home edit page issue tracker RegulaTome Annotation Documentation

Documentation for Relation Annotation for the RegulaTome corpus

Relation type hierarchy for relationships annotated in RegulaTome

- Complex_formation
- Regulation
  - Positive_regulation
  - Negative_regulation
- Regulation_of_gene_expression
  - Regulation_of_transcription
  - Regulation_of_translation
- Regulation_of_degradation
- Catalysis_of_posttranslational_modification
  - Catalysis_of_small_protein_conjugation_or_removal
    - Catalysis_of_small_protein_conjugation
      - Catalysis_of_Ubiquitination
      - Catalysis_of_SUMOylation
      - Catalysis_of_Neddylation
      - Other_catalysis_of_small_protein_conjugation
    - Catalysis_of_small_protein_removal
      - Catalysis_of_Deubiquitination
      - Catalysis_of_DeSUMOylation
      - Catalysis_of_Deneddylation
      - Other_catalysis_of_small_protein_removal
  - Catalysis_of_phosphoryl_group_conjugation_or_removal
    - Catalysis_of_Phosphorylation
    - Catalysis_of_Dephosphorylation
  - Catalysis_of_other_small_molecule_conjugation_or_removal
    - Catalysis_of_small_molecule_conjugation
      - Catalysis_of_Methylation
      - Catalysis_of_Acylation
        - Catalysis_of_Acetylation
        - Catalysis_of_Palmitoylation
        - Catalysis_of_Myristoylation
      - Catalysis_of_lipidation
        - Catalysis_of_prenylation
          - Catalysis_of_farnesylation
          - Catalysis_of_geranylgeranylation
      - Catalysis_of_ADP-ribosylation
      - Catalysis_of_Glycosylation
      - Other_catalysis_of_small_molecule_conjugation
    - Catalysis_of_small_molecule_removal
      - Catalysis_of_Demethylation

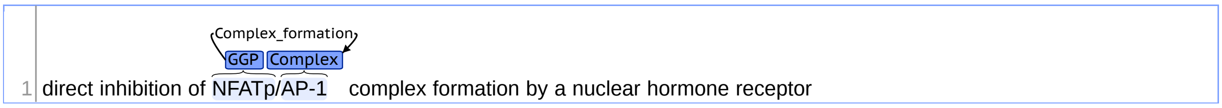
        - Catalysis_of_Deacylation
            - Catalysis_of_Deacetylation
            - Catalysis_of_Depalmitoylation
        - Catalysis_of_Deglycosylation
        - Other_catalysis_of_small_molecule_removal
- Out-of-scope

General guidelines

- Annotations should be made according to the annotator's best understanding of the **author's intended meaning in context**. For example, relations expressed using ambiguous verbs such as **"associate"** that express complex formation in some contexts but not others should be annotated if and only if the annotator interprets the authors as intending to describe complex formation. The annotators should only use the text excerpt they have available to make this judgement.
- Annotators should treat Gene or Gene Product, Protein-containing Complex and Chemical named entities as being **masked**, i.e. they shouldn't annotate relationships between entities just based on their names, when they would be unable to make the same annotations for two other entities. Protein Family should be treated as **NOT** being masked and it will be a question of experimental setting later as to whether the information that can be extracted from the names is important or not.

Detailed guidelines

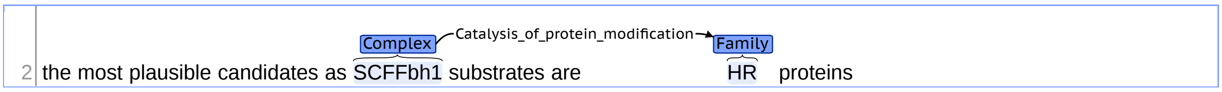
1. Complex formation relations can be annotated between two different protein mentions, but also between the same mentions, when the masked entities could be viewed as two different entities. However, statements such as "homodimerization of A" **are not annotated** as *Complex formation*.
2. Complexes of more than two proteins are annotated by creating **all binary relations** between the components.
3. Nominalized expressions ("interaction of A and B", "A/B interaction", "A:B complex") and noun phrases with **any surface word** that can be understood as implying the existence of a complex ("A/B complex", "A/B heterodimer") are **annotated** as expressing complex formation relations. However, **in the absence of any such word**, text such as "A/B" is not annotated. The text A-B will be annotated based on the understanding of the annotator from the entire context (abstract or paragraph) and not based on former biological knowledge.

