Peer-reviewed version available at *Bioinformatics* **2020**; doi:10.1093/bioinformatics/btaa622

The Minimum Information about a Molecular Interaction Causal Statement (MI2CAST)

Vasundra Touré[1,*], Steven Vercruysse[1], Marcio Luis Acencio[2], Ruth C Lovering[3], Sandra Orchard[4], Glyn Bradley[5], Cristina Casals-Casas[6], Claudine Chaouiya[7], Noemi del-Toro[4], Åsmund Flobak[2,8], Pascale Gaudet[9], Henning Hermjakob[4], Luana Licata[10], Astrid Lægreid[2], Christopher J. Mungall[11], Anne Niknejad[12], Simona Panni[13], Livia Perfetto[4], Pablo Porras[4], Dexter Pratt[14], Denis Thieffry[15], Paul D. Thomas[16], Dénes Türei[17], Julio Saez-Rodriguez[17], Martin Kuiper[1]

s

[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] Functional Gene Annotation, Preclinical and Fundamental Science, Institute of Cardiovascular Science, UCL, University College London, London, UK.

[4] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK.

[5] Computational Biology, Functional Genomics, GSK, Stevenage, UK.

[6] Swiss-Prot group, SIB Swiss Institute of Bioinformatics, 1 Michel Servet, 1211 Geneva 4, Switzerland.

[7] Aix Marseille Univ, CNRS, Centrale Marseille, I2M, Marseille, France.

[8] The Cancer Clinic, St. Olav's Hospital, Trondheim University Hospital, Trondheim, Norway.

[9] SIB Swiss Institute of Bioinformatics, Geneva, Switzerland.

[10] Department of Biology, University of Rome Tor Vergata, Via della Ricerca Scientifica, 00133 Rome, Italy.

[11] Division of Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA, USA.

[12] Vital-IT group, SIB Swiss Institute of Bioinformatics, Quartier Sorge, Amphipole building, 1015 Lausanne, Switzerland.

[13] University of Calabria, Dip. of Biology, Ecology and Earth Science, Via Pietro Bucci Cubo 6/C Rende (CS), Italy.

[14] Department of Medicine, University of California San Diego, La Jolla, CA 92093, USA.

[15] Institut de Biologie de l'ENS (IBENS), Département de biologie, École normale supérieure, CNRS, INSERM, Université PSL, 75005 Paris, France.

[16] Division of Bioinformatics, Department of Preventive Medicine, University of Southern California, Los Angeles, CA USA.

[17] Institute of Computational Biomedicine, Heidelberg University, Faculty of Medicine, 69120 Heidelberg, Germany.

Abstract

A large variety of molecular interactions occurs between biomolecular components in cells. When one or a cascade of molecular interactions results in a regulatory effect, by one component onto a downstream component, a so-called 'causal interaction' takes place. Causal interactions constitute the building blocks in our understanding of larger regulatory networks in cells. These causal interactions and the biological processes they enable (e.g., gene regulation) need to be described with a careful appreciation of molecular interactions that occur between entities. A proper description of this information enables archiving, sharing, and reuse by humans and for computational science. Various representations of causal relationships between biological components are currently used in a variety of resources. Here, we propose a checklist that accommodates current representations, and call it the Minimum Information about a Molecular Interaction CAusal STatement (MI2CAST). This checklist defines both the required core information, as well as a comprehensive set of other contextual details valuable to the end user and relevant for reusing and reproducing causal molecular interaction information. The MI2CAST checklist can be used as reporting guidelines when annotating and curating causal statements, while assuring uniformity and interoperability of the data across resources.

Keywords: causal statement, causal interaction, directed molecular interaction, minimum information, standardization, systems biology.

Background

Causal interactions describe interacting biomolecules involved in processes where the state of one biomolecule is affected by the state of another biomolecule, resulting in a behavioral change of the system. A formal description of such causal interactions is referred to as a *causal statement*. A causal statement describes a binary interaction between two biological entities (e.g., gene, protein, RNA), where, given a certain context, the source entity (i.e., the regulator) influences the activity (either directly or by affecting the quantity) of a target entity, which itself may have an altered influence over further downstream targets. For instance, the protein LYN phosphorylates PTPN6 at the C-terminal Tyr-564 site, stimulating PTPN6's tyrosine phosphatase activity [1]. In other words: the kinase activity of LYN (source entity in an active state) can cause an increase in the phosphatase activity of PTPN6 (change of state of target entity). Additional aspects relating to the why, when and where of the causal interaction are important elements that together capture the context in which this causal interaction occurs (e.g., taxon, cell type, experimental condition).

The Proteomics Standards Initiative Molecular Interaction (PSI-MI) community was initially driven by the need to curate undirected molecular interactions [2,3]. Yet since most physical interactions are known to be involved in regulatory processes, several knowledge base resources started to collect causal interactions by incorporating directionality information as well [4–6]. Therefore, the PSI-MI standard has been extended to also represent the causality of interactions

