). The following section aims at providing a short description of the most prominent formats (detailed description can be found through the links provided in Supplementary File S2), the type of metadata these formats can handle and the list of databases providing causal interactions in these formats.

The Simple Interaction File (SIF)

The Simple Interaction File is a simple, space- or tab-delimited format composed of three elements per line: the source node, the interaction type and the target node. Each line of data corresponds to a single interaction. SIF is commonly used in Boolean modeling and constitutes the most simple representation of causal interactions. However, there is no standard agreement for annotating these interactions in SIF. It is nevertheless common to represent the source and target node with a name (e.g., MYC, for HGNC-named genes [78]) or an identifier (e.g., P01106, for a UniProt ID [79]]) of the matching biological entity, and the interaction type is annotated with terms such as “activates” / “inhibits”, “increases” / “decreases”, “up-regulates” / “down-regulates” or even symbols like “->” / “-|” to represent the causal effect of the interaction. Still, the data annotation should be consistent throughout each individual data resource. As the SIF file contains only the most important and basic information about an entity-based causal interaction (i.e., no contextual details are stored), it can be seen as a format for data users once they have filtered out the data of interest for their study (e.g., selected for appropriate biological context) from original knowledge sources. A SIF file can be easily used in visualization tools for displaying and analyzing networks (e.g., Cytoscape [80], and GINsim [5]). Several databases allow export into a SIF format, such as Cell Collective [35], OmniPath [60,61], and CBN [25].

The PSI-MITAB2.8, a tab-delimited standardized format

The Molecular Interactions community from the Human Proteome Organization Proteomics Standards Initiative (HUPO PSI-MI) recently invested efforts in the representation of causal details in molecular interactions [81]. Initially focused on the representation of molecular interactions, the directionality and effects were not supported by the PSI-MITAB2.7, a tab-delimited format used for the curation of molecular interactions. A few causal events were actually expressed as free text in the “Interaction annotation(s)” column (e.g., “P40763 decreases expression at the protein level of ENSG0000012658”) in various molecular interaction databases. Using PSICQUIC [22], a web service to programmatically access molecular interaction databases supporting PSI-MITAB (e.g., IntAct [82,83], BioGRID [84], InnateDB [85]), it was possible to retrieve 2,916 causal annotations (queried on 12/01/2020) by applying the filter ‘annot:causality’. However, no formal structure was defined to represent causal information. A recent upgrade of the PSI-MITAB format included the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 July 2020 doi:10.20944/preprints202007.0123.v1

support for both directionality and effects of molecular interactions, thus supporting representation of entity-based causality. This extension called 'CausalTAB' [18] or 'PSI-MITAB2.8', extends the PSI-MITAB2.7 with four columns to support causal information: the 'Biological effect of interactor A' and 'Biological effect of interactor B' informing about the activated molecular function (i.e., entity's activity) in the causal interaction, the 'Causal regulatory mechanism' reporting the biological mechanism underlying a causal interaction, and the 'Causal statement' informing about the effect of the causal interaction (e.g, up-regulation or down-regulation). In addition, the MI controlled vocabulary has been extended with a "causal interaction" branch to incorporate terminology to define regulatory terms [18]. PSI-MITAB2.8 is a standard format supported by the SIGNOR database, the PSICQUIC web service, and possibly other resources will enable the storage and retrieval of causality in this format.

The Biological Expression Language (BEL), a triplet-oriented format

The Biological Expression Language (BEL, https://bel.bio/) is developed to express causative or correlated relations observed in a specific biological context using BEL statements [32]. A BEL statement is a semantic triplet of information composed of a subject (the regulator), a predicate (e.g., a causal interaction like increase, decrease, and variations of these) and an object (the target). For the subject and the object, a reference to the biological entities using namespaces (controlled vocabularies of origin where it originates) is given and their related biological conditions are annotated with BEL functions (e.g. protein modification, binding, kinase activity; or epigenetic modifications [86]). Thus, BEL enables the representation of both activity- and process-based causal interactions. In addition, contextual annotations (e.g, evidence, experimental context, cell line) can be associated with a BEL statement thanks to the use of BEL annotations. The main advantage of a BEL statement lies in the fact that it presents a relatively simple and flexible syntax (e.g., PD-like and AF-like statements can be mixed) meaningful both to humans and computers. PyBEL, a Python package, has been developed to enable the parsing of BEL scripts, validation and conversion to other formats [32]. Several resources provide export of causal interactions in the BEL format, which are listed at: https://cthoyt.com/2020/04/30/public-bel-content.html).

