CHEMDNER-patents: Gene and Protein Related Object (GPRO) annotation manual

Version 1.0 (3rd June 2015)

This document describes the guidelines used for the construction of the annotations of mentions of gene and protein related objects (named as GPROs throughout this manuscript) of the CHEMDNER-patents corpus (*the so-called GPRO-patent corpus*). It provides the basic details of the GPRO annotation task and the conventions that should be followed during the corpus construction process. The GPRO annotation guidelines have been refined after iterative cycles of annotations of sample documents. It also incorporated suggestions made by curators as well as observations of annotation inconsistencies encountered when comparing results from different human curators. In brief, the annotated GPROs include genes, gene products (proteins, RNA), DNA/protein sequence elements and protein families, domains and complexes. The aim of the iterative manual annotation cycles was to improve the quality and consistency of the guidelines, in order to make them more intuitive and easier to follow. During the preparation process of the guidelines some rules had to be reformulated to make them more explicit and additional rules were added when necessary to better cover the practical annotation scenario and for being more complete.

The manual annotation task basically consisted of labeling or marking up manually through a customized web-interface the mentions of GPROs (and related entities) in text. This was done following a set of rules that will be specified in more detail below. The text that was labeled consisted of patent abstracts (titles and abstracts in English) from patents published between 2005 and 2014 that had been assigned to the IPC codes A61P and A61K31.

The selected GPRO entity mentions were classified by hand into one of eight GPRO entity mention classes (see table 1).

GPRO class	Description	Examples
C2:NO CLASS (NC) (tag: no class*)	Names of individual protein domains and names of sequence or structural motifs. Here also DNA/RNA structural motifs should be labeled. Identifiers of protein domains (PFAM).	Epidermal growth factor-like domain; Sh2 domain; PDZ domain; MAM domain; SH3 domain; CARD domains; CD74-homology domain; ATPase-like motifs
C1:NESTED MENTIONS (NM) (tag: systematic*)	Nested mentions of a single entity. This would correspond to nested mentions where the actual decomposed entity mentions could be normalized to one unique entity.	ribosomal protein S6 (p70S6) kinase
C1:IDENTIFIER (ID)	Database identifiers of	P35354; PGH2_HUMAN

(tag: identifier)	genes or gene products (proteins, RNA). This includes identifiers and/or accession numbers from UniProt, GenBank, RefSeq, Ensembl, PDB, HGNC. Also model organism database identifiers should be tagged (e.g. from MGI, RGD, FlyBase, ...) Here SNP identifiers should be included too. This class includes mentions of enzyme commission numbers, so called EC numbers.	(UniProt); 3NSS (PDB) ; ADA71175; NM 006475 (GenBank); rs12979860; rs12153855; rs17207923; EC:2.7.11.1; EC 2.3.2.5 , FBgn0003638 (Flybase)
C2:SEQUENCE (SE) **(tag: formula*)**	Mentions of protein (amino acid) sequences, nucleotide sequences, mutation and residue mentions of DNA, RNA and proteins. Also includes: promoter sequences, sequence motifs.	CACGTG; SQEY motif; VGFPV motif; 5′-C/U-U-G/U-U-3′; C3592T; G3602A; G12V; Gly12Val; substitution of Glycine 12 to Valine
C1: FULL NAME (FN) **(tag: trivial*)**	Full name of a GPRO, including names of precursor proteins (in case of cleaved proteins). It also includes multi-word terms referring to specific gene/protein named entities. It covers single word GPRO what do not correspond to abbreviations or symbols.	ribosomal protein S6; DAP kinase-related apoptosis-inducing protein kinase 2; Serum amyloid A3; cyclooxygenase-2; Heat shock protein 90; human deoxycytidine kinase protein; human somatostatin 2 receptor protein; thromboxane synthase; Kirsten rat sarcoma viral oncogene homolog; tumor protein p53
C1:ABBREVIATION (AB) **(tag: abbreviation)**	Abbreviation of full name GPROs, GPRO acronyms, gene/protein symbols or symbolic names. Also two-word GPROs where one of the words is an Abbreviation and the other word is a number or single letter (e.g. Cox 2; H Ras; miR 145).	Drak2; p70S6; SAA; COX2; miR-145; SETD1B; MIR4304; TNXB; NOTCH4; IFNα-7; grk5; KCTD1; HSATU68; SCN9A, p21, rad51, v-Ki-ras2, H-Ras
C2:FAMILY (FA) **(tag: family)**	GPRO families that can be associated to some gene/protein family (or group of GPROs). It includes groups of genes/proteins at the	death-associated protein family; Bcl-2 protein family; sirtuin deacetylase protein family; TRP protein family; alpha chemokine family; PIM family kinase; FOXO family; HER-family tyrosine

	sequence or taxonomic level. It comprises plural mentions of proteins assuming that they potentially refer to various different GPRO entities.	kinases; cdc2-like kinases; Janus kinases; tyrosine kinases; phosphoinositide 3-kinases; mammalian T-type calcium channels; mammalian xanthine oxidase; Mnk homologous proteins
C2:MULTIPLE (MU) (tag: multiple)	Mentions that do correspond to GPROs that are not described in a continuous string of characters. This is often the case of mentions of multiple GPROs joined by coordinated clauses.	Interleukin 1 and 2; BRCA 1 or 2; Rab1B, -5, -7, -8, or -11A; alpha, beta, or gamma PKC

Table 1. GPRO classes defined for the CHEMDNER-patents task. For each GPRO class a short description and illustrative example cases are provided. C1: GPRO entity mention type 1, C2: GPRO entity mention type 2 (described in section 2).