through a direction and sign (up- or down-regulation) [7]. The extraction and annotation of causal interactions are predominantly performed via detailed manual curation of scientific publications [4]; but as techniques to infer causality through natural language processing [8] or 'omics data using prior knowledge [9–11] are also maturing, their results should also be supplied with the essential context details. Current formats of causal statements range from the simplest, with only two entities and the causal regulatory effect (e.g., the Simple Interaction Format (SIF) with "A activates B" or "A -> B"), to more complex statements including contextual description (e.g., BEL (Biological Expression Language, https://bel.bio/) [12], GO-CAM [13], and PSI-MITAB2.8 [7]). At present, various resources host molecular causal relationships (e.g., IntAct [14], SIGNOR [4,15], Causal Biological Network [16], SignaLink [5], TRRUST [17], TFacTS [18], DoRothEA [19]), each adhering to some of the formats mentioned above and annotated with specific controlled vocabularies (CVs) or ontologies (PSI-MI CV, Gene Ontology [20]). However, the contextual information provided in different resources can be depicted using different nomenclatures, or be incomplete or inconsistent, resulting in incompatibilities or conflicting information that hinders data integration and can complicate network building [6]. For example, entity A can be annotated to activate entity B in one database and inhibit entity B in another. Causal statements expressing these seemingly conflicting events are not necessarily incorrect, provided that there is sufficient context description to distinguish when each case occurs. A first step to improve the description of these interactions and their regulatory context is to standardize the different pieces of information and assemble them in a checklist. By adequately annotating and archiving the necessary and sufficient details, causal interactions can be efficiently shared and processed with computers (e.g., for regulatory network assembly) and humans alike (e.g., for designing experiments).

In response to the "reproducibility crisis" in science [21], novel projects focus on setting up formal structures for data management with collaborations between domain experts [22–24]. For instance, the description of molecular interactions has been formalized by the Human Proteome Organization (HUPO) PSI-MI community [25], leading to standard guidelines (MIMIx [26]), exchange formats (PSI-MI TAB [7], PSI-MI XML [27]), and CVs (PSI-MI CV [28]). These standards are adopted by biological databases (e.g., IntAct [14], the IMEx Consortium [29], SIGNOR [4,15], Reactome [30]), and researchers are called upon to describe their data following these standards. Developing a standardized framework for specific fields increases interoperability between resources [23,31] and helps to improve data findability, reuse and, reproducibility. Ontologies and CVs foster unambiguous semantics for the data, underpinned by unique identifiers (e.g., the Gene Ontology [20,32]), as their terms are used in annotation processes to attribute information to biological entities. Checklists with contextual details to be included in the description of data have been developed (e.g., MIAME [33], MIMIx [26]) and form a fundamental basis for the development of guidelines [34]. When semantics and checklists have been agreed upon, standard formats can be built for syntactic support, enabling the storage and exchange of information. The corresponding annotation guidelines advise the curators on the steps and necessary fields to complete in order to deliver valuable data. Finally, tools ranging from annotation tools to third-party software that can read, write and validate files, endorse these guidelines and formats.

We define here the Minimum Information about a Molecular Interaction CAusal STatement

(MI2CAST), as a foundation for a formal, consistent and intelligible data capture of causal interactions in molecular biology. It is developed to accommodate the needs of a data user, while taking into account the practical experience from biological curators. MI2CAST considers terms used in formats mentioned previously (e.g., PSI-MITAB2.8, BEL, GO-CAM) and covers the full range of metadata that should ideally be annotated during the curation process to enrich the description of a molecular causal interaction. MI2CAST checklist advises: 1) the molecular biologists to experimentally assess and describe a list of criteria, when conducting experiments, necessary to contextualize causal interactions; 2) the curators to consider and extract a list of metadata while curating causal interactions; 3) the data consumers to access persistent information and fully contextualized data to be able to select causal statements that comply with the system analyzed in their case study. These guidelines do not dictate the format in which one should represent causality, but rather guide on concepts that should be archived together with the causal interaction. Complying with these guidelines should be considered as good practice for the annotation of causal statements to generate high quality statements.

The Minimum Information about a Molecular Interaction Causal Statement (MI2CAST)

The MI2CAST checklist structures the information describing the causality resulting from a molecular interaction (Figure 1). There is not one single way to represent causal statements, but different alternatives should share a core of mutually compliant information. Their representation depends on the research interest, the available knowledge, and specific use cases. For instance, the molecular biologist might be interested in the fine details of the mechanistic events that lead to the expression of a gene (e.g., epigenetic modifications), while a modeler may be interested only in the resulting activation changes (e.g., signaling cascade of interactions between proteins) in addition to context that enables to assess the strength of the evidence. The purpose of MI2CAST is to support and increase compatibility of these different representations. In addition, there is a minimum level of context description that seems essential for any subsequent reuse of annotated causal interactions. The MI2CAST guidelines lay out these annotation tasks in four rules covering different aspects of a causal interaction (Figure 1). Each rule specifies terms corresponding to the metadata to annotate, for which recommendations on ontologies and CVs to use are included. Rule 1, Rule 2 and Rule 3 cover the most essential information, while Rule 4 recommends annotation of additional and optional details that increases the information content of a causal statement. Note that different instances of a causal interaction should be provided when the context is different, even if the involved entities are the same.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 April 2020 doi:10.20944/preprints202004.0480.v1

Peer-reviewed version available at *Bioinformatics* **2020**; doi:10.1093/bioinformatics/btaa622