The Gene Ontology Causal Activity Model (GO-CAM), an OWL/RDF format

The Gene Ontology Causal Activity Model (GO-CAM) [33] is a formalism developed for structuring biological processes by defining triplets of information (subject, predicate, object) using the Web Ontology Language (OWL) representation [58]. GO-CAM is based on the Gene Ontology terminology and uses the Relation Ontology terms (RO) for the annotation of causal interactions [87]. The RO ontology contains a branch called "causally related to" that groups terms for representing causal relations between material entities, between processes, and between both a material entity and a process. Still, GO-CAM is mainly based on an activity-level or reaction-level representation of causality rather than an entity-level based causality (even though the latter can be inferred). In this activity-based view, causality is represented as follows: a specific molecular function (e.g., a phosphorylation) that is enabled by an entity regulates a specific molecular

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 July 2020 doi:10.20944/preprints202007.0123.v1

function of another entity. GO-CAM is robust in terms of computational aspects: OWL is a powerful representation that structures knowledge in a way that allows for reasoning and offers the potential for real-time consistency checking during a curation process. But its understanding by humans is cumbersome at best. To address this and to help curators, the Noctua annotation platform is being developed (http://noctua.geneontology.org/), providing a graphical user interface for assembling, editing and interpretation of GO-CAM models.

The Systems Biology Markup Language qualitative models (SBML qual), an XML-based standard format

The SBML qual [36] is an XML-based standard developed by the Consortium for Logical Models and Tools (CoLoMoTo, http://www.colomoto.org/) and adopted by the SBML community, which is part of the COMBINE initiative [88]. It constitutes a more computational simulation-oriented format that is primarily meant to support exports of boolean models. The mathematical framework encoded in this format facilitates the simulation of the annotated models. Consequently, SBML qual is suited for storing causal interactions, but predominantly for linked sets of these, not individual causal interactions. Only entity-based causal interactions can be stored in the SBML qual format and this provides ready-to-use models for modeling software that support this standard (e.g., CellNOpt [89], GINsim [5], BoolNet [90],VisiBool [91], Cell Collective [35]).

Discussion

The different views on causal interactions

Over time, scientists have adopted different visions to describe causality in causal interaction data resources, based on their biological interest and subsequent data analysis purposes (see Figure 2). Data resources with entity-centric description of causal interactions are predominant. In these representations, the focus is commonly put on proteins and transcripts as actors of the causal interaction (protein-centric). The protein-centric causal statements may not report in detail gene regulatory events (see Figure 2, example 1.1). For instance, in SIGNOR, the causal relation between a transcription factor and a target gene is assessed by the transcription factor (source) indirectly regulating the gene product. In a gene-centric view, the mechanistic details of gene regulation would be represented and the gene itself would be represented as the target entity (see Figure 2, example 1.2). The gene does not have an activity per se, but the binding of a transcription factor onto the gene induces the process of transcription and thus the expression of the RNA. In this view, specificities on the gene can be represented (e.g., methylated gene) and the exact biological mechanisms could be unambiguously described, but it would make the modeling process more complex because more entities would be introduced. Other resources, such as GO-CAM, focus on the biological activities of gene products, and represent these activities as being the source and target ‘entities’ of the causal interaction (see Figure 2, example 2). These activities are usually linked to a gene product. Finally, process-based causality additionally incorporates one

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 July 2020 doi:10.20944/preprints202007.0123.v1

or two biological processes as entities, usually as a target entity to describe a phenotypic outcome (see Figure 2, example 3).

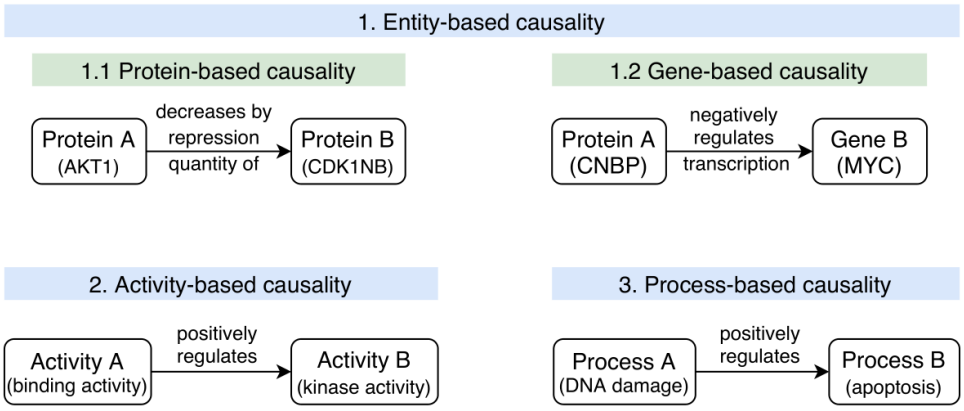

Figure 2. Examples of causal interactions with different types of representations of the interactors and the effect.

1.1. Protein-based causal interaction where both entities are gene products: Protein A (AKT1) represses the activity of Protein B (CDK1NB) by decreasing the quantity of protein by repressing it (meaning that this is most likely an indirect regulation with an intermediate entity, i.e., the gene that enables the production of Protein B).