2. GPRO entities

The definition of GPRO entity mentions that were annotated for the CHEMDNER-patents task was primarily concerned with capturing those types of mentions that are of practical relevance (both for end users of the extracted data as well as for the named entity recognition systems). Therefore the covered GPRO entities had to be annotated at a sufficient level of granularity to be able to determine whether the labeled mention can or can not be linked to a specific gene or gene product (represented by an entry of a biological annotation database). The annotation carried out for the CHEMDNER GPRO task was exhaustive for the types of GPRO mentions that were previously specified. This implies that mentions of other entities such as chemicals or substances should not be labeled as GPROs.

We distinguish two types of GPRO entity mention types:

(1) *GPRO entity mention type 1:* covering those GPRO mentions that can be normalized to a bio-entity database record. GPRO type 1 includes the following classes: NESTED MENTIONS, IDENTIFIER, FULL NAME and ABBREVIATION

(2) *GPRO entity mention type 2:* covering those GPRO mentions that in principle cannot be normalized to a unique bio-entity database record. GPRO type 2 includes the following classes: NO CLASS, SEQUENCE, FAMILY and MULTIPLE.

In case of FULL NAME GPROs, they often correspond to descriptive terms denoting a particular GPRO entity. Descriptive GPRO names can be of various kinds; such type of names might refer to the *protein/gene function* (e.g. growth hormone, Gate keeper of apoptosis-activating protein), to *interaction properties* (e.g. fatty acid-binding protein, JNK-interacting protein 4), to their *cellular/subcellular localization* (Sperm surface protein, ER-resident protein ERdj3), to their *tissue/cell type/disease* expression (Hepatointestinal pancreatic protein), to their *species* of origin (HIV-1 envelope protein), to *similarity* to other proteins (Human liver DnaJ-like protein, Jerky protein homolog-like), to *physical properties* (Cell-scattering factor 140 kDa subunit), to their *association/implication in diseases* (e.g. Breast Cancer Type 2 susceptibility protein, Human lung cancer oncogene 7 protein), to their *structural or domain* properties (FYVE-RING finger protein Momo, Sterile alpha motif- and leucine zipper-containing kinase AZK), phenotypic characteristics (such as mutant phenotypes), uncharacterized entities (Uncharacterized protein C17orf80),etc,.. .

The FULL NAME class comprises also single word names that do not correspond to abbreviations. All FULL NAME GPROs can be normalized/linked to an entity database record.

The annotation process itself also relied heavily on (1) common sense, (2) domain background knowledge and (3) consultation of external resources of the annotators when labeling the GPRO entity mentions. A prerequisite in order to be able to carry out the manual annotation task was that the annotators must have an academic training in biology (molecular biology, genetics) or biochemistry to make sure the annotations are correct and of high quality. This also allowed us to have shorter and more compact annotation rules rather then requiring very detailed guidelines for non-experts.

GPROs for this task had to refer to names of specific genes/proteins/RNAs; specific classes of genes/proteins/RNAs or fragments of specific genes/proteins/RNAs.

General genes/proteins/RNA-related concepts (isolated terms like ‘gene’, ’receptors’, ’proteins’, ’mRNA’, ’peptide’, ’sequence’, ’transcript’, ’gene product’, ’domain’, ’isolate’, etc.), lipids, small organic molecules are excluded from the annotation task. GPRO concepts were annotated if they could be directly or indirectly linked to one or more GPRO entities (e.g. FAMILY type detailed below).

In order to label GPRO entity mentions a set of annotation rules were defined. Example cases were provided when possible to aid in understanding the different rules. The correct GPRO cases are marked in each example case.

As the annotation of GPROs depends on the specific the specific context of mention (the patent abstract text), we tried to provide in the guidelines proper descriptions of commonly encountered the context situations.

We have revised several previously existing corpora, their descriptions and guidelines (when accessible) for the preparation of the GPRO-patent corpus annotation guidelines. The corpora that we have revised included: GENETAG corpus (Tanabe et al. 2005), Gene Normalization corpus of BioCreative II, GENIA corpus (Kim et al. 2003), Yapex corpus (Franzén et al., 2002), JNLPBA corpus (Kim et al. 2004), MedTag corpus (Smith et al. 2005), ProSpecTome corpus (Kabiljo et al. 2005) and PennBioIE corpus (Mandel et al 2006).

The GPRO mention annotation rules are structured into the following 6 classes:

1. *G-rules (general rules)*
2. *P-rules (positive rules)*
3. *N-rules (negative rules)*

4. *C-rules (class rules)*
5. *O-rules (orthography/grammar rules)*
6. *M-rules (multi-word rules)*

We introduce some basic terms that are important for labeling GPRO-related mentions; these include core terms, feature terms and qualifier/modifier terms.

- *Core terms*: terms distinguishable by containing sentence-medial capital letters, numbers, non-alphanumeric characters. They are the actual core part of a name of a GPRO. They are very specialized lexical items. Examples: p53, BRCA2.
- *Feature terms*: a fixed set of keywords describing the function of nature of a core term. They can be *general feature terms* (gene, protein, transcript) or *specific feature terms* (those that refer to a group of genes or gene products that share an evolutionary common origin or correspond to some group of GPROs that can be grouped together base don sequence characteristics, e.g. kinases).
- *Qualifier terms*: they provide further constraints that are important for the database normalization of a GPRO (examples: human, mouse) or indicate some crucial modification of a GPRO (e.g. post-translational modifications like ‘phosphorylated’).