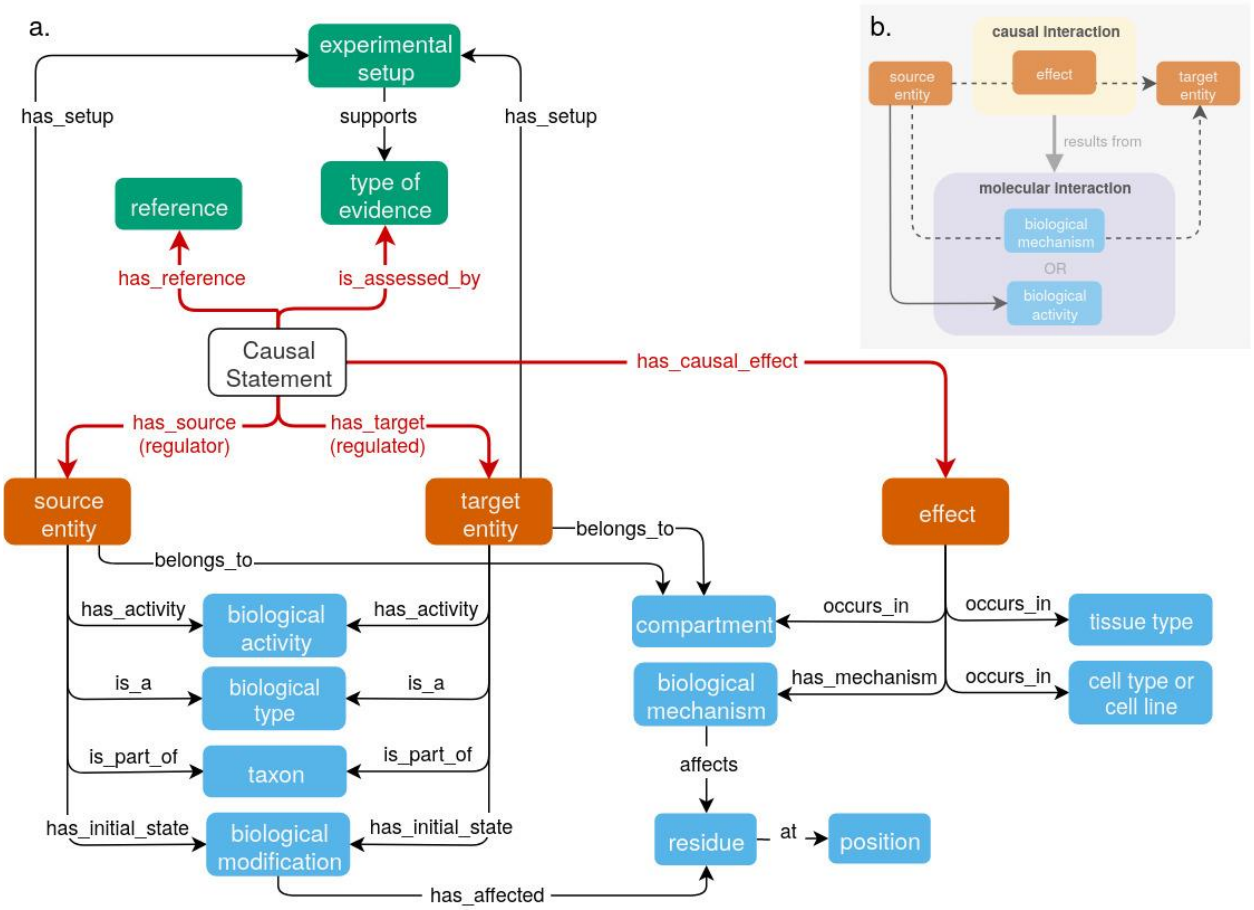

Figure 1. Data structure diagram documenting the causal statement terms and their relationships.
a. Red arrows represent the minimal and mandatory annotation about a causal statement: the source entity, the target entity and the effect of the causal interaction (orange boxes belonging to Rules 1 and 2), as well as the provenance of the causal statement (green boxes belonging to Rule 3). The black arrows correspond to useful but optional annotations about the entities or effect (the blue boxes belonging to Rule 4). b. A causal effect, or 'causal interaction', between two entities is the result of an associated 'molecular interaction' between them, which is specified through either a mechanism or the activity of the source entity (see Rule 4.1).

The MI2CAST guidelines are structured into four rules.

Rule 1: The source and target entities must be specified

All molecular interaction causal statements must provide reference identifiers of at least a source entity and a target entity. The source entity corresponds to the regulatory element of a causal statement and controls the state (activity or quantity) of the target entity. The target entity corresponds to the regulated element of a causal statement and is controlled by the source entity. The direction of the interaction is implicit: the molecular state change is exerted by the source entity and affects the target entity. For a causal interaction to occur, it is therefore assumed that the annotated context about the source entity (see Rule 4 below) is required for the target entity to be affected.

The source and target can be any molecular entity, for instance a protein, although in reality, molecular entities may not always refer to individual physical entities but rather to populations of individuals of a specific class of molecules: when it is stated that A regulates B in context C, it is actually a population of A that regulates the size or activity of the population of B, in context C.

In addition, we offer basic support for other entities than biomolecules as well. This enables annotation of the effect that a biomolecule exerts on an observable phenomenon (e.g., a phenotype like DNA repair or apoptosis), or vice versa. Causally relating a biomolecule and a phenotype is abundantly used in biology, as it enables: 1) to grasp knowledge about a process when the curator lacks information regarding downstream molecular events, 2) to more easily assess the final outcome of a signaling network (e.g., cell survives, apoptosis is activated) during analysis, and 3) to highlight relevant paths of information flow that may otherwise be challenging to extract from dense signaling networks. MI2CAST also specifies how to capture relevant context, but for molecular entities only; and its Rules 3-4 below apply only to biomolecules.