   direct inhibition of NFATp/AP-1 complex formation by a nuclear hormone receptor

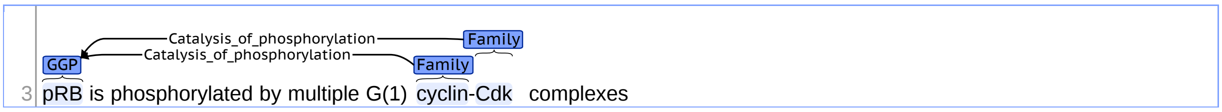
   Exception: Sentences like *"A is phosphorylated in vitro using B/C (or B-C)"* where B is a kinase and C is a cyclin have been annotated as *"B complex formation C"*. These can be reverted or we can check the error rate in this specific subproblem.
4. Relations **should not be interpreted as combinations**, on the contrary each annotated relation should be **valid on each own** (e.g. *"A positively regulates the proteolytic degradation of B and that leads to the rapid depletion of B"*, should be annotated as *"A regulation of degradation B"* and *"A negative regulation B"* and not the combination of relations *"A positive regulation B"* and *"A regulation of degradation B"*).
5. **Co-immunoprecipitation** can be used as an indicator of complex formation between two named entity mentions
6. *"A regulation of degradation B"* does not necessarily imply *"A negative regulation of B"*, so this should be annotated with care.
7. Post-translational modifications should **not** receive a binding annotation unless binding is clearly mentioned in context. Post-translational modifications (PTMs) imply transient interactions which will not be present in physical interaction databases, so they shouldn't be annotated as such. For an example of a corner case see Specific examples

8. The following are generally understood as implying *Complex formation*:
    - consitutive association
    - stable association
9. The following are generally understood as **NOT** implying *Complex formation*:
    - synergize
    - stabilize
10. If **part of a protein/complex** has the ability to **form a complex**, then the ability of the entire protein/complex to do the same can be extrapolated from that.
11. Subcellular localization is not annotated for *Complex formation* even if the structure is made of proteins.
12. When an entity is a substrate of another entity then the relation connecting them is **Catalysis of protein modification**. For example:



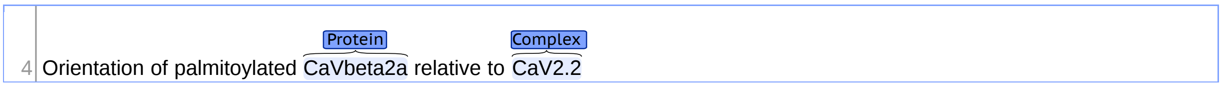
2 the most plausible candidates as SCFFbh1 substrates are HR proteins

13. If a cyclin-kinase complex phosphorylates protein A then **Catalysis of phosphorylation** should be added both for the *kinase* **AND** the *cyclin*. For example:



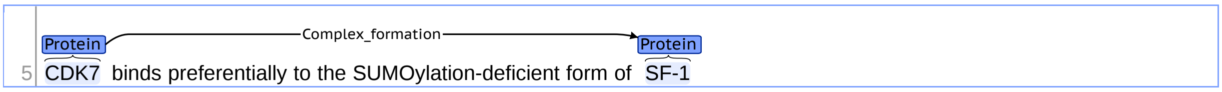
3 pRB is phosphorylated by multiple G(1) cyclin-Cdk complexes