General annotation rules (G-Rules)

G1. Use of external knowledge sources
In case the curator is not sure if a mention corresponds to a GPRO or he does not know what kind of GPRO mention it is, he should consult external knowledge resources: UniProt, Wikipedia, NCBI, OMIM, GeneCards, model organism databases (e.g. MGI, SGD, RGD, FlyBase, etc) or other resources including the web (e.g. scientific papers, Wikipedia).

G2. Unclear mentions
Do not tag unclear cases. If the annotator is not sure about a given mention, even after consulting some external sources, the corresponding mention should remain unlabeled.

G3. Guideline under-specification
In case the annotator encounters cases of mention types that could be related to GPROs but the guidelines do not specify their labeling, these should be reported together with examples to request for refinement of the annotation rules.

G3. Exhaustive annotation
Annotate each and every GPRO mention found in the title and abstract, regardless if they are mentioned once or multiple times. Try to be consistent also in the way you normalize the mentioned entities and how you define the entity boundaries.

G4. Each GPRO mention can only be marked as a single GPRO class.
This means that a specific GPRO mention cannot be labeled for instance as FAMILY and FULL NAME at the same time (one mention - one class).

Positive annotation rules (P-Rules)

P1. GPRO Nouns referring to genes, gene locus, gene products (RNA, protein), gene/gene product families, mutations, sequences and domains.
Label GPRO nouns referring to genes, gene products (proteins, RNA), DNA/protein sequence elements (sequences and mutations) and protein families, domains and complexes (see table 1). Mentions of a specific gene, polypeptide or RNA product (including also rRNA genes, mRNAs, tRNA genes, snRNA genes, snoRNA genes, microRNAs, pseudo-genes, mitochondrial genes/gene products, ribosomal protein genes/gene products) and also of a specific non-protein-coding gene, or gene encoded in the mitochondrial genome is a sufficient mention as long as it clearly maps to a single gene/gene product (i.e. database identifier). Alternative transcripts or post-translational variants of a gene/protein are valid mentions of entities.

TP53; PDE4; IGF-1R; FGF18; COX-1; acetylcholinesterase; mGluR5 ; 28SrRNA ; mir-125 ; snoRNA:U3:9B ; mitochondrial NADH-ubiquinone oxidoreductase chain 4 ; RpL3 ;

In general, tag the minimal text span referring to the entity mention (flanking organism source names immediately adjacent to the GPRO and PTM prefixes should be included in the text span, see O6),

P2. GPRO class names
Label GPRO class names where the definition of the class includes information on structural/sequence GPRO family classes. ® GPRO = FAMILY class

cytokines; CCR; VEGF; chemokine; cathepsin, globin

P3. Enzyme classes
-Mentions of enzyme names and classes (with the exception of isolated mention of ‘enzyme/s’ itself (except when the core term is highly ambiguous or has less than 3 letters).

transaminase; lipase; decarboxylase; alanine racemase, tartrate epimerase; methylmalonyl CoA epimerase; UDP-glucose 4-epimerase
enzyme *not tagged*
T enzyme (Comment: core term ‘T’ is highly ambiguous and also has less than 3 letters; T enzyme corresponds to 1,4-α-d-glucan 6-α-d-glucosyltransferase)

P4. Natural antibodies
Natural antibody names should be tagged. Do not label mentions of antibodies/reagents that are used to study some target protein (only label the target).

P5. Hormones
Tag mentions of peptide hormones.
Insulin, Glucagon, growth hormone, oxytocine
estrogen *not tagged*

progesterone	*not tagged*
thyroid hormone	*not tagged*
ileal hormones	*not tagged*

P6. Sequence mentions

Specific sequence mentions that could be potentially mapped to protein, gene or genome sequences should be tagged in case they contain at least either a wild type residue/nucleotide mention together with positional information (number) or if there are a sequence of more than 3 sequence elements (residues or nucleotides). In the first case only the actual residue and number mentions should be tagged.

TGGAAATTCC
TGG *not tagged (should have more than 3 sequence elements)*

P7. Channels, pores, receptors and natural antigens

Channels, pores, receptors and natural antigens should always be tagged if they are mentioned together with adjacent modifiers that make them more specific. In case it is not clear which is the specific protein/gene record (e.g. in UniProt or other databases) they should be labeled as class FAMILY.

ion channel	FAMILY
potassium channel	FAMILY
K+ ion channel	FULL NAME (Q12809)
proton pump	FAMILY
human nuclear pore protein	FAMILY
Nuclear pore complex protein	FAMILY

Negative Rules (N-rules)

List of GPRO-related mentions that should not be tagged.

N1. Other terms different from GPRO nouns

Do not tag adjectives (if isolated/outside from GPRO nouns), pronouns, verbs, other terms (reactions).

- Enzymatic Reactions:

Dehydrogenation	*not tagged*
Methylation	*not tagged*
hydrolysis	*not tagged*

-Other terms

peptoid	*not tagged* (be careful, peptoids are poly-N-substituted glycines, do not confuse with peptides!)

N2. Experimental/reagent methods

Do not tag names that are part of experimental methods (if the entity corresponds to an experimental tool), including names restriction enzyme names.

DNase I hypersensitive sites HS2, HS3 and HS4	*not tagged*
polymerase chain reaction	*not tagged*
GFP (Green Fluorescent Protein)	*not tagged*

in vitro kinase assay	*not tagged*

N3. General sequence motifs

Very general motifs or type should not be tagged.

cAMP responsive element	*not tagged*
mu element	*not tagged*
4-mer RNA motif	*not tagged*
short RNA motif	*not tagged*
RNA oligonucleotides	*not tagged*

N4. Transposon/satellite rule

In general repetitive sequences should not be tagged.