An exhaustive list of entity classes that can be part of causal statements is provided in Figure 2, together with recommendations of comprehensive and widely used ontologies and CVs to describe them. For instance, if the source entity is a known ‘mRNA’, it is recommended to use an ‘Ensembl transcript’ identifier. If the exact mRNA entity is not known, the ‘Ensembl gene’ identifier should be provided, and the “biological type” of the entity (see Rule 4.2 below) must be specified (e.g., ribonucleic acid, messenger rna). For chemicals that do not have a ChEBI identifier, a PubChem identifier would be an alternative. When the entity is a protein, it can often be present in different isoforms. If the isoform is known, it is recommended to provide the UniProtKB isoform accession number, otherwise the generic UniProt identifier. In addition, it is recommended to annotate a protein with a UniProtKB/Swiss-Prot reviewed identifier, when available, instead of a UniProtKB/TrEMBL (i.e., unreviewed) identifier. In the case of a ‘family’ (i.e., group of entities with similar functions, sequence or structure) or a ‘transient complex’ (i.e., group of entities that interact together temporarily [35]), the list of individual entities should be provided (e.g., if a complex has proteins as components, one should provide UniProtKB identifiers for the components of the complex). To be able to distinguish between a complex and a family, the “biological type” of the entity must be provided (see Rule 4.2 below). The phenotype is a distinct type of entity that refers to biological processes associated with molecular events (e.g., TP53 activates apoptosis). This list does not preclude the use of other identifiers (see Supplementary File S2 for a more extensive list of identifiers), as long as appropriate ones are provided.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 April 2020 doi:10.20944/preprints202004.0480.v1

Peer-reviewed version available at Bioinformatics 2020; doi:10.1093/bioinformatics/btaa622

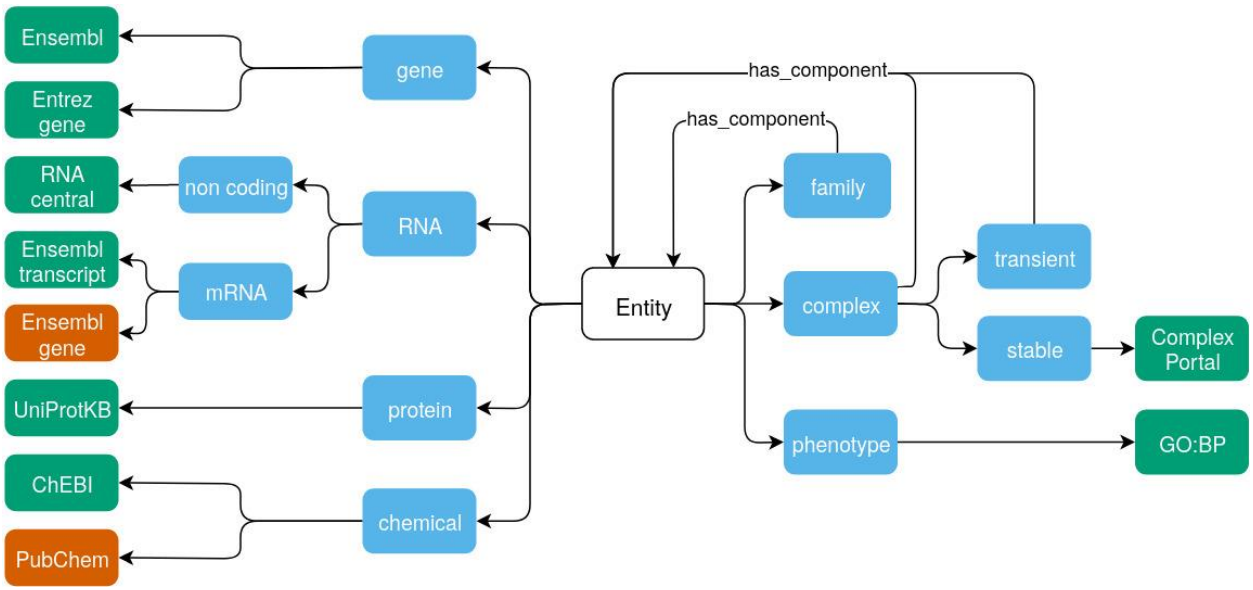

Figure 2. Diagram of entity types and related databases for identifier origin.
Blue boxes show the different entity types; green boxes, primarily recommended databases; and orange boxes, alternative databases. For each entity type, identifiers from specific databases are recommended.

Rule 2: The effect of the interaction must be specified

All causal statements must provide the regulatory effect of the molecular interaction. This effect describes the causal nature of the interaction between the source and the target. It should preferably include the regulatory outcome exerted by the source on the target, i.e., positive or negative, if known. It can also specify whether the interaction is direct or indirect. A direct interaction involves physical contact between the entities. An indirect interaction implies that source and target are not necessarily in direct contact; e.g., the causation could be mediated by intermediate steps that are not spelled out. For instance, when a transcription factor positively regulates a protein via transcription, it is an indirect interaction because the transcription factor acts on a gene in order to produce the protein.
In general, the effect implies an increase or decrease in a particular activity of the target, which will affect the process that this target is involved in. It is recommended that one of the following ontologies and CVs are used to annotate the effect information:

- the “causal statement” branch of PSI-MI (purl.obolibrary.org/obo/MI_2233),
- the “causally related to” branch of the Relation Ontology (purl.obolibrary.org/obo/RO_0002410).