1.2. Gene-based causal interaction where Protein A negatively regulates the transcription of Gene B.

2. Activity-based causality where an Activity A positively regulates an Activity B. Usually, the entities that perform these activities are known and annotated.

3. Process-based causality where a biological Process A (DNA damage) positively regulates a biological Process B (apoptosis).

Interoperability between data formats

Data interoperability is defined as the ability provided by a certain data format to different computer systems or tools, to exchange, comprehend, and perform meaningful processing of this data. After reviewing the different formats for storing causal interactions, it is noticeable that they all share a common core of information: two interacting entities (source and target, which can be physical entities or activities) and the regulation type (effect) of this interaction. This set of information is covered in the SIF format, commonly used as input format by network modelers in computational modeling frameworks that connect interactions together. However, the interesting aspects reside in their differences. These differences are influenced on one side by the data annotators, who format the knowledge that is accessible from primary data (e.g., literature, experiments), and on the other side by the end users, who need specific knowledge defining the context in which a causal interaction is applicable. These contextual differences can be observed in Table 2. For instance, BEL and PSI-MITAB2.8 explicitly mention and structure the post-translational modifications (PTMs) (e.g., phosphorylation, acetylation) and cellular locations of the biological entities. Notice that these defining details or ‘context’ enable end users to build

better informed models, as it informs them on the validity of chaining together several individual causal interactions, whereby the target entity from one causal interaction is used as the source entity for the next one, in the assembled model. Very likely, such a 'reused' entity's detailed PTM state should be the same in both of the originally annotated causal interactions, since the PTM often determines (one of) a protein's active versus inactive states; and also both of its original cellular compartments may be informative for assessing the validity of the assembled logical model, or for fine-tuning it to better match available experimental data. Notice also that the description of PTMs can also vary between formats, meaning that there is no agreed consensus to represent this information either. An effort to address this issue has been proposed by Danos et al. [92] in the context of rule-based modeling. Furthermore, BEL, GO-CAM, and PSI-MITAB2.8 allow to depict with precision the biological activities of the entities involved, using Gene Ontology terms. Notice that BEL in addition supports representing some of these activities using an internally defined vocabulary of a dozen high-level, shorthand terms (e.g., 'kin' for kinase activity), which PyBEL can map to GO terms (https://pybel.readthedocs.io/en/latest/_modules/pybel/language.html).

The absence of common guidelines leads to possible data interoperability issues [93]. The building of these different data resources remains a niche activity, making it arduous for data users to build data integration protocols that are able to cope with the diversity of public data. In addition, their subsequent data analysis approaches are limited by the dataset with the least expressive value. And even if the level of annotation detail is the same, the use of a specific ontology or controlled vocabulary can vary across resources. This results in non-compatible data sets, and it would demand rigorous mapping services between these sets to meet the needs of the data users. For instance, the same biological entity may be annotated and referred to with different identifiers: BEL and PSI-MITAB2.8 commonly use UniProt IDs [79] for proteins but BEL also enables the use of HGNC IDs [78] for the annotation of gene products, and whereby the context in which an identifier is used clarifies whether the identifier refers to a gene or to a specific type of gene product (e.g., having kinase activity implies that the referred biological entity is a protein). For these reasons, data aggregation between databases can be a challenging task, or at least require additional data processing tasks. This is further compounded by the fact that comparing causal statements between resources may lead to seemingly conflicting causal interactions because of the heterogeneous contextual information (and the extent to which it is supported in different formats), thereby leaving the data analyst with the task to solve possible ambiguities.

The unification of the description of molecular causal interactions goes through agreed-on guidelines for scientists to lean on when producing, accessing or using causal statements. To this end, the Minimum Information about a Molecular Interaction Causal Statement (MI2CAST, see Table 2) [3] has been recently put in place through the collaboration of scientific communities involved in causal representation (e.g., NTNU, IMEx consortium, GREEKC consortium (http://greekc.org/), Swiss Bioinformatics Institute), which organized discussions between data curators, data providers and data users with the desire to improve and homogenize the curation of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 July 2020 doi:10.20944/preprints202007.0123.v1

causal molecular interactions. The establishment of these guidelines encourages the communities (PSI-MITAB2.8, GO-CAM, BEL) to update data formats and curation protocols to comply with MI2CAST and offer better interoperability possibilities between resources supporting these formats. For instance, BEL is upgrading the representation of biological activities (mentioned in 'Interoperability between data formats' section) to enable the use of GO identifiers, similarly to PSI-MITAB2.8 and GO-CAM.

Table 2. Comparative table of annotations handled by different formats storing molecular causal interactions. Data and metadata types, described in the MI2CAST guidelines, that can currently be annotated and stored in each format are listed. The '+' sign means that the format stores explicitly this type of data or metadata. The mention '*(work in progress)*' indicates that the community maintaining the data format is working towards providing support for that term.