LTR retrotransposon	*not tagged*
L1 element	*not tagged*
Copia retrotransposon	*not tagged*

N5. Pronouns/Anaphora rule

Do not tag anaphoric expressions referring to GPROs, only non-anaphoric explicit mentions of GPROs should be considered. Mentions that do not use a specific name such as ‘the gene’ or ‘the protein’ and pronouns should not be annotated.

“BRCA2 is expressed in breast tissue ….**this** gene can…” *do not tag “this”*

N6. Very general sequence elements

Sequence element names that cannot be linked specifically to a particular GPRO entity or group of GPRO entities should not be tagged. Do not annotate general sequence feature descriptors.

LTR	*not tagged*
Long terminal repeat	*not tagged*
Purine-rich binding sites	*not tagged*
HIV long terminal repeat	Should be tagged as it could be mapped to a specific viral genomic sequence position

N7. Experimental antibodies

Experimentally generated antibodies, antibodies used for disease diagnostics or therapy should not be tagged themselves. In case their target is a GPRO, than only the GPRO should be labeled (and not the Antibody mention). Isolated word tokens like *antinuclear*, *antibody* or antibody related prefixes like *anti*- or *Ab*- should not be labeled.

Antibody	*not tagged*
anti-BRCA2	

N8. Chromosomes/genomes

Do not tag mentions of chromosomal positions or chromosomes or genomes. This means that if a mention refers to a whole genome, chromosome or large locus (that covers different genes), chromosomal abnormalities it should not be labeled.

Simian immunodeficiency Virus (SIV) genome

Human genome	*not tagged*
Chromosome 5	*not tagged*
Chr 5	*not tagged*

“RNAi Modulation of RSV, PIV and Other Respiratory Viruses and Uses Thereof … methods that are useful in reducing RSV or PIV mRNA levels, RSV or PIV protein levels and viral titers in a subject … “-> not tagged (comment: refers to the entire set of mRNAs or proteins of a particular species/virus)

Del(9)4H 5	*not tagged* (refers to a deletion involving Chr 9, the 4th deletion)
MatDi(12)	*not tagged* (refers to maternal disomy for Chr 12)
Tel19q2	*not tagged* (refers to the second telomere mapped at the distal end of Chr 19)

N9. Functional gene/protein groups

Do not annotate functional types of GPROs that cannot linked to a particular group of entities that share an evolutionary origin (sequence similarity, orthologues and homologues).

DNA-binding proteins	*not tagged*
Transcription factors	*not tagged*
transmembrane proteins	*not tagged*
secreted proteins	*not tagged*
penicillin-binding proteins	*not tagged*
oncogenes	*not tagged*
anti-apoptotic proteins	*not tagged*
ion channel	*tagged (exception of protein groups: receptors, channels, pores and natural antigens. Class FAMILY)*

N10. Synthetic peptides and synthetic experimental fusion genes

Do not label synthetic peptide or construct mentions as well as synthetic fusion genes.

Melanoma-based polyepitope construct proteins	*not tagged*

N11. Cell lines

GPRO mentions embedded in cell line or cell type names should not be tagged, with the exception of GPRO mentions that are explicitly described to be expressed in a cell line/type.

NIT-1 cells	*not tagged*
CD4= lymphocytes	*not tagged*
Fz4-WT–expressing cells	(is not a cell line therefore the gene is tagged)

N12. Transgenic animal models

Do not label names of transgenic animal models.

H-2 class I negative mice	*not tagged*
HLA-A2.1 transgenic HHD mice	*not tagged*

N13. General vague domains
Do not tag very general/vague domain mentions/descriptions.

a targeting domain	*not tagged*
extracellular domain	*not tagged*
extracellular cystein-rich domain	*not tagged*
transcription activator domain	*not tagged*

N14. Epitopes
Do not tag epitope mentions.

human tumor-associated CTL epitopes	*not tagged*

N15. Genetic diseases
General mentions of genetic diseases that may be caused by variations in a particular polypeptide, gene, chromosomal abnormalities etc. mentioned in the context of the disease and not named as a name for protein or gene do not count as mentions. Only if a specific GPRO is mentioned in a disease context in which that GPRO should be tagged.

GLC1A glaucoma	
BRCA1 breast cancers	
Angelman syndrome	*not tagged*
Sickle-cell disease	*not tagged*
5p deletion syndrome	*not tagged*
Alpha 1-antitrypsin deficiency	

N16. Entity surrogates
Do not tag surrogates of GPRO entities if they do not specifically mention a GPRO.

protein inhibitor mimicking organic molecule compound	*not tagged*

Nevertheless in case they do mention a GPRO as in the examples below you have to tag them.

NCX inhibitors
CCR antagonist
IGF-1R inhibitor
H4 antagonists
Alpha-2 adrenergic agonist

N17. Isolated nonspecific general GPRO feature concepts
Do not tag <u>isolated</u> <u>*general*</u> feature terms. Feature terms (e.g. protein, gene) consist of those terms that determine the semantic category of the GPRO, they often appear in text before (modifier, e.g. *protein* p53) or after (head, e.g. p53 *protein*) a GPRO

mention. General feature terms are those that do not provide any protein/gene family information. If they are mentioned without an adjacent GPRO they should be never tagged.