14. Synthetic lethal interactions are genetic and thus are **NOT** annotated as Complex formation.
15. If protein A acts as an **effector** for protein B then *B regulation A* is annotated.
16. If **transfection** with a protein leads to it obtaining a relationship with another protein entity, then that relationship should be annotated.
17. **Chemicals** COVALENTLY bound to other entities are not annotated as **Complex formation**, since complex formation is non-covalent.
18. Orientation of Protein A relatively to Protein B is not enough cue to annotate **Complex formation**. For example:



4 Orientation of palmitoylated CaVbeta2a relative to CaV2.2

19. Incorporation of a small molecule/protein congugate to a Protein (i.e. a PTM) is **Out-of-scope** and should not be annotated as **Complex formation**
20. Proteoforms (e.g. proteins with PTMs, or isoforms), should receive annotations as if they were the main isoform/unmodified protein. For example:



5 CDK7 binds preferentially to the SUMOylation-deficient form of SF-1

21. **Complex formation** should be annotated when a **Chemical** binds to any other entity (**Protein**, **Family** or **Complex**) unless it is clearly stated that the bond is covalent (either by the fact that it is a PTM or covalently bound is mentioned in the text).
22. **Entity A is post-translationally modified by Entity B** should be interpreted as *B Catalysis of Protein Modification A*, UNLESS **B is a protein conjugate**, where B should receive an attribute *Small mol PTM* and the relationship between A and B is **Out-of-scope**
23. The interactions between members of **transient intermediate complexes** as part of catalytic reactions should **NOT** be annotated neither between Protein-Protein (e.g. kinase-substrate), nor between Protein-Chemical.
24. **Chemical A modulates, inhibits, acts as an agonist/antagonist for Protein B**: On top of any regulatory relationships, a **Complex formation** relationship between A and B should be annotated. This rule applies mostly to drugs.

Negation and speculation

1. Statements explicitly **denying** a relationship, for example the formation of a complex (*"A does not bind B"*)

are **not annotated** in any way. However, if the negated statement is qualified with conditions in a way that implies that the proteins would normally e.g. form a complex, the statement is annotated as if the negation were absent. For example in the sentence *"When A is phosphorylated, it fails to form a complex with B"*, it is implied that under other circumstances A and B would form a comples and *A Complex Formation B* is annotated.
2. Statements expressed **speculatively** or with **hedging** expressions (e.g. *"may form a complex"*) are **annotated** identically to affirmative statements (in effect, **speculation and hedging are ignored**).

Complex formation relationships

Undirected binary relation associating two proteins that form a complex. Annotated for any statement implying the existence of a complex, including statements explicitly discussing the dissociation of a complex. Relevant ontology terms:

- GO:0065003 (**protein-containing complex assembly**): The aggregation, arrangement and bonding together of a set of macromolecules to form a protein-containing complex.
- GO:0032984 (**protein-containing complex disassembly**): The disaggregation of a protein-containing macromolecular complex into its constituent components.
- GO:0032991 (**protein-containing complex**): A stable assembly of two or more macromolecules, i.e. proteins, nucleic acids, carbohydrates or lipids, in which at least one component is a protein and the constituent parts function together.

Note that by contrast to the scope of GO:0032991 (protein-containing complex) and related terms, the annotated complex formation relation is restricted to cases where both of the associated constituents are *proteins*, *protein complexes*, *protein families* or *chemicals*.

Regulation relationships

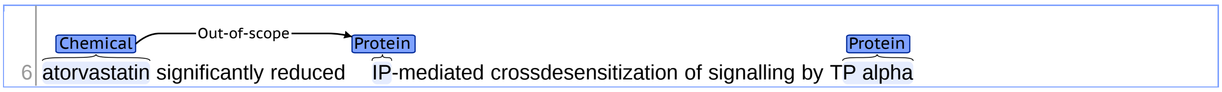
Regulation relationships are generally annotated in cases where we know that entity A has an effect on entity B, even in cases where we don't know the type of effect A has on B, but we know it is upstream. The relevant GO term upon which are annotations are based is Regulation of biological process. Some more specific rules had to be added for this class for annotation consistency:

1. **Protein X can do sth to Protein Y, in a Chemical Z dependent manner**: Z>regulates>X
2. **Protein X responds to Chemical Y**: Y>regulates>X
3. **Changes in protein level of protein A lead to changes in the molecular function of protein B**: A>regulates>B
4. **Protein A levels are regulated in a protein-B dependent manner**: B>Regulates>A (As they can be regulated through expression, i.e. positively or through degradation, i.e. negatively)
5. **Protein A participates in the assembly of Complex B**: A>Regulates>B
6. **Protein A structure is regulated by Protein B**: B>Regulates>A (not sure if the structure is regulated positively or negatively)
7. **Protein A and protein B gain a function upon forming a complex and this function affects another entity C**: A>regulates C AND B>regulates>C
8. **Targeting of protein A to a specific subcellular location is mediated/regulated/blocked/allowed (including secretion) by protein B**: B>regulates>A
9. **Protein A mediates the activation/inhibition of protein B**: A>Regulates>B
10. **Activity of entity A is sensitive to entity B**: B>Regulates>A
11. **PTM of protein A controls protein B**: A>Regulates>B
12. **Protein A phosphorylates Protein B and the phosphorylation leads to regulation of Protein C**: A>Regulates>C, B>Regulates>C, A>Catalysis of phosphorylation>B
13. **Protein A acts upstream of Protein B**: A>Regulates>B
14. **Chemical A-responsive Protein B**: A>Regulates>B

15. **Protein A cleaves Protein B**: A>Regulates>B, because the cleavage is not necessarily leading to up or down regulation of the protein. We choose to annotate it because most of the times the cleavage will have an effect on the protein. The only reason to annotate **Negative Regulation** and **Regulation of degradation** in these cases is if it is clearly mentioned that the cleavage from Protein A leads to degradation of Protein B.

Special cases for Regulation

1. **Protein A induces a PTM of protein B**: This is not annotated, unless changes in the levels/structure of protein B related to the PTM are described in the sentence, where *Positive/Negative regulation* should be annotated.
2. The same rule applies if the **PTM** is a **small protein conjugation/removal**. So there shouldn't be any extra annotations regarding regulations of protein levels for *Catalysis of (de)ubiquitination*, *Catalysis of (de)SUMOylation* or *Catalysis of (de)NEDDylation*, unless clearly stated in a sentence.
3. **Entity A-mediated PTM of Entity B**: *A>Catalysis of PTM>B*. But **A-promoted PTM of B** is not such a strong hint, thus no relationship will be annotated in such cases.
4. **Regulation of a PTM of Protein A by Protein B** should **NOT** be annotated as B>Regulation>A. This should also be applied when a single ubiquitin molecule is added (a single ubiquitin does not necessarily imply a change in function). But **regulation/induction of polyubiquitination** implies **regulation of degradation**, as the polyubiquitin is a signal for the proteasome SHOULD BE annotated as **Regulation of degradation**.
5. **Proteasome** is the only complex that performs protein degradation, so **Regulation of degradation** is **NOT** annotated in that case (as we are annotating Proteins who regulate, not perform, the degradation process). Same applies if a protein is hydrolyzing another protein.
6. **Protein X degrades Protein Y** will be annotated as **X>Negative Regulation>Y** even if the complex is **proteasome**.
7. Regulation of degradation of **misfolded** proteins should **NOT** be annotated as regulation of degradation.
8. **Chemical A reduces Protein B-mediated process**: *OOS* since we don't know if this is regulating the actual protein mentioned or something downstream in the process.