	SIF	SBMLqual	PSI-MITAB2.8	GO-CAM	BEL
Source entity	+	+	+	+	+
Interaction effect	+	+	+	+	+
Target entity	+	+	+	+	+
Reference			+	+	+
Evidence			+	+	+
Experimental setup			*(work in progress)*		+
Biological mechanism			+	+	+
Biological activity			+	+	+
Biological type			+		+
Biological modification			+	+	+
Taxon entity			+	+	+
Taxon interaction			+		+
Tissue type			*(work in progress)*	+	+
Cell type / Cell line			*(work in progress)*	+	+
Cellular Compartment		+	*(work in progress)*	+	+

Causal interaction formats for logical modeling

In logical modeling, the focus is put on entities, representing 'nodes' in a regulatory network. On the one hand SIF, PSI-MITAB2.8, and SBML qual are focused towards an entity-based

 doi:10.20944/preprints202007.0123.v1

representation of causality, which makes these formats more amenable for combining causal regulatory interactions into logical statements as inputs for logical modeling. On the other hand, GO-CAM and BEL allow the representation of causality from an activity-based view for the biological entities, and a process-based view for biological reactions, meaning that the focus is not put on biological entities per se but rather on functions or actions that they can perform. In this case, the use of statements for logical modeling may require a minor post-processing of the data to extract entities' information, as logical models do not explicitly focus on activities or processes, or at least they do not represent that level of details. It should be noted though that BEL in principle supports both views; and that information about the biological activities is also supported in PSI-MITAB2.8 as a defining detail.

It is important to realize that different formats may serve different purposes and needs. For storing any type of contextualized causal interactions, the PSI-MITAB2.8, BEL, and GO-CAM formats seem to be well suited as they can handle fine grained details describing a causal interaction. These formats facilitate the pre-processing work for the modelers, by providing sufficient information to select causal interactions of interest through filtering processes: a more defined context allows to more easily assess whether a causal interaction is useful for the specific biological study. The SIF and SBML qual formats seem to be better suited for handling a collection of causal interactions that together form a model for a specific context or case study, and in particular SBML qual can already store logical formulae describing the causal regulations. These files are mainly used as inputs of modeling or simulation tools; and after running the simulations, also to further analyze the outcomes of causal interaction networks for a given biological situation. In theory, PSI-MITAB2.8, BEL and GO-CAM can also be used as input files in modeling tools when interactions of interest have been selected. These statements of course only apply when a particular tool supports the import of these formats. It should be noted that for instance, in the case of PSI-MITAB2.8, additional steps are required to correctly assess information about complexes (i.e., collections of entities acting together), because this format stores a protein complex as a list of several binary interactions. These binary interactions need to be combined when modeling, in order to correctly assess the cases in which an entity is acting in combination with other entities (i.e., AND logic formulae will be added between the components of a complex). This exemplifies that different formats have different trade-offs. Managing biomolecular process information necessitates representing entities of diverse compositions, and in diverse and intricate states (even after translating their diverse interactions into causal regulations). Tabular data (like PSI-MITAB2.8) is a convenient input format for data analysis algorithms, but due to the format's inherent limitations for representing this kind of information variability, (as shown for protein complexes) it can still require an extra preprocessing step. A non-tabular format (like BEL) can support more of this information variability, but then again, may equally require a dedicated preprocessing step before serving as input for modeling algorithms. Next, note that causal interactions describe mainly the effect of a source entity on a target entity. However, when building a model that combines causal interactions, a target entity usually has more than one source entity

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 July 2020 doi:10.20944/preprints202007.0123.v1

(regulator). The way regulators affect the target entity is defined through the use of logical connectors (e.g., AND, OR, NOT) that together define when the target entity is active or not. These logics are not described in causal interactions and need to be assessed by modelers either through computer algorithms or manual curation.

To facilitate the validation of logical models through the use of causal interactions, a roadmap to guide the Curation and Annotation of Logical Models in biology (CALM, [94]) has been recently put in place to propose the adoption of good practices when building logical models. It suggests to include integrated pipelines to facilitate data reproducibility, to follow minimum requirement guidelines and use standards (i.e., SBML qual) for increasing data interoperability, to put effort in the development of automatic annotation tools for reducing the gaps between curation and model annotation, and to systematically use a common repository (i.e., Cell Collective [35], Biomodels [95]) for sharing the produced models.