When evidence or knowledge of a physical association mechanism is available, or the regulatory outcome is known, one should use a lower-level (i.e., more specific) term from the ontology or CV.

Rule 3: The provenance and evidence types of the annotation must be specified

A basic rule of any annotation is to keep track of provenance (i.e., reference to scientific reports),

as it allows consumers of a causal statement to check the quality of an annotation, and the supporting evidence. This evidence may either be curated from high-confidence biological or other assays, or from computational analysis. Provenance and evidence type affect trust that a user may have, and influence the decision of incorporating a causal statement in a model. For instance, one may consider a manually curated statement more valuable and trustworthy than an automatically generated one because of the errors that may be associated with computational inference and text-mining extraction of causal interactions [36].

3.1 The reference

When a causal statement is curated manually from an experiment, it is always extracted from a description of that experiment, usually a publication, so the reference to that publication or other source must be provided. If a combination of several articles has led to the finding of a causal interaction, then the full list of these publications must be provided. Each of the publications in the list provides a necessary but not sufficient part of the evidence, and the full list is a minimal set of articles that provide sufficient evidence to support the causal interaction. For instance, if an article reports a causal effect between two molecules, and a second one reports that they bind, then one may infer from both that there is a causal effect that is direct. Statement trustworthiness should not be assessed by counting references within a statement. Still, multiple statements expressing the same causal interaction, but each with different reference(s), could make it more trustworthy.

It is recommended that the PubMed identifier is used for the article(s) curated. A Digital Object Identifier (DOI) can also be provided in case of articles not referenced in MEDLINE, e.g., to refer to manuscripts available in preprint servers.

3.2 The type of evidence

The type of evidence for declaring and annotating a causal statement must be provided. This information corresponds to the experimental or other data on which the causal interaction is based. The causal statement may be electronically inferred (e.g., through text mining or in silico study), observed during a certain experiment (in which case *in vivo* and *in vitro* studies can be specified), just mentioned (author statement) in a paper, or based on combinatorial evidence as in Rule 3.1 (e.g., a causal interaction assessed from the necessary combination of an author statement and the results of an experiment). In the latter case, multiple evidence identifiers can be recorded. When the evidence type is an experiment, the annotation can either be at the less specific level (e.g., experimental evidence, ECO:0000006) or as specific as possible (e.g., a yeast 2-hybrid evidence used in manual assertion, ECO:0005805). The type of evidence should be specified with terms from the Evidence & Conclusion Ontology (ECO [37]).

3.2.1 The experimental setup

If the type of evidence is an experiment (Rule 3.2), then it may help to record particular experimental conditions that support the experiment, in order to enable users to select causal statements that meet a particular confidence level. An experimental setup specifies the design

of any of the two entities. For example, with a reporter gene assay annotated as the type of evidence, the following annotations could support this experiment type: the source entity was overexpressed, and both the source and the target entities were engineered (MI:0331). The following recommended ontology/CV should be used to capture the experimental condition:

- the Evidence & Conclusion Ontology (ECO [37]),
- the PSI-MI "experimental preparation" branch (purl.obolibrary.org/obo/MI_0346).

Rule 4: The defining contextual details should be specified

While causal statements as they are defined above are already useful for building mechanical models, their relevance becomes even greater when they indicate the experimental context of the corresponding observation. If applied to any other context, the possibility exists that the causal interaction does in fact not occur in this other context. Defining the contextual details may also help to disambiguate statements that would otherwise appear to be conflicting, because the presence of a causal effect in a given interaction can vary depending on the context. This information benefits data users, who may need to select relevant causal interactions valid for specific conditions. The better the contextual information is, the lower the chance that the causal statement is taken as generally valid and wrongly applied. All molecular interaction causal statements should therefore provide the contextual details that are essential to define the specific circumstances in which the causality has been observed (e.g., interaction observed in a particular cell type). The context can be attributed to the source entity, the target entity, or the interaction itself. The relevance of context being very much dependent on the biological aspects, it is not necessary to apply every single contextual term for all causal interactions. The conditions under which the context and particulars are optional or required to be annotated are described in the following sections (Rules 4.1 to 4.5).

4.1 The biological activity of an entity, or the mechanism of an interaction

If the causal statement relates to a specific biological activity of an entity (source and/or target), then the biological activity of that entity should be provided. A biological activity corresponds to a biological event, behaviour or function of a molecular entity that is involved in the causal interaction, as for instance the two activities in: "A, having kinase activity, regulates B, having DNA binding transcription factor activity". This information enables translation from an entity-based view (used in causal statements) to an activity-based view (used in GO-CAMs [13]). We recommend using:

- the Gene Ontology Molecular Function terms for proteins and RNA gene products [20],
- the ChEBI "role" branch (purl.obolibrary.org/obo/CHEBI_50906) for chemicals (for roles that correspond to a particular activity that the chemical has, e.g., "catalyst"),
- the Sequence Ontology for genes [38] (for roles of gene features that can be causally affected, e.g., "binding_site").

If biological activities are not available, then the mechanism of the causal interaction should be provided, if known. The mechanism describes how the source exerts a biological effect on the

target, for instance through a transcriptional regulation or through a phosphorylation reaction. We recommend using:

- the PSI-MI "causal regulatory mechanism" branch (purl.obolibrary.org/obo/MI_2245),
- the PSI-MI "interaction type" branch (purl.obolibrary.org/obo/MI_0190),
- the Gene Ontology Biological Process (GO:BP) branch.