*subunit, mRNA, promoter, protein, gene, receptor, peptide, sequence, transcript, gene product, domain, isolate, dimer, homodimer, oncogene, polypeptide, mutant, mutation, element, activity, signaling, expression, complex, multimer, molecule, polynucleotide, amino aci*d*, pore, channel, receptor, antigen, mutated, mutant*

Exception: viral gene/protein mentions.
Most viral proteins or genes do have very general names such as envelope polyprotein, matrix protein, RNA polymerase, glycoprotein, replication protein, capsid protein, etc,.. Despite those general names when it is possible to normalize those mentions to some concrete database record (from the context it is clear to which virus species and database record they belong to) those mentions should be labeled and linked to their corresponding database identifier. When the context only specifies the virus family or genus those mentions should be labeled with the class FAMILY (see rule C11.)

N18. Single residues
Do not label amino acid or nucleotide mentions that do not refer to parts of genes or proteins (requires associated sequence position specifications). Watch/careful: unless they are found in protein mutation.

l-histidine	*not tagged*

N19. Context Criteria.
Words are not GPRO if they are not GPRO in context, even if they are co-incidentally the same set of characters (synonyms and metaphors). For example, there is a Drosophila gene with the name ‘wing’; it should only be tagged when from the context it is clear that it refers to a gene, otherwise it would not be tagged (e.g. when it refers to a part of the body).

The wing of the bird was large …	*not tagged*
The drosophila wing gene is expressed in …	

N20. Prefixes based on phenotype, EST or STS.
Some generic gene symbol/name prefixes have been used for genes sharing a common mutant phenotype or originally identified by virtue of an EST or STS. Those should not be tagged. Some examples of those are: *anon-* (anonymous gene), *BEST* (Berkeley Drosophila Genome Project EST cluster-based gene), *e(a)m, E(a)m* (enhancer), *fs(n)m, Fs(n)m* (female sterile), *l(n)m* (lethal)

Class Rules (C-rules)

List of rules that specify how to label each mention according to the class of GPRO that they belong to.

C1. One mention-one class
Each annotated GPRO mention should be labeled only with one of the following GPRO class tags: NO CLASS (NC), NESTED MENTIONS (NM), IDENTIFIER (ID), SEQUENCE (SE), FULL NAME (FN), ABBREVIATION (AB), FAMILY (FA) and MULTIPLE (MU).

C2. Bacterial enzymes
Bacterial enzyme mentions that are formed of multiple proteins -> tag as FAMILY .
We introduce this rule because in most cases bacterial enzymes are composed of multi-protein complexes.

C3. Name aggregations
Aggregated names corresponding to multiple GPROs should be annotated as FAMILY if the entity mention does not contain any whitespace or special character, else it should be annotated as MULTIPLE.

queCDEF	should be annotated as FAMILY as it corresponds to queC, queD, queE and queF.
queC and D	should be annotated as -> MULTIPLE

C4. Regulon/operon
Mentions of regulons or operons should be annotated as FAMILY

C5. Protein families
Mentions of protein families or gene groups should be tagged as FAMILY. Functional groups should generally not be tagged.

C6. Domain names
References to domain names should be tagged as SEQUENCE if it is clear that they refer to a domain and not to a particular gene/protein.

C7. AND rule
In case the actual name of a protein has as part of the proper name (e.g. contained as such in a database record) ‘and’ it should be labeled as TRIVIAL, otherwise as MULTIPLE

C8. Nested mentions of same entity
Nested mentions of the same entity should be labeled as class NESTED MENTION.

Serum Amyloid (SAA) A gene
ribosomal protein S6 (p70S6) kinase

C9. Determiner-rule
In case a potential GPRO mention is preceded by a determiner like ‘a’ or another linguistic element indicating that there might be various different GPROs associated to that mention, than it should be labeled as FAMILY (unless the various different GPROs are isoforms of the same gene: isoform exception).

a TRPV1-binding protein -> FAMILY (comment: if it indicates one of several/many)

C10. Unclear domains

In case it is not clear if the authors mean a domain or sequence motif, it should be annotated as class SEQUENCE if is contains amino acid or nucleotide sequences, otherwise it should be labeled as 'NO CLASS'.

C11. Non-species level taxonomic information

If the context explicitly refers to a taxonomic source of the GPRO mention that is not at the level of species (e.g. they refer explicitly to mammalian, eukaryotic, primate), than the GPRO mention should always be annotated as FAMILY.

mammalian TP53 -> FAMILY

C12. Pathway rule

Mentions of GPRO names referring to pathways should be annotated as FAMILY as they refer to a group of proteins.

inhibition of the VEGF pathway gene expression and inhibition -> FAMILY

C13. Protein complex rule

Protein complexes that have a proper name not formed by the individual constituent proteins should be labeled as FAMILY. If a descriptive term (e.g. 'complex') is immediately adjacent to this kind of mention it should be included in the markup. In case the protein complex members are explicitly given, they should be marked up as separate entities.

TFTC complex
BRCA1–RAP80 complex

C14. NO CLASS (NC)

This class covers names of individual protein *domains* and names of *sequence* or *structural motifs*. This class includes also *DNA/RNA structural motifs* and identifiers of protein domains (PFAM).

Epidermal growth factor-like domain
Sh2 domain
PDZ domain; MAM domain; SH3 domain; CARD domains; CD74-homology domain; ATPase-like motifs

C15. NESTED MENTIONS (NM)

This class covers *nested mentions of a single entity*. This would correspond to nested mentions where the actual decomposed entity mentions could be normalized to one unique entity.

ribosomal protein S6 (p70S6) kinase

C16. IDENTIFIER (ID)

Database identifiers of genes or gene products (proteins, RNA). This includes identifiers and/or accession numbers from UniProt, GenBank, RefSeq, Ensembl, PDB, HGNC. Here SNP identifiers should be included too. This class includes mentions of enzyme commission numbers, so called EC numbers.