Conclusion

Causal effects between molecular biological entities play an increasingly important role in the analysis and modeling of biological systems. Their relevance for a specific type of analysis often depends on the specific conditions and context in which they were experimentally observed. Given the extensive variation in metadata describing these conditions and contextual details, in combination with the large number of resources available online, it is a daunting task for anyone interested in using these causal molecular interactions to find, grasp, and understand these differences. This manuscript provides a comprehensive review to help data users appreciate this diversity. We presented data resources of causal interactions, software and pipelines that have been developed to infer causality either from datasets, pathway resources or a combination of both, and data formats handling causal information. A summary of these resources and tools are shown in Table 3. This set of knowledge constitutes a key element to facilitate the building of logical models to better predict the behavior of a cell system.

Given the range of existing resources, tools and formats, it would serve data users well if the different resources would provide thorough documentation of the curation rules followed to generate the causal statements (e.g., annotations following MI2CAST, additional contextual information, ontologies/controlled vocabularies used). This would give sufficient information to assess if (and how) data are mappable between different resources, thereby improving their interoperability, when possible, by developing tools that would enable to switch between data formats (and resources) without loss of information.

Table 3: Summary table of the resources, inference methods and formats related to biological causal interactions presented in the manuscript.

Resources	Inference Methods	Formats

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 July 2020 doi:10.20944/preprints202007.0123.v1

Explicit causality	Implicit causality	Integrated resources	Pipeline / Software	For data and metadata storage	For logical modeling
CBN Cell Collective GO-CAM SignaLink SIGNOR	ACSN Disease Maps KEGG Reactome	BioGateway OmniPath Pathway Commons NDEx	CARNIVAL CaSQ CausalR Causaly CoRegNet INDRA-IPM Whistle	BEL GO-CAM PSI-MITAB2.8	BEL GO-CAM SBML qual SIF

Key points

- Causal molecular interactions are key concepts supporting the assembly of regulatory networks for logical modeling.
- Data resources provide a wide range of causal statements that are commonly manually curated, but tools and algorithms also support the automatic inference of causal statements from biological datasets with implicit causal information.
- Several formats have been developed for the storage of causal statements, with different ranges of contextual information.
- Improved interoperability between resources is dearly needed to facilitate the usability of the available data.

Acknowledgements

The authors would like to thank Charles Tapley Hoyt for providing feedback on the manuscript.

Funding

This study is supported by the Norwegian University of Science and Technology's Strategic Research Area "NTNU Health" (V.T., Å.F.), the ERACoSysMed grant COLOSYS (V.T., M.K.), and the Gene Regulation Ensemble Effort for the Knowledge Commons CA15205 (M.K.).

References

1. Bleske-Rechek A, Morrison KM, Heidtke LD. Causal Inference from Descriptions of Experimental and Non-Experimental Research: Public Understanding of Correlation-Versus-Causation. J. Gen. Psychol. 2015; 142:48–70
2. David, Hume. A Treatise of Human Nature. 1740;
3. Touré V, Vercruysse S, Acencio ML, et al. The Minimum Information about a Molecular Interaction Causal Statement (MI2CAST). Preprints 2020;
4. Wang R-S, Saadatpour A, Albert R. Boolean modeling in systems biology: an overview of methodology and applications. Phys. Biol. 2012; 9:055001
5. Naldi A, Berenguier D, Fauré A, et al. Logical modelling of regulatory networks with GINsim 2.3. Biosystems 2009; 97:134–139