Note that biological activity of the source entity corresponds to the mechanism of the interaction. Either choice is just a syntactic variation of similar information. For example, "A, having kinase activity (GO:0016301), regulates B" corresponds to "A regulates, through a phosphorylation reaction (MI:0217), B". The biological activity of the target entity specifies what function is affected as a result of activity of the source or mechanism.

However, annotating only the biological mechanism of an interaction does not necessarily properly describe its impact on any particular biological activity of the target entity. Therefore, when the biological mechanism results in a modification (or state-change) of the target (e.g., a phosphorylation event), it is recommended to annotate as precisely as possible how it modifies the target (e.g., with both residue type and position) so as to capture information about the state of the target entity that results from the causal regulation (see also 4.3). Precise description of the target entity's resulting state supports inference of what biological activity may be affected.

Note that biological activity is required in GO-CAMs (which center around activities, as in "biological activity X of entity A regulates biological activity Y of entity B"), while it is optional in causal statements (which center around entities, as in "entity A regulates entity B"). In causal statements, biological activity or mechanism may not always be known. Still, in order to enable a better assessment of the effects of a causal interaction, on the molecular interaction level, we recommend that the activity or mechanism is always annotated when this knowledge is available.

4.2 The biological type of an entity

In MI2CAST, the biological type of an entity corresponds to its biological nature, such as gene, RNA, protein, complex. The biological type of an entity is usually defined indirectly, by the identifier provided by the database that aims to list all entities of a certain type. In most cases, it is therefore not needed to further define the biological type. For instance, a UniProt identifier classifies an entity as a protein.

In some cases, however, the biological entity involved in a causal interaction may not yet have a unique identifier assigned to it (see preferred database IDs, Rule 1). A pragmatic solution is to search another database for a gene or gene product that is related to the intended entity, as this will at least allow the use of an identifier rather than make no annotation at all. For example, in the absence of a corresponding UniProt ID, an Entrez ID could be used to annotate a protein. In these cases, the correct and intended biological type of the entity must be provided. In the example, the Entrez ID would have to be accompanied by "has biological type: protein" to clarify that actually the associated gene product is meant. Likewise, when a complex is not referenced in the Complex Portal database [39], it can be specified as a general entity that has a list of components, but should then be annotated with the 'complex' (MI:0314) biological type. For

biological type, we recommend to use the terms provided by the PSI-MI interactor type branch (purl.obolibrary.org/obo/MI_0313).

4.3 The biological modification of an entity

A causality may depend on an entity (source and/or target) having a particular physical modification or conformation prior to its engagement in the causal interaction. Modifications include physical configurations (e.g, post-transcriptional modifications, post-translational modifications, covalent links to other molecules, methylations of genes) that lead to conformational changes (e.g, open, closed) necessary for a causal interaction to occur.
If the causality depends on an entity having a particular biological modification, then that state ideally is provided with as much precision as available, and represented by:

- a modification type (e.g., phosphorylation of a protein, methylation of a gene or RNA), specified by PSI-MOD for proteins [40], and the SO for genes,
- a modified residue (i.e., amino acid, nucleotide), for which we recommend using ChEBI,
- a number indicating the protein sequence position of the residue that is modified.

4.4 The taxon of an entity or interaction

For both the source and target entity, the taxon is usually defined through its identifier (e.g., UniProt ID, Ensembl ID). In the case of heterologous system assays, each entity needs to be annotated with its species of origin.
It may be useful for a data user to select causal statements based on taxon ID of the interaction as well. However, as MI2CAST focuses on knowledge that cannot be computationally inferred and needs to be captured by curation, only the entities' taxon information needs to be annotated. Of course, any data exchange format based on MI2CAST can still require the inclusion of the taxon at the interaction level. A taxon for the causal interaction as a whole would correspond to the organism in which the interaction has its 'native function'. For example, if the observed molecular interaction takes place between a source and target entity of the same taxon, then the causal interaction's taxon would be inferred as being the same. Alternatively, if a causality was observed via an assay in which source and target originate from different taxa (i.e., a heterologous assay), then based on entity homology the causal statement could be computationally inferred as to be valid for both taxa. The NCBI Taxonomy [41] is recommended to capture the taxon.

4.5 The location of an interaction or entity

The annotations of physical location specify the precise localization where a causal interaction was observed or where an entity was located. For instance, if two human genes were inserted into the genome of a yeast cell (i.e., a heterologous system) and were observed to be involved in a causal interaction, then the concept of 'yeast' would be conveyed via a location term (i.e., cell type, or cell line). We define different levels of locational definitions, from the highest level being the tissue (Rule 4.5.1) to the most detailed level being the cellular compartment (Rule 4.5.3).

4.5.1 The tissue type of an interaction

If the tissue type in which the causal interaction has been observed is known, an established ontology identifier should be provided. BRENDA [42] or Uberon [43] are recommended to capture the tissue type for metazoans, the Plant Ontology [44] for plants, and the Fungal Anatomy Ontology (purl.obolibrary.org/obo/fao.owl) for fungi.

4.5.2 The cell type or cell line of an interaction

If the cell type or cell line in which the causal interaction occurs is known, it should be provided. The Cell Ontology [45,46] or BRENDA [42] are recommended to capture the cell type. The Cellosaurus [47] or BRENDA [42] are recommended to capture the cell line.