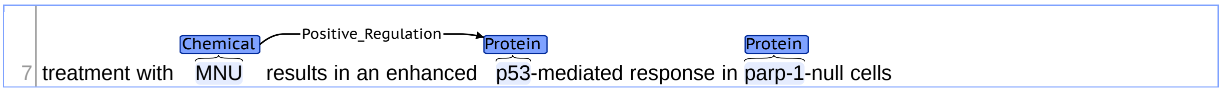


atorvastatin significantly reduced IP-mediated crossdesensitization of signalling by TP alpha

9. **Protein A function was increased/reduced in Protein Y mutants/cells**: **OOS** because one cannot necessarily extrapolate that it's the protein the cell line is named after that is responsible for the effect.
10. **Protein A acts as a counterpart of Protein B**: **OOS**, since "acts as a counterpart" means that they have a similar function or position in a different place.
11. **Protein A is reduced by Protein B**: **OOS**. The word "reduced" standalone without mention to levels could refer to the chemical reaction of reduction. This should **NOT** be annotated as *Negative regulation*.
12. If **Protein X up/down regulates responses induced by Protein Y stimulation**: **OOS** since we don't know if it is protein Y that is regulated.

Positive/Negative regulation relationships

1. **Protein A is essential/required/needed in the assembly of Complex B**: A> Positively regulates>B
2. The word **chaperone** can be used as a hint to annotate *Positive regulation*
3. **Protein A is necessary for Protein B to perform a function**: A> Positive Regulation>B
4. **Protein A function is restored in the presence of Protein B**: B>Positively Regulates>A
5. **Protein A structure is maintained by Protein B**: B>Positively Regulates>A
6. **Protein A directly activates/inhibits protein B**: A>Positively/Negatively Regulates>B
7. **Protein A levels increase/decrease because of protein B**: B>Positively/Negatively Regulates>A
8. **Cell lines**: when we have cell lines (including mutant cell lines), but it's not the gene the cell line is named after that is affected annotate as **Positive regulation**

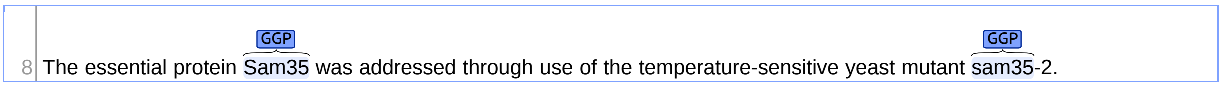


treatment with MNU results in an enhanced p53-mediated response in parp-1-null cells

9. **Protein A catalyzes the modification of Chemical B**: A>Negatively Regulates>B. The idea is that if a Chemical is modified, then we end up with another chemical, which means that the actual levels of the original chemical mentioned are lower. The same is not true for Proteins because having a different proteoform (e.g. with a PTM), doesn't change the Protein in the same sense. **Catalysis of protein modification** should **NOT** be annotated though, because PTMs are by definition for Proteins (including Protein Families and Complexes) so those are not annotated for chemical entities.
10. **Protein A transactivates Protein B**: A>Positive Regulation>B, A>Regulation of transcription>B (Transactivation refers to the increased rate of transcription)
11. **Protein A targets Protein B for ubiquitin-mediated degradation**: A Negative Regulation B
12. **If Protein X causes the aggregation of Protein Y**: X Negative Regulation Y
13. **If Protein X activity is attenuated by Protein Y**: Y Negative Regulation X
14. **Suppressed/Repressed activity of a Protein/Protein Family A from Protein B**: B Negative Regulation A
15. **A restrains expression of B**: A Negative Regulation B, A Regulation of gene expression B
16. **A opposes the activity of B**: A Negative Regulation B
17. **A is a biosynthesis inhibitor of B**: A Negative Regulation B