6. Thomas R, Kaufman M. Multistationarity, the basis of cell differentiation and memory. II. Logical analysis of regulatory networks in terms of feedback circuits. Chaos Interdiscip. J. Nonlinear Sci. 2001; 11:180–195
7. Glass L, Kauffman SA. The logical analysis of continuous, non-linear biochemical control networks. J. Theor. Biol. 1973; 39:103–129
8. Collombet S, Oevelen C van, Ortega JLS, et al. Logical modeling of lymphoid and myeloid cell specification and transdifferentiation. Proc. Natl. Acad. Sci. 2017; 114:5792–5799
9. Rodríguez-Jorge O, Kempis-Calanis LA, Abou-Jaoudé W, et al. Cooperation between T cell receptor and Toll-like receptor 5 signaling for CD4+ T cell activation. Sci. Signal. 2019; 12:
10. Mendoza L. A network model for the control of the differentiation process in Th cells. Biosystems 2006; 84:101–114
11. Zhang R, Shah MV, Yang J, et al. Network model of survival signaling in large granular lymphocyte leukemia. Proc. Natl. Acad. Sci. 2008; 105:16308–16313
12. Flobak Å, Baudot A, Remy E, et al. Discovery of Drug Synergies in Gastric Cancer Cells Predicted by Logical Modeling. PLOS Comput. Biol. 2015; 11:e1004426
13. Vogt T, Czauderna T, Schreiber F. Translation of SBGN maps: Process Description to Activity Flow. BMC Syst. Biol. 2013; 7:115
14. Aghamiri SS, Singh V, Naldi A, et al. Automated inference of Boolean models from molecular interaction maps using CaSQ. Bioinformatics 2020;
15. Novère NL, Hucka M, Mi H, et al. The Systems Biology Graphical Notation. Nat. Biotechnol. 2009; 27:735–741
16. Perfetto L, Briganti L, Calderone A, et al. SIGNOR: a database of causal relationships between biological entities. Nucleic Acids Res. 2016; 44:D548-554
17. Licata L, Lo Surdo P, Iannuccelli M, et al. SIGNOR 2.0, the SIGnaling Network Open Resource 2.0: 2019 update. Nucleic Acids Res. 2020; 48:D504–D510
18. Perfetto L, Acencio ML, Bradley G, et al. CausalTAB: the PSI-MITAB 2.8 updated format for signalling data representation and dissemination. Bioinformatics 2019; 35:3779–3785
19. Lo Surdo P, Calderone A, Iannuccelli M, et al. DISNOR: a disease network open resource. Nucleic Acids Res. 2018; 46:D527–D534
20. Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res. 2020; 48:D845–D855
21. Fazekas D, Koltai M, Türei D, et al. SignaLink 2 – a signaling pathway resource with multi-layered regulatory networks. BMC Syst. Biol. 2013; 7:7
22. del-Toro N, Dumousseau M, Orchard S, et al. A new reference implementation of the PSICQUIC web service. Nucleic Acids Res. 2013; 41:W601–W606
23. Demir E, Cary MP, Paley S, et al. The BioPAX community standard for pathway data sharing. Nat. Biotechnol. 2010; 28:935–942
24. Hucka M, Finney A, Sauro HM, et al. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics 2003; 19:524–531
25. Boué S, Talikka M, Westra JW, et al. Causal biological network database: a comprehensive platform of causal biological network models focused on the pulmonary and vascular systems. Database J. Biol. Databases Curation 2015; 2015:
26. Schlage WK, Westra JW, Gebel S, et al. A computable cellular stress network model for non-diseased pulmonary and cardiovascular tissue. BMC Syst. Biol. 2011; 5:168
27. S Park J. Construction of a Computable Network Model of Tissue Repair and Angiogenesis in

the Lung. J. Clin. Toxicol. 2013; s12:
28. De León H, Boué S, Schlage WK, et al. A vascular biology network model focused on inflammatory processes to investigate atherogenesis and plaque instability. J. Transl. Med. 2014; 12:185
29. Gebel S, Lichtner RB, Frushour B, et al. Construction of a Computable Network Model for DNA Damage, Autophagy, Cell Death, and Senescence. Bioinforma. Biol. Insights 2013; 7:BBI.S11154
30. Westra JW, Schlage WK, Frushour BP, et al. Construction of a computable cell proliferation network focused on non-diseased lung cells. BMC Syst. Biol. 2011; 5:105
31. Westra JW, Schlage WK, Hengstermann A, et al. A Modular Cell-Type Focused Inflammatory Process Network Model for Non-Diseased Pulmonary Tissue. Bioinforma. Biol. Insights 2013; 7:BBI.S11509
32. Hoyt CT, Konotopez A, Ebeling C. PyBEL: a computational framework for Biological Expression Language. Bioinformatics 2018; 34:703–704
33. Thomas PD, Hill DP, Mi H, et al. Gene Ontology Causal Activity Modeling (GO-CAM) moves beyond GO annotations to structured descriptions of biological functions and systems. Nat. Genet. 2019; 51:1429–1433
34. Ashburner M, Ball CA, Blake JA, et al. Gene Ontology: tool for the unification of biology. Nat. Genet. 2000; 25:25–29
35. Helikar T, Kowal B, McClenathan S, et al. The Cell Collective: Toward an open and collaborative approach to systems biology. BMC Syst. Biol. 2012; 6:96
36. Chaouiya C, Bérenguier D, Keating SM, et al. SBML qualitative models: a model representation format and infrastructure to foster interactions between qualitative modelling formalisms and tools. BMC Syst. Biol. 2013; 7:135
37. Rougny A, Touré V, Moodie S, et al. Systems Biology Graphical Notation: Process Description language Level 1 Version 2.0. J. Integr. Bioinforma. 2019; 16:
38. Kanehisa M, Furumichi M, Tanabe M, et al. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017; 45:D353–D361
39. Jassal B, Matthews L, Viteri G, et al. The reactome pathway knowledgebase. Nucleic Acids Res. 2020; 48:D498–D503
40. Kuperstein I, Bonnet E, Nguyen H-A, et al. Atlas of Cancer Signalling Network: a systems biology resource for integrative analysis of cancer data with Google Maps. Oncogenesis 2015; 4:e160
41. Kutmon M, Riutta A, Nunes N, et al. WikiPathways: capturing the full diversity of pathway knowledge. Nucleic Acids Res. 2016; 44:D488–D494
42. Slenter DN, Kutmon M, Hanspers K, et al. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Res. 2018; 46:D661–D667
43. Ostaszewski M, Gebel S, Kuperstein I, et al. Community-driven roadmap for integrated disease maps. Brief. Bioinform. 2018; 20:659–670
44. Mazein A, Ostaszewski M, Kuperstein I, et al. Systems medicine disease maps: community-driven comprehensive representation of disease mechanisms. Npj Syst. Biol. Appl. 2018; 4:21
45. Ostaszewski M, Mazein A, Gillespie ME, et al. COVID-19 Disease Map, building a computational repository of SARS-CoV-2 virus-host interaction mechanisms. Sci. Data 2020; 7:136
46. Fujita KA, Ostaszewski M, Matsuoka Y, et al. Integrating Pathways of Parkinson's Disease in a Molecular Interaction Map. Mol. Neurobiol. 2014; 49:88–102