P35354; PGH2_HUMAN (UniProt); 3NSS (PDB); ADA71175; NM 006475 (GenBank); rs12979860; rs12153855; rs17207923; EC:2.7.11.1; EC 2.3.2.5

In case of EC numbers that have less than 4 digits they should be labeled as FAMILY.

C17. SEQUENCE (SE)

Mentions of protein (amino acid) sequences, nucleotide sequences, mutation and residue mentions of DNA, RNA and proteins. Also includes: promoter sequences, sequence motifs.

CACGTG; SQEY motif; VGFPV motif; 5′-C/U-U-G/U-U-3′; C3592T; G3602A; G12V; Gly12Val;
substitution of glycine 21 to valine

C18. FULL NAME (FN)

Full name of a GPRO, including names of precursor proteins (in case of cleaved proteins). It also includes multi-word terms referring to specific gene/protein named entities. It covers single word GPRO what do not correspond to abbreviations or symbols.

ribosomal protein S6;
DAP kinase-related apoptosis-inducing protein kinase 2; Serum amyloid A3; cyclooxygenase-2; Heat shock protein 90; human deoxycytidine kinase protein; human somatostatin 2 receptor protein; thromboxane synthase; Kirsten rat sarcoma viral oncogene homolog; tumor protein p53

C19. ABBREVIATION (AB)

Abbreviation of full name GPROs, GPRO acronyms, gene/protein symbols or symbolic names. Also two-word GPROs where one of the words is an Abbreviation and the other word is a number or single letter (e.g. Cox 2; H Ras; miR 145).

Drak2; p70S6; SAA; COX2;
miR-145; SETD1B; MIR4304; TNXB; NOTCH4; IFNα-7; grk5; KCTD1; HSATU68; SCN9A, p21, rad51, v-Ki-ras2, H-Ras

C20. FAMILY (FA)

This class covers GPRO families that can be associated to some *gene/protein family* (or *group of GPROs*). It includes groups of genes/proteins at the *sequence or taxonomic level*. It comprises *plural mentions* of proteins assuming that they potentially refer to *various different GPRO* entities. Note: Terms that indicate that the GPRO refers to a family of group should be part of the entity mention tag.

death-associated protein family; Bcl-2 protein family; sirtuin deacetylase protein family; TRP protein family; alpha chemokine family; PIM family kinase; FOXO family; HER-family tyrosine kinases; cdc2-like kinases; Janus kinases; tyrosine kinases; phosphoinositide 3-kinases; mammalian T-type calcium channels;

mammalian xanthine oxidase; Mnk homologous proteins; collagen

C21. MULTIPLE (MU)

This class addressed mentions that did correspond to GPROs that are not described in a continuous string of characters. This is often the case of mentions of multiple GPROs joined by coordinated clauses or enumerations of GPROs names (often used to avoid redundancies). Also parts of names divided by long text passages fall into this class. The dependencies of the partial GPROs mentions are not captured in this version of the task. Such MULTIPLE mentions could be decomposed later defining the dependencies, chaining rules or alternative allowed mentions in a second step if needed.

Interleukin 1 and 2
BRCA 1 or 2
Rab1B, -5, -7, -8, or -11A
alpha, beta, or gamma PKC

C22. Fusion proteins and protein complexes

Mentions of fusion proteins (e.g. Bcr-Abl) should be labeled as two separate entities except for cases where the fusion is a name that cannot be directly decomposed into its constituents. The same is true for protein complexes. In case the protein complex cannot be decomposed into its constituents it should be labeled as FAMILY.

C23. Mutation mentions

Mutations are only tagged if at least the wild type residue and its position are specified. If this is the case the wild type residue, the mutant residue (if described) and the sequence position should be tagged. Mentions of sequence positions without specifying the residue should not be tagged. Those mentions would be labeled as class *SEQUENCE.*

Orthography/Grammar Rules (O-Rules)

O1 Other languages

Names in other languages than English should be annotated regardless the language according to the general annotation rules and GPRO classes.

O2 Mis-spellings & conversion errors

Mentions of GPROs (as long as they follow some of the other mention rules) that are misspelled should be tagged. This also includes mentions suffering from automatic conversion errors generated by text conversion programs.

BRCAl	*where 1 is "one" not "l"*
Interleukin2receptor, …	*where it should be "*Interleukin 2 receptor*"*

O3 "A B" wrong space

White space-separated words that should properly be a single word should be marked up as single entity.

… the BRCA 1 group was …

O4 GPRO named after people

Mentions of GPRO named after people should be tagged if they clearly refer to a GPRO entity.

O5 Sentence boundary

GPRO entity mentions cannot span multiple sentences.

O6 Not flanking white space characters

Do not tag white space characters flanking the GPRO. Annotators should try to define the mentions precisely, and not include flanking whitespace or other spacing characters.

O7 Not Commas, full stops, brackets

Do not include as part of the GPRO: off commas, full stops, brackets, and references to papers etc. that aren't a part of the name itself. Do include as part of GPRO the square brackets around complexes.