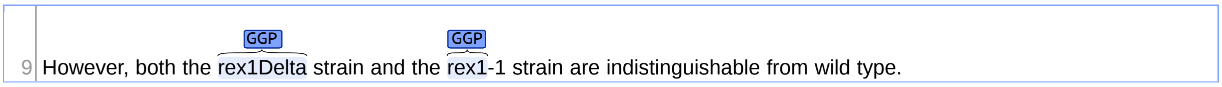
Named Entity annotation rules

1. Entity name mentions like *ubiquitin* or reporter genes (e.g. *GFP*) which are *GGPs* but are blacklisted by tagger, will be assigned the **blacklisted** attribute
2. Histones:
   - Tag *H2*, *H3* etc. when they appear standalone
   - Include *histone* in the span when it appears with one of the names (e.g. *histone H3*)
   - Tag *histone* as **Protein family or group** when it appears standalone
   - Methylated histones are also tagged as **GGP**
3. *Amino acid residues* should not be annotated as **chemical** when they are part of a polypeptide chain
4. *Glycosylphosphatidylinosiol* (GPI) should not be annotated as **chemical** as it cannot be a standalone chemical
5. Determiners like *the* should not be included in the entity span of **GGP**, **Protein-containing complex** and **Protein family or group**
6. Mutants of specific proteins will receive **GGP** annotations and an *Entity Attribute*: **Mutant**
7. In order for the annotated text to be as close as possible to the ideal NE annotation produced by the NER system, cases where only part-of **mutant names** are standalone entities, only these mentions should be annotated, e.g. in the following example from 18039934 *sam35* and **NOT** *sam35-2* is annotated as a ggp

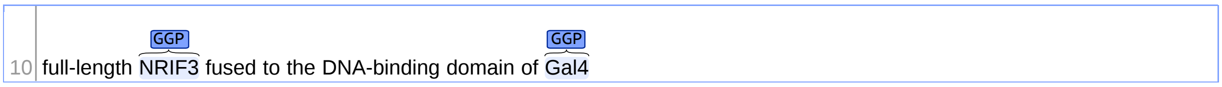
   The essential protein Sam35 was addressed through use of the temperature-sensitive yeast mutant sam35-2.

   An exception is when **mutant names are a single word**, and then they are annotated as one mutant entity e.g. **rex1Delta** in the following sentence from 16100378

   However, both the rex1Delta strain and the rex1-1 strain are indistinguishable from wild type.

8. Named entities that are part of antibodies should be annotated as the corresponding NE type and should receive a *Note: antibody*
9. rRNAs and tRNAs are annotated as **GGP** with the **noncoding** attribute
10. **Fusion proteins** should be treated as two entities for the purposes of annotation and during the creation of the training dataset. These should get an *Entity Attribute*: **Fusion**. The reporter protein in fusion should get a note: **not tagged by tagger** if it is not detect by tagger. E.g. in this document **NRIF3** will receive an *Entity Attribute*: **Fusion** and **Gal4** will receive an *Entity Attribute*: **Fusion** and a *Note: not tagged by tagger* 11713274

    full-length NRIF3 fused to the DNA-binding domain of Gal4

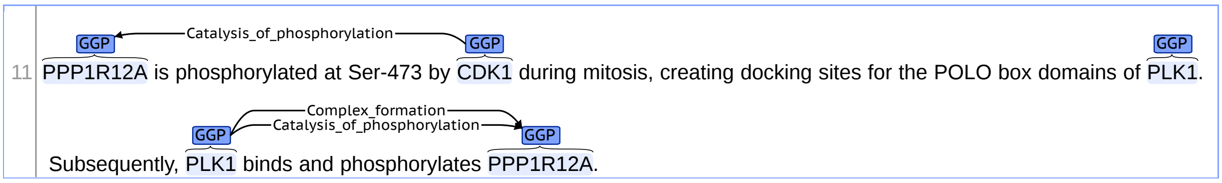
11. *Domains* and other *protein regions* should **NOT** be annotated as *GGP*
12. *FLAG* and *6xHis* are polypeptide protein tags and should receive an *OOS* annotation, or should not be annotated at all.
13. *ATP* and *ADP* are annotated as **OOS**
14. *GTP* and *GDP* are annotated as **Chemicals** due to their signalling function
15. Palmitate, Myristate, Acetate and ADP-ribose should receive a **Chemical** annotation. When they are mentioned as [3H]chemical, only the chemical is annotated (e.g. **myristate** in [3H]myristate)
16. **prenyl** is a functional group and thus **NOT** annotated as chemical. Similarly **geranylgeranyl** and **farnesyl** are also protein groups and are **NOT annotated as Chemical**.
17. **HMR** and **HML** loci will be annotated as GGP, but no annotation will be added to their chromosomal location (e.g. 17p13.1)
18. **Cyclic AMP** is annotated as Chemical and **AMP** annotated as OOS
19. **Zinc** as part of **zinc finger** domain won't be annotated as Chemical
20. **Leucine** as part of **leucine zipper/latch** won't be annotated as Chemical
21. **POU** will be annotated as Family when part-of **POU transcription factor**, annotated as OOS when part of **POU domain**
22. **ankyrin** will be annotated as OOS when part-of **ankyrin repeats**'
23. **RING** standalone will NOT be annotated as Family
24. **phospholipid** will be annotated as **Chemical** with **blacklist** attribute

