SIGNOR curation manual, July 2021

Introduction to SIGNOR Data Curation

The first step of the curation process consists in the systematic search of articles reporting evidences of causal relationships between biological entities (proteins, RNAs, chemicals, etc), specifying the effect (up/down-regulation) and the mechanisms (phosphorylation, binding, chemical activation, etc) i. To this end literature repositories, mainly PubMed, are searched either manually or semi-automatically, using text mining tools, to identify papers that potentially describe causal interactions. Each selected paper is then validated for reporting relevant content and annotated by trained curators.

During the curation process, curators use a tab delimited text file (or a spreadsheet) organized in three parts (entity, relationship and reference), every part contains mandatory fields to be filled with fundamental information describing a causal relation. In order to ensure data interoperability and reproducibility, a controlled vocabulary (CV) has been developed to represent the direction and sign of causal interactions in a structured format.

Each interaction curated in SIGNOR 2.0 must report information on the following mandatory fields:

ENTITIES

-Entity name

-EntityType (e.g. chemical, protein, protein family, complex, miRNA, phenotype, stimulus);

•Identifiers (e.g.UniprotKB id, ChEBI, Signor ID ...);

RELATIONSHIP

•Effect (e.g. up/down- regulates);

•Mechanism (e.g. binding, phosphorylation, transcriptional activation ….);

•Modified residue (i.e. Ser36)

•Organism, cell line, tissue

•Direct (yes/no)

REFERENCE

•PMID

•Short sentence

ENTITIES

Type

In the Entity part curators annotate information about the Entity Type of both source and the target. Entity Type in SIGNOR 2.0 may belong to ten different categories:

Protein Complex

miRNA

Small Molecule

Protein family

Fusion protein

Chemical

Phenotype

Stimulus

Antibody

Identifiers

Each entity is associated with a unique identifier and linked to a reference database. Please use them in column TYPEA or TYPEB along with DATABASEA or DATABASEB, respectively. Here are a few examples, please use the following terms and case for TYPE and DATABASE.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ENTITYA or ENTITYB | TYPE A or TYPEB | IDA or IDB (examples…) | DATABASEA or DATABASEB |  |
| EGFR | protein | P00533 | UNIPROT | UniProtKB |
| AMPK | complex | SIGNOR-C15 | SIGNOR | SIGNOR ID |
| miR-27b | mirna | MI0000440 | miRBase | mirBase and RNA Central |
| ATP | smallmolecule | CHEBI:15422 | ChEBI | ChEBI and PUBCHEM |
| ramucirumab | antibody | DB05578 | DRUGBANK | DRUGBANK |
| ErbB receptor family | proteinfamily | SIGNOR-PF36 | SIGNOR | SIGNOR ID |
| Differentiation | phenotype | SIGNOR-PH37 | SIGNOR | SIGNOR ID |
| DNA\_damage | stimulus | SIGNOR-ST1 | SIGNOR | SIGNOR ID |
| afatinib | chemical | CHEBI:61390 | ChEBI | ChEBI or PUBCHEM |
| BCR-ABL | fusion protein | SIGNOR-FP6 | SIGNOR | SIGNOR ID |

ABOUT PROTEINS…

* Interactions identified in animal models are ALWAYS remapped to the HUMAN orthologue PROTEINS.
* The targets of transcriptional regulation events are mapped to the gene product of the regulated gene.

example the interaction “P53 binds to the MDM2 promoter and activates MDM2” is captured as:

[Q00987](https://www.uniprot.org/uniprot/Q00987) (Protein TP53) up-regulates [P04637](https://www.uniprot.org/uniprot/P04637) (protein MDM2) by transcriptional regulation.

* Protein isoforms are annotated with the format:

<canonical ID>\_<isoformID> (example: P12345\_P12345-2)

* Protein chains that are processed post translationally are captured using PRO id with the format:

<canonical ID>\_<PRO id> (example: Q9NQ88-PRO\_0000179957)

* when an interaction is reported to involve a protein that belongs to a family (i.e AKT1 and AKT2) without specifying which family member, the interaction must be assigned to the family. Example: “akt increased estrogen receptor 1 activity” is captured as

SIGNOR-PF24 (protein family AKT) up-regulates P03372 (protein ESR1)

* when an interaction is reported to involve a protein that belongs to a family and is specified which family member is involved, the interaction must be assigned to the protein. Example: “akt2 increased estrogen receptor 1 activity” is captured as

[P31751](https://www.uniprot.org/uniprot/P31751) (protein AKT2) up-regulates P03372 (protein ESR1)

* when there is evidence that all the members of a family are involved in the interaction with another protein, the interaction is also assigned to the family.
* when an interaction is reported to involve a protein that belongs to a complex and is not specified which complex subunit is involved, the interaction must be assigned to the Complex. Example: “mTORC1 increased estrogen receptor 1 activity” is captured as SIGNOR-C3 (complex mTORC1) up-regulates P03372 (protein ESR1)
* when an interaction is reported to involve a protein that belongs to a complex and it is specified which complex subunit is involved, the interaction must be assigned to the Protein. Example: “mTOR increased estrogen receptor 1 activity” is captured as

P42345 (protein MTOR) up-regulates P03372 (protein ESR1).