47. Mazein A, Knowles RG, Adcock I, et al. AsthmaMap: An expert-driven computational representation of disease mechanisms. Clin. Exp. Allergy 2018; 48:916–918
48. Singh V, Ostaszewski M, Kalliolias GD, et al. Computational Systems Biology Approach for the Study of Rheumatoid Arthritis: From a Molecular Map to a Dynamical Model. Genomics Comput. Biol. 2018; 4:
49. Singh V, Kalliolias GD, Ostaszewski M, et al. RA-map: building a state-of-the-art interactive knowledge base for rheumatoid arthritis. Database 2020; 2020:
50. Serhan CN, Gupta S, Perretti M, et al. The Atlas of Inflammation-Resolution (AIR). bioRxiv 2020; 2020.01.27.921882
51. Zhou S, Appleman VA, Rose CM, et al. Chronic platelet-derived growth factor receptor signaling exerts control over initiation of protein translation in glioma. Life Sci. Alliance 2018; 1:e201800029
52. Wentker P, Eberhardt M, Dreyer FS, et al. An Interactive Macrophage Signal Transduction Map Facilitates Comparative Analyses of High-Throughput Data. J. Immunol. 2017; 198:2191–2201
53. Jagannadham J, Jaiswal HK, Agrawal S, et al. Comprehensive Map of Molecules Implicated in Obesity. PLOS ONE 2016; 11:e0146759
54. Tripathi S, Flobak Å, Chawla K, et al. The gastrin and cholecystokinin receptors mediated signaling network: a scaffold for data analysis and new hypotheses on regulatory mechanisms. BMC Syst. Biol. 2015; 9:40
55. Tortolina L, Duffy DJ, Maffei M, et al. Advances in dynamic modeling of colorectal cancer signaling-network regions, a path toward targeted therapies. Oncotarget 2015; 6:5041–5058
56. Mizuno S, Iijima R, Ogishima S, et al. AlzPathway: a comprehensive map of signaling pathways of Alzheimer's disease. BMC Syst. Biol. 2012; 6:52
57. Funahashi A, Morohashi M, Kitano H, et al. CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. BIOSILICO 2003; 1:159–162
58. McGuinness DL, Van Harmelen F. OWL web ontology language overview. W3C Recomm. 2004; 10:2004
59. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res. 2019; 47:D330–D338
60. Türei D, Korcsmáros T, Saez-Rodriguez J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. Nat. Methods 2016; 13:966–967
61. Ceccarelli F, Turei D, Gabor A, et al. Bringing data from curated pathway resources to Cytoscape with OmniPath. Bioinformatics 2020; 36:2632–2633
62. Paz A, Brownstein Z, Ber Y, et al. SPIKE: a database of highly curated human signaling pathways. Nucleic Acids Res. 2011; 39:D793–D799
63. Antezana E, Blondé W, Egaña M, et al. BioGateway: a semantic systems biology tool for the life sciences. BMC Bioinformatics 2009; 10:S11
64. Durinx C, McEntyre J, Appel R, et al. Identifying ELIXIR Core Data Resources. F1000Research 2017; 5:2422
65. Holmås S, Riudavets Puig R, Acencio ML, et al. The Cytoscape BioGateway App: explorative network building from an RDF store. Bioinformatics 2020; 36:1966–1967
66. Pratt D, Chen J, Welker D, et al. NDEx, the Network Data Exchange. Cell Syst. 2015; 1:302–305
67. Rodchenkov I, Babur O, Luna A, et al. Pathway Commons 2019 Update: integration, analysis and exploration of pathway data. Nucleic Acids Res. 2020; 48:D489–D497

