Documentation for Physical Protein Interaction Relationship Annotation of the ComplexTome corpus and trigger word annotation

Relationship Annotation

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 all named entities as being **masked**. **Masked** is a term adopted from large language model training, and it means that the entities should be treated as if they are not visible, but the annotators knows their place in text (e.g. *mutations in* ***p53*** *have been associated with lung cancer* should be treated as *mutations in* ***[MASK]*** *have been associated with lung cancer*). This means that annotators 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.

Complex formation definition

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 gene 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.

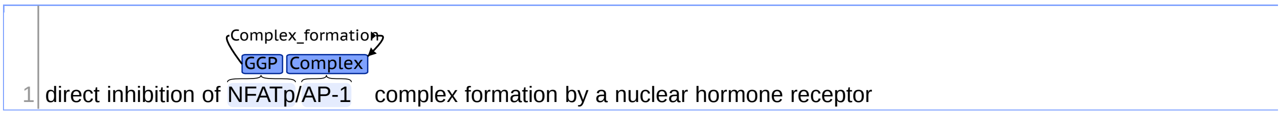
Note that by contrast to the scope of GO:0032991 (**protein-containing complex**) and related terms, the annotated complex formation relation is restricted to cases where both of the associated constituents are *proteins*, *protein complexes*, *protein families*, *groups of proteins* or *chemicals*.

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.

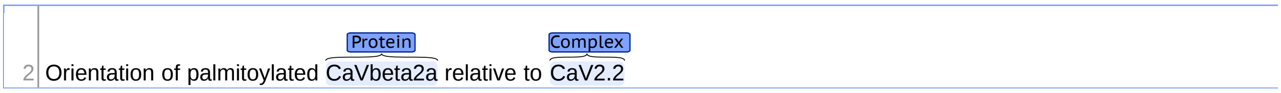
However, statements such as “homodimerization of A” **are not annotated** as *Complex formation*, since self-loops are not annotated in the corpus.

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.



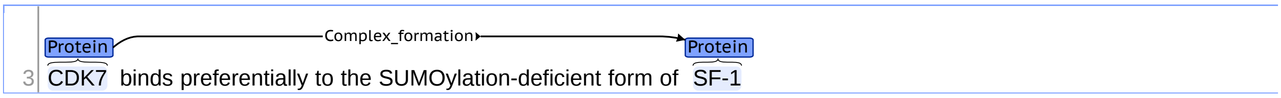
direct inhibition of NFATp/AP-1 complex formation by a nuclear hormone receptor

4. Relations **should not be interpreted as combinations**, on the contrary each annotated relation should be **valid on each own**.
5. **Co-immunoprecipitation** can be used as an indicator of complex formation between two NE mentions.
6. Post-translational modifications should **not** receive a binding annotation unless binding is clearly mentioned in context. 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
7. The following are generally understood as implying *Complex formation*:
   - consitutive association
   - stable association
8. The following are generally understood as **NOT** implying *Complex formation*:
   - synergize
   - stabilize
9. Incorporation of a small molecule/protein congugate to a Protein (i.e. a Post-translational modification) is **Out-of-scope** and should not be annotated as **Complex formation**
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** and not **Complex formation**. Thus no annotation is added in such cases.
13. Synthetic lethal interactions are genetic and thus are **NOT** annotated as Complex formation.
14. **Chemicals** COVALENTLY bound to other entities are **NOT** annotated as **Complex formation**, since complex formation is non-covalent interactions.
15. Orientation of *Protein A* relatively to *Protein B* is not enough cue to annotate **Complex formation** e.g. from 19858358



Orientation of palmitoylated CaVbeta2a relative to CaV2.2

16. Proteoforms (e.g. proteins with PTMs, or isoforms), should receive annotations as if they were the main isoform/unmodified protein e.g. from 19015234



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

17. **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 post-translational modification or covalently bound is mentioned in the text).

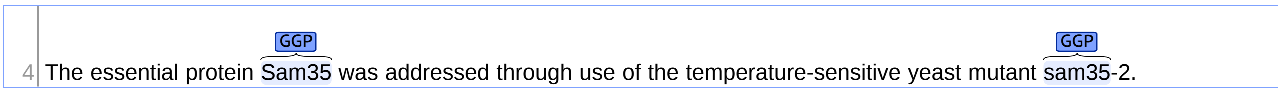
18. The interactions between members of **transient intermediate complexes** as part of catalytic reactions should **NOT** be annotated neither between Protein-Protein (e.g. kinase-substrate), nor between Protein-Chemical entities.
19. **Chemical A modulates, inhibits, acts as an agonist/antagonist for Protein B**: A **Complex formation** relationship between A and B should be annotated (this rule applies mostly to drugs.)

Negation and speculation

1. Statements explicitly **denying** the formation of a complex (e.g. "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 form a complex, the statement is annotated as if the negation were absent (e.g. *"When A is phosphorylated, it fails to form a complex with B"*).
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**).