NB: If the sentence indicates that this interaction occurs when the protein is part of the complex (e.g. “mTOR increased estrogen receptor 1 activity when part of the mTORC1 complex”) then the complex id shall be reported in the column MODULATOR\_COMPLEX (or TARGET\_COMPLEX if the subunit is the regulated entity). See SIGNOR-153450.

Directionality

In Signor 2.0, each interaction has a direction, with a regulatory entity (Entity A) that acts on a regulated entity (Entity B).

* All the information concerning the **regulator** will be captured in columns A-D of excel file (ENTITYA | TYPEA | IDA | DATABASEA);
* All the information related to the **regulation target** will be captured in columns E-H of excel file (ENTITYB | TYPEB | IDB | DATABASEB);

RELATIONSHIP

Effect

In the effect part, curators provide information about the effect (positive or negative) that the regulatory entity has on a regulated entity. A regulatory entity can act by up-regulating or down-regulating the function of another entity. It is also possible to specify if the regulation acts on the activity or on the quantity, i.e by modulating the expression or thestability (e.g. down-regulates quantity by expression or up-regulates quantity by stabilization).

**up-regulates** -> generic, used when no additional info is provided

**up-regulates activity** -> used when the action of the regulator affects the ACTIVITY of the target (e.g. “mTOR increased estrogen receptor 1 activity”)

**up-regulates quantity** -> generic, used when the action of the regulator enhances the QUANTITY of the target with no reference to the underlying mechanism.

**up-regulates quantity by expression** -> used when the action of the regulator positively modulates the transcription or the translation of the target (e.g. “P53 binds to the MDM2 promoter and activates MDM2 transcription”)

**up-regulates quantity by stabilization** -> used when the action of the regulator is to repress the degradation of the target (e.g. “ProtA protect ProtB from proteasomal degradation”)

**down-regulates** -> generic, used when no additional info is provided

**down-regulates activity** -> used when the action of the regulator affects the ACTIVITY of the target (e.g. “mTOR reduced estrogen receptor 1 activity”)

**down-regulates quantity** -> generic, used when the action of the regulator reduces the QUANTITY of the target with no reference to the underlying mechanism.

**down-regulates quantity by repression** -> used when the action of the regulator represses the transcription or the translation of the target (e.g. “P53 binds to the MDM2 promoter and inhibits MDM2 transcription”)

**down-regulates quantity by destabilization** -> used when the action of the regulator is to induce the degradation of the target (e.g. “ProtA induces ProtB degradation”)

**form complex ->** used to indicate the formation of a complex by its subunits

About the curation of complexes…

A complex needs to be defined in the database( See the paragraph ‘How to define an entity in signor’). Once the entity has been created it is **MANDATORY** to add the relationships ‘form complex’ between each subunit and the complex.

Example. *C-Fos dimerizes with c-Jun to form the transcription activator protein-1 (AP-1) which binds to the specific recognition site....* This info is captured as 2 relationships :

* P05412 (JUN protein) form complex SIGNOR-C154 (AP1 complex) by binding
* P01100 (FOS protein) form complex SIGNOR-C154 (AP1 complex) by binding

Mechanism

In the mechanism part, curators add details about the molecular mechanism of the underlying relationships (e.g. phosphorylation, transcriptional regulation, etc.). Depending on the type of mechanism, the field ‘DIRECT’ associated with the interaction can be YES or NO. The following table provides the dependencies between the ‘MECHANISM’ field and the ‘DIRECT’ field.

The field MECHANISM can be blank only when the relationship is indirect (DIRECT=’NO’).

|  |  |  |
| --- | --- | --- |
| MECHANISM NAME |  DIRECT |  |
| catalytic activity | YES |  |
| post translational modification and enzymatic reactions\*  | YES |  |
| chemical inhibition | YES |  |
| chemical activation | YES |  |
| gtpase-activating protein | YES |  |
| guanine nucleotide exchange factor | YES |  |
| binding | YES |  |
| relocalization | YES |  |
| post transcriptional regulation | NO |  |
| translation regulation | NO |  |
| blank | NO |  |
| transcriptional regulation | YES/NO | if DNA binding DIRECT='YES', else DIRECT='NO'. |

post translational modification and enzymatic reactions\* = catalytic activity; oxidoreductase activity; post translational modification; Farnesylation; deacetylation; demethylation; dephosphorylation; destabilization; acetylation; cleavage; desumoylation; deubiquitination; glycosylation; hydroxylation; neddylation; trimethylation; ubiquitination; s-nitrosylation; tyrosination; lipidation; oxidation; small molecule catalysis; methylation; palmitoylation; phosphorylation; sumoylation

NB: Transport activities and recruitment activities should be captured as ‘relocalizations’.