68. Domingo-Fernández D, Mubeen S, Marín-Llaó J, et al. PathMe: merging and exploring mechanistic pathway knowledge. BMC Bioinformatics 2019; 20:243
69. Hoyt CT, Domingo-Fernández D, Mubeen S, et al. Integration of Structured Biological Data Sources using Biological Expression Language. bioRxiv 2019; 631812
70. Todorov PV, Gyori BM, Bachman JA, et al. INDRA-IPM: interactive pathway modeling using natural language with automated assembly. Bioinformatics 2019; 35:4501–4503
71. Liu A, Trairatphisan P, Gjerga E, et al. From expression footprints to causal pathways: contextualizing large signaling networks with CARNIVAL. Npj Syst. Biol. Appl. 2019; 5:1–10
72. Bradley G, Barrett SJ. CausalR: extracting mechanistic sense from genome scale data. Bioinformatics 2017; 33:3670–3672
73. Nicolle R, Radvanyi F, Elati M. CoRegNet: reconstruction and integrated analysis of co-regulatory networks. Bioinformatics 2015; 31:3066–3068
74. Catlett NL, Bargnesi AJ, Ungerer S, et al. Reverse causal reasoning: applying qualitative causal knowledge to the interpretation of high-throughput data. BMC Bioinformatics 2013; 14:340
75. Margolin AA, Nemenman I, Basso K, et al. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. BMC Bioinformatics 2006; 7:S7
76. Martin F, Thomson TM, Sewer A, et al. Assessment of network perturbation amplitudes by applying high-throughput data to causal biological networks. BMC Syst. Biol. 2012; 6:54
77. Babur O, Luna A, Korkut A, et al. Causal interactions from proteomic profiles: molecular data meets pathway knowledge. bioRxiv 2018; 258855
78. Yates B, Braschi B, Gray KA, et al. Genenames.org: the HGNC and VGNC resources in 2017. Nucleic Acids Res. 2017; 45:D619–D625
79. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019; 47:D506–D515
80. Kohl M, Wiese S, Warscheid B. Cytoscape: Software for Visualization and Analysis of Biological Networks. Data Min. Proteomics Stand. Appl. 2011; 696:291–303
81. Deutsch EW, Orchard S, Binz P-A, et al. Proteomics Standards Initiative: Fifteen Years of Progress and Future Work. J. Proteome Res. 2017; 16:4288–4298
82. Hermjakob H, Montecchi-Palazzi L, Lewington C, et al. IntAct: an open source molecular interaction database. Nucleic Acids Res. 2004; 32:D452–D455
83. Kerrien S, Aranda B, Breuza L, et al. The IntAct molecular interaction database in 2012. Nucleic Acids Res. 2012; 40:D841–D846
84. Oughtred R, Stark C, Breitkreutz B-J, et al. The BioGRID interaction database: 2019 update. Nucleic Acids Res. 2019; 47:D529–D541
85. Breuer K, Foroushani AK, Laird MR, et al. InnateDB: systems biology of innate immunity and beyond—recent updates and continuing curation. Nucleic Acids Res. 2013; 41:D1228–D1233
86. Khanam Irin A, Tom Kodamullil A, Gündel M, et al. Computational Modelling Approaches on Epigenetic Factors in Neurodegenerative and Autoimmune Diseases and Their Mechanistic Analysis. J. Immunol. Res. 2015; 2015:e737168
87. Mungall C, Osumi-Sutherland D, Overton JA, et al. diatomsRcool. (2020, February 26). oborel/obo-relations: 2020-02-26 release (Version v2020-02-26). 2020;
88. Hucka M, Nickerson DP, Bader GD, et al. Promoting Coordinated Development of Community-Based Information Standards for Modeling in Biology: The COMBINE Initiative. Front. Bioeng. Biotechnol. 2015; 3:
89. Terfve C, Cokelaer T, Henriques D, et al. CellNOptR: a flexible toolkit to train protein signaling networks to data using multiple logic formalisms. BMC Syst. Biol. 2012; 6:133

90. Müssel C, Hopfensitz M, Kestler HA. BoolNet—an R package for generation, reconstruction and analysis of Boolean networks. Bioinformatics 2010; 26:1378–1380
91. Dilek A, Belviranli ME, Dogrusoz U. VISIBIOweb: visualization and layout services for BioPAX pathway models. Nucleic Acids Res. 2010; 38:W150–W154
92. Danos V, Feret J, Fontana W, et al. Rule-Based Modelling of Cellular Signalling. CONCUR 2007 – Concurr. Theory 2007; 17–41
93. Bachman JA, Gyori BM, Sorger PK. Assembling a phosphoproteomic knowledge base using ProtMapper to normalize phosphosite information from databases and text mining. bioRxiv 2019; 822668
94. Niarakis A, Kuiper M, Ostaszewski M, et al. Setting the basis of best practices and standards for curation and annotation of logical models in biology—highlights of the [BC]2 2019 CoLoMoTo/SysMod Workshop. Brief. Bioinform. 2020;
95. Malik-Sheriff RS, Glont M, Nguyen TVN, et al. BioModels—15 years of sharing computational models in life science. Nucleic Acids Res. 2020; 48:D407–D415