Specific rules for complexes/families and plural form annotations

1. If a term is in Gene Ontology and is assigned a Protein-containing complex annotation then it is considered a Complex in this annotation effort,
2. If a term is found in Gene ontology but it is NOT a **protein-containing complex**, then it will **NOT** be considered a *Complex* in this effort
3. If a term is not at all present in Gene Ontology then other resources in the field will be used to decide whether it should be considered a *Complex* or not (e.g. Complex Portal, Reactome).
4. For cases where it is difficult to distinguish *family* from *domain* mentions, the field type in Pfam could be used to aid in making a decision (if available).
5. When a *protein-containing complex* name coincides with the name of the GGPs comprising it, seperated by ***ANY*** punctuation, annotation of the **GGP** NEs is preferred over annotation of the protein-containing complex. Two notable exceptions are *Arp2/3* and *SWI/SNF* where a single **Protein-containing complex** NE is annotated instead
6. **Complexes** (**Protein-containing complex**) and protein **Families** (**Protein family or group**) are annotated if there is data verifying the existence of them. As a general rule for **Complexes** Gene Ontology - Cellular Component or another database of Complexes is enough to annotate an entity as Complex. For protein **Families** information is gathered from InterPro or Wikipedia or specific publications for very rare cases.
7. The words *"complex"* and *"family"* _ should **not** be part of the entity annotations.
8. Annotations should be applied to all variants of a name: e.g. **NF kappaB**, **NF-kappaB**, **NFkappaB** should all be marked as **Protein-containing complex**

Specific Examples discussed

1. Instances of binding and phosphorylation should be seperately annotated as two events when binding is clearly mentioned in text.

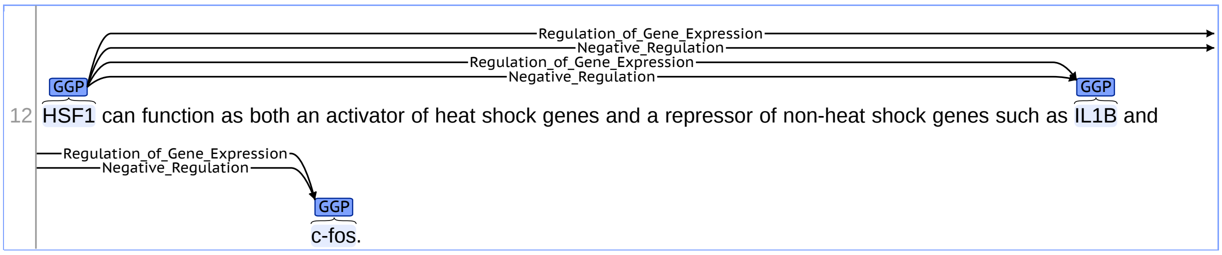


11 PPP1R12A is phosphorylated at Ser-473 by CDK1 during mitosis, creating docking sites for the POLO box domains of PLK1.

Subsequently, PLK1 binds and phosphorylates PPP1R12A.