Modified residue

If in the supporting manuscript the amino acid residue that undergoes the PTM is identified, this information is also annotated.

It is strongly recommended to check whether the modified residue numbering is aligned to the Uniprot Sequence especially when the manuscript reports experiments in a species different from *Homo sapiens*

The modified residue is captured in the field RESIDUE in the following Format:

<three-letters aminoacid><modification position>

Example: Tyr345

NB: multiple modifications should be captured as different entries (rows)!

Example. *Ser-2093 is efficiently phosphorylated by GSK-3β and, to a minor extent, residues Thr-2068 and/or Ser-2070 and Thr-2074 of Notch2 are also targets for GSK-3β-dependent phosphorylation.We also find that GSK-3β-dependent phosphorylation of Notch2 is inhibiting transcriptional activation of different Notch target genes.* This is captured as four different lines:

P49841 (GSK3B) down-regulates activity Q04721 (NOTCH2) by phosphorylation at Ser2093

P49841 (GSK3B) down-regulates activity Q04721 (NOTCH2) by phosphorylation at Ser2070

P49841 (GSK3B) down-regulates activity Q04721 (NOTCH2) by phosphorylation at Thr2074

P49841 (GSK3B) down-regulates activity Q04721 (NOTCH2) by phosphorylation at Thr2068

The field SEQUENCE is automatically filled.

Organism, cell lines/tissues

The SIGNOR curation policy aims at capturing this information in the ‘TAX\_ID’, ‘CELL\_DATA’ and ‘TISSUE\_DATA’ fields of the excel file. This information is not mandatory but highly recommended.

The organism identity is captured in the field TAX\_ID. Use -1 when the interaction is demonstrated by 'in vitro' experiments .

Cell lines utilized in the experiment that support the causal interactions are captured in the field CELL\_DATA using the BRENDA IDs (e.g. BTO:0000007 for HEK293 cells, BTO:0000567 for Hela Cells).

The tissues where the relationship was experimentally determined are captured in the field TISSUE\_DATA using the BRENDA IDs (e.g. BTO:0000142 for ‘brain’, BTO:0000763 for ‘lung’).

Direct

In the ‘direct’ field, curators annotate whether the manuscript reports evidence that the annotated relationship is direct.

We define as DIRECT those interactions where the regulator acts immediately upstream of the target and no intermediate regulation is inferable from the sentence or the evidence. On the contrary, we define ‘INDIRECT‘ relationships where the regulator doesn’t act directly on the target implying intermediate steps of regulation. Some mechanisms, such as transcriptional regulation, can have associated either the interaction type ‘direct or indirect’ depending on the reported information. We consider ‘direct’ those transcription relationships in which it is clear that the regulator (transcription factor) binds the promoter of the target gene by activating it (e.g. P53 binds to the MDM2 promoter and activates MDM2), while ‘indirect’ when no such evidence is provided.

When the interaction is DIRECT ‘YES’ will be entered in the column ‘DIRECT’ of the excel file.

When the interaction is INDIRECT ‘NO’ will be entered in the column ‘DIRECT’ of the excel file.

Modulator and Target COMPLEX

When an interaction is reported to involve a protein that belongs to a complex and it is specified which complex subunit is involved, the interaction must be assigned to the Protein. However, if the sentence indicates that this interaction occurs when the protein is part of the complex (e.g. “mTOR increased estrogen receptor 1 activity when part of the mTORC1 complex”) then the complex id shall be reported in the column MODULATOR\_COMPLEX (or TARGET\_COMPLEX if the subunit is the regulated entity). See SIGNOR-153450.

Modulator and Target COMPLEX

The MODIFICATIONA and the MODIFICATIONB fields accommodate modifications in the source (MODIFICATIONA) or the target (MODIFICATIONB) required for the causal interaction to occur.

Example. EGFR phosphorylates GRB2 when

Format: <Modification>:<three-letters amino acid><modification position>

Phosphorylation:Ser443

REFERENCE

In the reference part, curators annotate the PubMed ID (PMID) of the supporting manuscript and a sentence, extracted from the same manuscript, possibly containing all (most of) the information captured in the entry. All publications MUST have an associated PMID cross-reference.

TYPE OF EVIDENCE ACCEPTED

The minimum requirement for an interaction to be captured from a reference is that the interaction is discussed in the reference (best in an article but reviews are acceptable). The types of evidence are ranked as follows:

1) Experimental evidence of the effect of the regulator on the target. HIGHLY ENCOURAGED.

2) Discussion in a review about the effect of the regulator on the target.

3) Causal effects mentioned *en passan*t with no further discussion. HIGHLY DISCOURAGED Example: ‘we treated cells with AKT inhibitor, GSK034629’