Breast Cancer Type 2 susceptibility protein (BCRA2)
BCRA1/BCRA2 complex
BCRA2 [4]

O8 Include prefixes for modifications (post-translational modifications) or species

Include in the GPRO label prefixes/postfixes that denote the (1) genus/species of a gene's origin (e.g. *D*, *Dro* or *Dm* for D. melanogaster) as well as those related to (2) the location of the gene (nuclear, mitochondrial, chloroplast; e.g. *mt* for mitochondrial genes) as well as those corresponding to post-translational modifications (PTMs) of proteins such as phosphorylations. These should only be tagged if they are adjacent to the GPRO, in case species or modifications are mentioned isolated, they should not be tagged.

hGluR6	*ABBREVIATION*
human GluR6	*FULL NAME*
pTSC1	*ABBREVIATION*
phosphorylated TSC1	*FULL NAME*
mt-Atp6	*ABBREVIATION*
n-Ta12 (nuclear encoded tRNA alanine 12)	*ABBREVIATION ; FULL NAME*
n-R5s104 (nuclear encoded rRNA 5S 10)	*ABBREVIATION ; FULL NAME*
.. of h1n1 influenza virus infection ..	*not tagged (isolated species mention)*

preventing human or animals diseases	*not tagged (isolated species mention)*

O9 Not Trademarks

Do not include trademark symbols as part of GPRO

O10 Not tailing hyphen/apostrophe

Do not tag tailing hyphens or the apostrophe-s in possessives.

GluR6-mutant	*ABBREVIATION*
BRCAl-expression	*ABBREVIATION*
SMA6-induced transcription	*ABBREVIATION*
PKM2's activity	*ABBREVIATION*

O11 Numbers in GPRO sequences and numbers as part of the GPRO name

Include numbers of positions of mutations

Tyr234Ser	*SEQUENCE*
Tyrosine 234 change to serine	*SEQUENCE*

Include when present the numbers that are required to specify the actual GPRO entity, often the case when a group of GPROs has multiple members:

BRCA genes	*FAMILY*
BRCA 1 protein	*ABBREVIATION*
BRCA 2 gene	*ABBREVIATION*

O12 Not quotation marks

Do not tag quotation marks as part of GPROs.

('mGluR2')	*ABBREVIATION*

O13 Alleles, superscripts and mutant symbols and phenotypes due to Mutations

The symbol for mutant alleles is usually formed by adding the gene symbol the original mutant symbol as a superscript. In the case of wild type alleles of a gene it is indicated by + as superscript to the mutant symbol and reversions to wild type are usually indicated by the symbol + with the mutant symbol as superscript. All these cases should not be labeled as part of the GPRO mention. Exception: when an allele name is an integral part of a gene symbol or name or it is embedded within the GPRO mention itself, nevertheless such cases are very rare. All pre- or postfix superscript allele, mutant or phenotype specifiers should NOT be included in the mention. (Careful: note that opposed to this, post-translational modifications (e.g. phosphorylation) should be included in the mention).

Kit^{+} wild type *Kit* locus
Grid2$^{ho\text{-}cpr}$.

Kit$^{W\text{-}v}$
Kit$^{W\text{-}sh}$
Myo5a^{d+}
Myo5a^{d+}
Crb1$^{rd8+em1Mvw}$
su(w^{a}) Exeptional case: D. Melanogaster gene: suppressor of white-apricot

O14 Transgenes
Transgenes are produced by homologous recombination as targeted events at particular loci. Transgenes are usually denoted by the prefix *Tg* (see examples below). This prefix should not be included as part of the GPRO mention.

O15 3 character core names
In case the GPRO mention is only 1 or 2 characters long, than immediate adjacent function terms should also be added to the labeled mention tag, both general feature terms (e.g. protein, gene, etc.) as well as specific feature terms (kinase, sulfurtransferase, carboxypeptidases).

F protein *(core term with < 3 characters add flanking feature term)
H1 receptor (Given some particular context when referring to the Histamine H1 receptor)

O16 Highly ambiguous core names
In case the GPRO mention, if taken out of context, is highly ambiguous (i.e. it corresponds to a common English word), than immediate adjacent function terms should also be added to the labeled mention tag, both general function terms (e.g. protein, gene, etc.) as well as specific function terms (kinase, sulfurtransferase, carboxypeptidases).

O17 Mutations/fragments as part of gene symbols
In case a mutation position or fragment specification is provided as part of a gene/protein symbol, it should not be labeled as part of the mention.

Commented example: rvtA11
In this example do not annotate *11* as part of the mention if it is clear from the context that it specifies a mutation position (and is not really part of the entity name).
rvtA would be a mention of type ABBREVIATION, while *11* would not be tagged unless in the context it is clear what was the wild type residue at position 11, in that case if would be a separate mention of type SEQUENCE together with the wild type name.

O18 Inhibitor/receptor rule (boundary)
Entity adjacent feature terms like *inhibitor* or *receptor* should only be tagged in case they are a proper part of the entity name (e.g. it could be found also in a corresponding database record of that entity).

MULTIWORDS: SINGLE ENTITIES vs. MULTIPLE ENTITIES

M1 The longest GPRO should always be tagged, but only including those words that are actually part of the GPRO name and taking into account the special rules for immediate adjacent flanking (before and after) words and prefixes. Non-essential parts of the GPRO entity and name modifiers should NOT be tagged:

GluR6 protein
F protein *(core term with < 3 characters add flanking feature term)
DAP kinase
acetylcholinesterase enzyme
BCRA2 molecule
SERPINF2 gene
ackA1 mRNA
AGTR1 transcript
zinc finger protein 16 expression
ATP6D polypeptide
human interleukin 1 protein
Wings apart-like protein homolog mutant
TP53 wild type
atf4 transcription factor
Anti- STATI2 Antibody
alpha-1-B glycoprotein
mutated GluR6
GluR6 mutant

Note: do not include feature terms if they do not provide specific information on the core term (Exceptions: if they are part of an accepted GPRO named contained in databases, the core term is less than 3 characters or highly ambiguous).