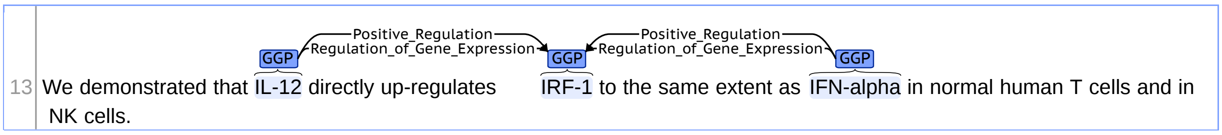
2. When *"regulation of expression"* is mentioned in text, if the annotator suspects that the intended meaning

of the authors is **Regulation of Transcription** based on the context of the document they have available, then the annotator should annotate **Regulation of Gene Expression**, with a **Note: "Potentially Regulation of Transcription"**.


Regulation_of_Gene_Expression
Negative_Regulation
Regulation_of_Gene_Expression
Negative_Regulation
GGP
GGP
Regulation_of_Gene_Expression
Negative_Regulation
GGP

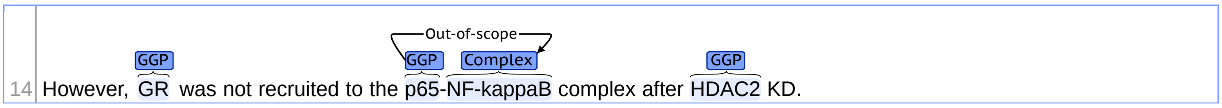
12 HSF1 can function as both an activator of heat shock genes and a repressor of non-heat shock genes such as IL1B and c-fos.

3. When the level at which the protein product is regulated at is not clear, then the general term *Regulation of Gene Expression* should be used.



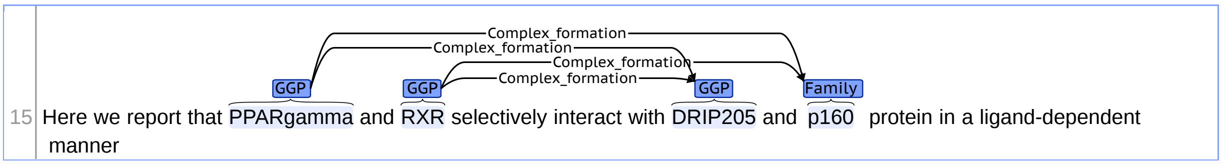
13 We demonstrated that IL-12 directly up-regulates IRF-1 to the same extent as IFN-alpha in normal human T cells and in NK cells.

4. In the current scheme we can annotate the semantics of e.g. *"A negatively regulates the expression of B"* by assigning **two relations**: *A negatively regulates B* AND *A Regulation of Gene Expression B*
5. The following is a **part-of relationship** and not complex formation, so it should be annotated as **Out-of-scope** 16380507



14 However, GR was not recruited to the p65-NF-kappaB complex after HDAC2 KD.

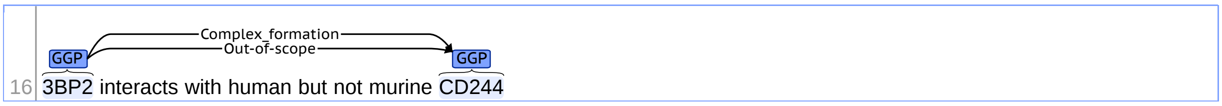
6. In document 19328066, *Mex67:Mtr2*, *NXF1:NXT1* and *TAP:P15* represent protein heterodimers and have been annotated as such, since the first sentence mentions **Mex67:Mtr2 heterodimer** denoting that in the document *:* can be used to represent *heterodimers*.
7. In this sentence, it looks like quite clear mentions of **Complex formation**, but *"selectively interact … in a ligand-dependent manner"* probably denotes the correct pairs, which are otherwise not clear.



15 Here we report that PPARgamma and RXR selectively interact with DRIP205 and p160 protein in a ligand-dependent manner

In the very next sentence, the authors explain in detail, that only *PPARgamma-DRIP205* and *RXR-p160* are interacting.

8. In this document, the *Complex formation* relationship occurs only for *human CD244*. Complex formation and negated Complex formation are in a single sentence with coordination.



16 3BP2 interacts with human but not murine CD244

For information on Annodoc, see http://spyysalo.github.io/annodoc/.