4.5.3 The compartment of an interaction or entity

If the causal interaction is specifically observed in a particular cellular compartment, this should be annotated. The compartment corresponds to the cellular localization where the causal interaction takes place. The interaction can involve multiple compartments (e.g., transport of entities). A compartment can be annotated at the interaction level or at the entity level, in cases where entities are located in different compartments. When the causal statement describes the translocation of a target entity into another compartment, the entity's original location should be annotated. The entity's new location could be conveyed by a translocation mechanism term (Rule 4.1; e.g., 'import into nucleus' (GO:0051170)). The terms provided by the Gene Ontology Cellular Component (GO:CC [20]) are recommended for cellular location annotations.

Conclusion

MI2CAST describes the Minimum Information about a Molecular Interaction CAusal STatement, consisting of a checklist of terms and identifiers recommended for annotations. It takes the form of a set of rules that serve as annotation guidelines. A causal interaction consists of compulsory information on the source entity, the target entity (Rule 1) and the regulation (Rule 2). The evidence supporting a causal interaction and its provenance (Rule 3) must also be reported. Annotations describing the defining context of a causal interaction (Rule 4) specify the conditions under which a causal interaction has been observed, together with more detailed information regarding its source entity, target entity and causal effect.

The MI2CAST guidelines have been developed in close collaboration with the GREEKC consortium (http://greekc.org/) and HUPO Proteomics Standards Initiative (HUPO-PSI) molecular interaction workgroup. PSI-MITAB2.8 has been specifically designed to hold MI2CAST-compliant data, enabling the capture of both sign and causality of an interaction. Interestingly, the SIGNOR database already compiles data pertaining to causal relationships between biological entities available in the PSI-MITAB2.8 format. Users will be able to access and merge these data using the MITAB2.8-compliant PSICQUIC webservice [48]. The addition of new terms to relevant CVs, such as PSI-MI and Sequence Ontology, as part of the development of these data standards, will enable a fuller description of the biological activity of an entity. MI2CAST remains dynamic, contingent on research insights, requests and the evolution of scientific discoveries in the field of molecular

and systems biology, and represents a next step in the global efforts to take care of valuable life science information. Future extensions could include the recording of logical operators and cells as valid entities for the annotation of cell-to-cell causal interactions (i.e., causality where neither entity is a biomolecule).

References

1. Yoshida K, Kharbanda S, Kufe D. Functional interaction between SHPTP1 and the Lyn tyrosine kinase in the apoptotic response to DNA damage. J. Biol. Chem. 1999; 274:34663–34668
2. Hermjakob H. The HUPO Proteomics Standards Initiative – Overcoming the Fragmentation of Proteomics Data. PROTEOMICS 2006; 6:34–38
3. 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
4. Perfetto L, Briganti L, Calderone A, et al. SIGNOR: a database of causal relationships between biological entities. Nucleic Acids Res. 2016; 44:D548-554
5. 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
6. Türei D, Korcsmáros T, Saez-Rodriguez J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. Nat. Methods 2016; 13:966–967
7. 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
8. Todorov PV, Gyori BM, Bachman JA, et al. INDRA-IPM: interactive pathway modeling using natural language with automated assembly. Bioinformatics 2019; 35:4501–4503
9. Babur O, Luna A, Korkut A, et al. Causal interactions from proteomic profiles: molecular data meets pathway knowledge. bioRxiv 2018; 258855
10. Chindelevitch L, Ziemek D, Enayetallah A, et al. Causal reasoning on biological networks: interpreting transcriptional changes. Bioinformatics 2012; 28:1114–1121
11. Bradley G, Barrett SJ. CausalR: extracting mechanistic sense from genome scale data. Bioinformatics 2017; 33:3670–3672
12. Hoyt CT, Konotopez A, Ebeling C. PyBEL: a computational framework for Biological Expression Language. Bioinformatics 2018; 34:703–704
13. 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
14. Orchard S, Ammari M, Aranda B, et al. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res. 2014; 42:D358–D363
15. Licata L, Lo Surdo P, Iannuccelli M, et al. SIGNOR 2.0, the SIGnaling Network Open Resource 2.0: 2019 update. Nucleic Acids Res.
16. 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:
17. Han H, Cho J-W, Lee S, et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. Nucleic Acids Res. 2018; 46:D380–D386