M2 Adjectives with valid GPROs

Adjectives are only to be annotated if i) precede/follow a valid GPRO entity and ii) add more precise information to this GPRO entity relevant for its database normalization (species/taxonomic information or adjectives that are part of a valid database record GPRO name) or if it refers to **post-translational modifications** of the GPRO (e.g. acetylation, amidation, disulfide bonds, glycosylation, methylation, phosphorylation, esterification, ubiquitination, hydroxylation, biotinylation, ADP-ribosylation or sulfation. Include also adjectives expressing the removal of PTMs, like dephosphorylation, deacetylation or deubiquitination). The whole concept (adjective + GPRO noun) should be tagged as a unique GPRO entity assignable to the GPRO class of the GPRO entity alone. This is independent on the origin of the root name of the adjective and on the adjective ending ("-ed", "-ing","-ic").

acetylated glycogen phosphorylase
modified NFAT4
phosphorylated NFAT4

phosphorylating STAT
N-linked glycosilated ABCG2
dephosphorylated STAT
ABCG2 glycosylated protein

M3 Adjectives with general feature terms

Adjectives are only to be annotated if they precede/follow a core GPRO term (or GPRO mention that contains a core term) and not if they precede/follow only an isolated general feature term.

It is a phosphorylated protein expressed in	*not tagged (general feature term)*

M4 Adjectives with specific feature terms

Adjectives that precede/follow an isolated specific feature term should be included in the tagged mention (FAMILY).

a phosphorylated human kinase	*FAMILY*

M5 Negative adjectives

"Negative" concepts that discard specific GPROs should not be tagged.

non-TP53	*tag only TP53*
noninsulin	*tag only insulin*

Exceptions here are negative adjectives that 'target' taxonomic or PTM related information, then tag the corresponding adjective:

non-human TP53	*tag also the negation (FAMILY)*
unphosphorylated NFAT4	*tag also the negation (FULL NAME, Q12968)*
non-phosphorylated NFAT4	*tag also the negation (FULL NAME, Q12968)*

M6 Enumerations and list of GPROs vs multiple entities:

If full names are enumerated, tag separately each individual GPRO:

BCRA1 and BCRA2
ULK4, ULK3
GLI1, GLI2 and/or GLI3
KIAA0261/KIAA1613

If GPROs are not described in a continuous string of characters tag the whole string (including words such as "and", "or" and commas) as a single entity of class type multiple. Avoid the generation of "half truths".

BCRA-1 and 2	*tag as MULTIPLE*
GLI1 as well as 2	*tag as MULTIPLE*
21 or 53 kDa proteins	*tag as MULTIPLE*

M7 ‘GPRO1 GPRO2’ : a single GPRO or two GPROS?

If there are two continuous words of type GPRO: “GPRO1” and “GPRO2”, each of which would individually be of class GPRO:

- if they denote a single entity - label as a unique single GPRO
- if they denote different GPRO entities® label as independent GPRO’s

BRCA1 BRCA2

M8 Related sequence and pseudogene symbols

In case of mentions of related sequence and pseudogene symbols do not split the hyphenated specifiers from the root.

Hk1-rs1	(corresponds to hexokinase-1 related sequence 1)
Hba-ps3	(corresponds to hemoglobin alpha pseudogene 3)

M9 Database GPRO names

Database record gene entry names might override the gene boundary definitions provided in the guidelines.

“Alpha-1-B glycoprotein is a 54.3 kDa protein in humans that is encoded by the A1BG gene.”

Alpha-1-B glycoprotein	-> tag as FULL NAME, normalize to UniProt: P04217
54.3 kDa protein	-> tag as FAMILY
A1BG	-> tag as ABBREVIATION, normalize to UniProt: P04217

GPRO Grounding guidelines

1. GPRO type 1 with database identifier

All those GPRO mentions of type 1, i.e. those GPRO mentions that can be normalized to a bio-entity database record (NESTED MENTIONS, IDENTIFIER, FULL NAME and ABBREVIATION) have to be associated to a valid database identifier.

2. Valid database list
Database identifiers that can be used fro the normalization of genes or gene products (proteins, RNA) include identifiers and/or accession numbers from UniProt, GenBank, RefSeq, Ensembl, HGNC, model organism database identifiers (e.g. from MGI, RGD, FlyBase,..) and SNP identifiers.

3. Normalization preference order
When possible follow the following order to preference:

UniProt > GenBank > HGNC > RefSeq > Ensembl > model_organism_db > other databases

4. One mention one identifier rule
Only one single database identifier should be used for a given GPRO mention.

5. All mentions of the same GPRO with same identifier
In case there are several mentions that are aliases or synonyms or typographical variants of the same entity (after checking the context of mention), try to be consistent in normalizing all of them to the same database identifier. In such cases, the mention that is the most discriminative one should be considered for the manual linking process.

'*Oct4* (also known as *Pou5f1*)'.

6. Normalization resources
You are free to use the necessary query terms (e.g. GPRO names/symbols) together with some other terms (e.g. species names or taxonomy identifiers) to consult the necessary external knowledge resources (UniProt, Wikipedia, NCBI, OMIM, GeneCards, model organism databases like MGI, SGD, RGD, FlyBase, miRBase, etc. or other resources including the web (e.g. scientific papers, Google) to assign the correct database identifier to the GPRO mention.

7. Multiple protein isoforms
If there are several isoforms and it is not sure to which one the authors refer then one should normalize the mention to a Gene Identifier and not to a protein identifier.

8. Following database cross-links for normalization
In case you are able to normalize a given GPRO mention to a database (e.g. GeneCards), and this database provides a link to another database that has a higher preferential normalization order (e.g. UniProt), you should use the high order database as a reference resource instead, but only if there is no additional level of ambiguity associated to the out-link resource. Here no more than a maximum of two steps of following database outlinks should be done.