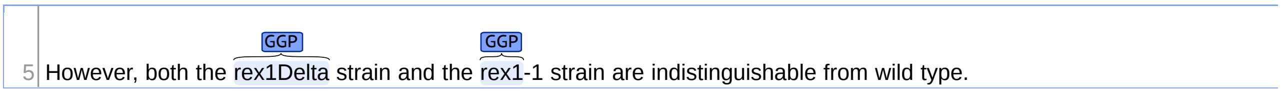
Named Entity annotation rules

1. Entity name mentions like *ubiquitin* or reporter genes (e.g. *GFP*) which are *GGPs* but are in the blocklist of our NER system, will be assigned the **blacklisted** attribute (see next section)
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.
    - We could then either go discontinuous or decomposed for mentions such as *histones H2A and H3*.
    - Methylated histones are also tagged as **GGP** even though our NER system will not detect them
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. *Domains* and other *protein regions* should **NOT** be annotated as *GGP*.
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. **sam35** and **NOT** *sam35-2* is annotated as a GGP in the following example

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

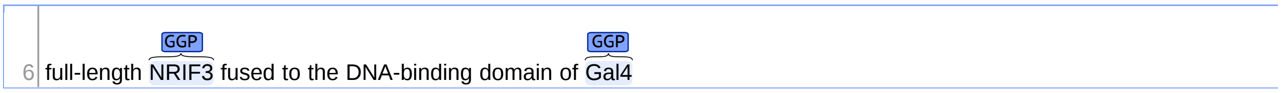
    An exception is when **mutant names are a single word**, and then they are annotated as one mutant entity e.g. **rex1Delta** in the following sentence:

    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 currently annotated as **GGP** with **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 an attribute: **blacklisted** if it is not detect by tagger. E.g. in the example below **NRIF3** will receive an *Entity Attribute*: **Fusion** and **Gal4** will receive an *Entity Attribute*: **Fusion** *Entity Attribute*: **Blacklisted**:


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

11. *FLAG* and *6xHis* are polypeptide protein tags and should receive an *OOS* annotation, or should not be annotated at all.
12. *ATP* and *ADP* are annotated as **OOS**.
13. *GTP* and *GDP* are annotated as **Chemicals** due to their function in protein signalling.

Named Entity Attributes

There are 5 Named Entity (NE) attributes in the corpus:

1. **Mutant**: used to mark NEs that are mutated forms or mutants of the annotated entity
2. **Fusion**: used to mark NEs which are part of fusion proteins
3. **Non-coding**: used as an attribute for GGPs to denote functional non-coding RNA molecules (e.g. transfer RNA, microRNA, piRNA, ribosomal RNA, and regulatory RNAs) among others.
4. **Small protein post-translation modification**: used as an attribute to denote GGPs that are covalently attached to other proteins as a result of a post-translational modification (e.g. ubiquitin, SUMO)
5. **Blacklisted**: used to denote NEs that belong to one of the annotated NE types, but which are not detected by our dictionary-based NER system, since they are part of its blacklist.

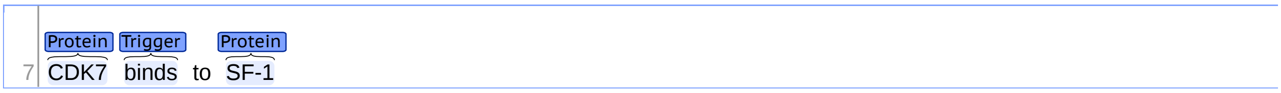
Specific rules for complexes/families and plural form annotations

- 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.
- 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
- 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).
- There is no clear distinction in Gene Ontology between small (e.g. NF-kappaB) and large (e.g. Nuclear Pore) complexes and for this reason, all these complexes will be treated the same and receive a *Complex* annotation
- 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)
- The words *"complex"*, *"family"* and *"group"* should **not** be part of the entity annotations.
- 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**

Trigger word annotation

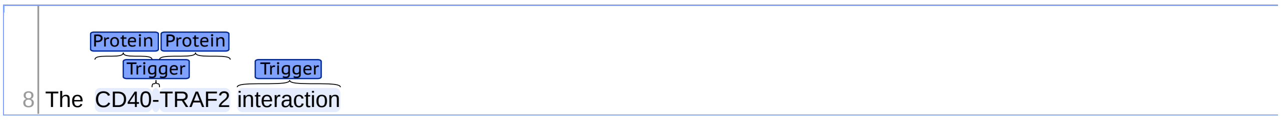
General guidelines

- When annotators have already identified a **Complex formation** relationship in text, it is possible to also annotate the specific word(s) which led them to make this annotation. The words that allow their interpretation of a relationship as **Complex formation** are called trigger words. An example of a trigger word annotation is shown below:



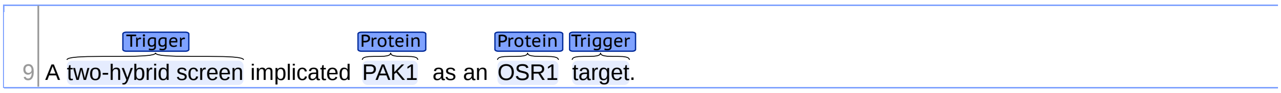
CDK7 binds to SF-1

- If two or more trigger words were considered as equivalently valid they will all be annotated.



The CD40-TRAF2 interaction

- If a trigger word is discontinuous, all the constituents of the trigger words will be annotated.



A two-hybrid screen implicated PAK1 as an OSR1 target.

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