18. Essaghir A, Toffalini F, Knoops L, et al. Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data. Nucleic Acids Res. 2010; 38:e120–e120
19. Garcia-Alonso L, Holland CH, Ibrahim MM, et al. Benchmark and integration of resources for the estimation of human transcription factor activities. Genome Res. 2019; 29:1363–1375
20. Ashburner M, Ball CA, Blake JA, et al. Gene Ontology: tool for the unification of biology. Nat. Genet. 2000; 25:25–29
21. Baker M. 1,500 scientists lift the lid on reproducibility. Nat. News 2016; 533:452
22. Mayer G. Data management in systems biology I - Overview and bibliography. ArXiv09080411 CoRR 2009; abs/0908.0411.
23. Dräger A, Palsson BØ. Improving Collaboration by Standardization Efforts in Systems Biology. Front. Bioeng. Biotechnol. 2014; 2:
24. National Academies of Sciences E and Medicine. Reproducibility and Replicability in Science (Consensus Study Report) 2019; Washington, DC: The National Academies Press.
25. Hermjakob H, Montecchi-Palazzi L, Bader G, et al. The HUPO PSI's Molecular Interaction format—a community standard for the representation of protein interaction data. Nat. Biotechnol. 2004; 22:177–183
26. Orchard S, Salwinski L, Kerrien S, et al. The minimum information required for reporting a molecular interaction experiment (MIMIx). Nat. Biotechnol. 2007; 25:894–898
27. Sivade MD, Alonso-López D, Ammari M, et al. Encompassing new use cases - level 3.0 of the HUPO-PSI format for molecular interactions. BMC Bioinformatics 2018; 19:134–134
28. Orchard S, Montecchi-Palazzi L, Hermjakob H, et al. The use of common ontologies and controlled vocabularies to enable data exchange and deposition for complex proteomic experiments. Pac. Symp. Biocomput. Pac. Symp. Biocomput. 2005; 186–196
29. Orchard S, Kerrien S, Abbani S, et al. Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. Nat. Methods 2012; 9:345–350
30. Fabregat A, Jupe S, Matthews L, et al. The Reactome Pathway Knowledgebase. Nucleic Acids Res. 2018; 46:D649–D655
31. Stanford NJ, Wolstencroft K, Golebiewski M, et al. The evolution of standards and data management practices in systems biology. Mol. Syst. Biol. 2015; 11:
32. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res. 2019; 47:D330–D338
33. Brazma A, Hingamp P, Quackenbush J, et al. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Nat. Genet. 2001; 29:365
34. Taylor CF, Field D, Sansone S-A, et al. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. Nat. Biotechnol. 2008; 26:889–896
35. Acuner Ozbabacan SE, Engin HB, Gursoy A, et al. Transient protein–protein interactions. Protein Eng. Des. Sel. 2011; 24:635–648
36. Britan A, Cusin I, Hinard V, et al. Accelerating annotation of articles via automated approaches: evaluation of the neXtA5 curation-support tool by neXtProt. Database 2018; 2018:
37. Giglio M, Tauber R, Nadendla S, et al. ECO, the Evidence & Conclusion Ontology: community standard for evidence information. Nucleic Acids Res. 2019; 47:D1186–D1194
38. Eilbeck K, Lewis SE, Mungall CJ, et al. The Sequence Ontology: a tool for the unification of genome annotations. Genome Biol. 2005; 6:R44

39. Meldal BHM, Bye-A-Jee H, Gajdoš L, et al. Complex Portal 2018: extended content and enhanced visualization tools for macromolecular complexes. Nucleic Acids Res. 2019; 47:D550–D558
40. Montecchi-Palazzi L, Beavis R, Binz P-A, et al. The PSI-MOD community standard for representation of protein modification data. Nat. Biotechnol. 2008; 26:864–866
41. Federhen S. The NCBI Taxonomy database. Nucleic Acids Res. 2012; 40:D136–D143
42. Gremse M, Chang A, Schomburg I, et al. The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. Nucleic Acids Res. 2011; 39:D507–D513
43. Mungall CJ, Torniai C, Gkoutos GV, et al. Uberon, an integrative multi-species anatomy ontology. Genome Biol. 2012; 13:R5
44. Cooper L, Walls RL, Elser J, et al. The Plant Ontology as a Tool for Comparative Plant Anatomy and Genomic Analyses. Plant Cell Physiol. 2013; 54:e1
45. Bard J, Rhee SY, Ashburner M. An ontology for cell types. Genome Biol. 2005; 6:R21
46. Diehl AD, Augustine AD, Blake JA, et al. Hematopoietic cell types: prototype for a revised cell ontology. J. Biomed. Inform. 2011; 44:75–79
47. Bairoch A. The Cellosaurus, a Cell-Line Knowledge Resource. J. Biomol. Tech. JBT 2018; 29:25–38
48. 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
49. Christensen J, Cloos P, Toftegaard U, et al. Characterization of E2F8, a novel E2F-like cell-cycle regulated repressor of E2F-activated transcription. Nucleic Acids Res. 2005; 33:5458–5470

Acknowledgements

The authors would like to acknowledge the MI group from the HUPO-PSI consortium, the IMEx consortium and the GREEKC consortium for fruitful discussions.

Conflict of interests

The authors declare that they have no conflict of interest.

Funding

This project was supported by the Norwegian University of Science and Technology's Strategic Research Area NTNU Health [to VT, AF]; the ERACoSysMed grant COLOSYS [to MK, VT]; the Gene Regulation Ensemble Effort for the Knowledge Commons CA15205 [to MK, VT, SV]; the Research Council of Norway [to AL, MK, MLA, SV]; Alzheimer's Research UK [grant number ARUK-NAS2017A-1 and ARUK-NSG2018-003 to RCL]; the National Institute for Health Research University College London Hospitals Biomedical Research Centre [to RCL]; EMBL Core funding [to HH, LP, PP, ND-T]; Open Targets OTAR02-048 [to LP]; the National Institutes of Health [grant number 5U24CA184427-06 to DP]. This work was also supported by the National Eye Institute

(NEI), National Human Genome Research Institute (NHGRI), National Heart, Lung, and Blood Institute (NHLBI), National Institute on Aging (NIA), National Institute of Allergy and Infectious Diseases (NIAID), National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institute of General Medical Sciences (NIGMS), National Cancer Institute (NCI) and National Institute of Mental Health (NIMH) of the National Institutes of Health under Award Number [U24HG007822] (the content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health) [to CC-C, SO]. Finally, GB is an employee of GSK.