Annotating Metabolic Processes: Guidelines

In the following sections, for each type of annotation, we describe the scope (i.e., what should be considered for marking up) and span (i.e., what exactly should be marked in text). In the examples provided, spans which should be annotated are shown inside square brackets, e.g., [tyrosine]; whereas those which should *not* be annotated are additionally crossed through, e.g., [~~antimicrobial compounds~~].

Annotation of chemical compounds (CCs)

Annotation scope

All mentions of specific chemicals, classes of chemicals and fragments of chemicals should be annotated as CCs. Note that for this annotation task, genes or gene products (e.g., proteins) are also considered as chemical compounds, but will be annotated separately as GGPs according to our guidelines for Annotation of genes or gene products.

✔Include:

1. Anything that can be represented with a chemical structure diagram, including structural classes.

   > However, Src treatment led to [tyrosine] phosphorylation of p27 and catalytic activation of assembled cyclin D1-Cdk4-p27 complexes.

   > The products of this enzyme, i.e. [inositides] phosphorylated in the D3 position of the [inositol] ring, may act as second messengers themselves.

2. Different types of chemical names (e.g., systematic, semi-systematic, trivial, brand), formulas, CAS registry numbers, SMILES and InChI strings.

   > Breast cancer is characterized, among others, by the concurrence of lipophilic xenobiotica such as [2,3,7,8-tetrachlorodibenzo-para-dioxin] with hypoxic tissue conditions.

   > Consistent with a role of this phosphatase on cell wall physiology, cells lacking Msg5 displayed an increased sensitivity to the cell wall-interfering compound [Congo Red].

   > Two novel tetrasaccharides were discovered with proposed structures of [DeltaUA2S-GlcNS-IdoUA2S-[(35)S]GlcNH(2)3S] and [DeltaUA2S-GlcNS-IdoUA2S-[3-(35)S]GlcNH(2)3S6S].

3. Abbreviations, including unofficial ones. The full name and abbreviation should be annotated separately.

   > Although 2-O-sulfated [L-iduronic acid] ([IdoA]) residues have been known to occur in [heparin], 2-O-sulfated [D-glucuronic acid] ([GlcA]) residues have been reported only recently.

✖Exclude:

1. General concepts (e.g., biological, laboratory roles), as well as nonspecific structural concepts.

   > Crude extracts and column fractions from the red algae Asparagopsis taxiformis and A. armata

from the Strait of Messina (Italy) were screened for the production of [~~antimicrobial compounds~~].

Their strategy comprised the coupling of an α-amino acid with a [~~heteroketene dimer~~].

2. Genes or gene products (GGPs), which are marked up separately.

   To prepare oligosaccharides, chitosan can be hydrolyzed by a mixture of different enzymes ([~~cellulose~~], [~~alpha amylase~~], [~~proteinase~~]).

3. Chemical reactions and chemical adjectives.

   For this purpose the chitosan or chitin methanolysis/hydrolysis products are [~~trimethylsilated~~] or [~~acetylated~~].

4. Compound numbers arbitrarily assigned by authors.

   In this study, we found that [epimuqubilin A] ([~~1~~]) possessed potent NO production in LPS-stimulated RAW 264.7 cells and [sigmosceptrellin A] ([~~6~~]) exhibited significant selectivity in the GSK-3β pathway. Compounds [~~1~~] and [~~6~~] have been reported in treatment of human African trypanosomiasis against Trypanosoma brucei.

5. Descriptive references, even if by reading them the annotator can identify an entity in a relevant database.

   [Smenospongine], [~~a sesquiterpene aminoquinone isolated from the marine sponge of Dactylospongia elegans~~], induced G1 arrest in K562 cells.

Annotation span

In general, only the minimal span of text containing the name of the CC should be annotated and nothing else.

✔Include:

1. Complicated structures (e.g., prepositional phrases and coordinating conjunctions) only if they appear as part of a CC name.

   The [Active Peptide from Shark Liver] ([APSL]) was expressed in E. coli BL21 cells.

2. The common fragment(s) in the case of coordinated names and entity ranges. The number of annotated spans should correspond to the number of entities identified by the annotator, except in entity ranges where only the initial and final entities in the range should be counted.

   [Calyculins] and Related Marine Natural Products as Serine-Threonine Protein Phosphatase PP1 and PP2A Inhibitors and Total Syntheses of [[[Calyculin A], B], and C]

   Due to these structural similarities to [manoalide] (1), [[petrosaspongiolides M]–R] have received special attention from the scientific community to study their inhibitory activity against PLA2 from different resources to point out their specificity.

3. Intervening text corresponding to embedded abbreviations/full forms/aliases within the name.

   [N,N-Didesmethylgrossularine (DDMG) -1], a compound with a rare α-carboline structure, was isolated from an Indonesian ascidian Polycarpa aurata.

✖Exclude:

1. Modifiers which are not part of the name.

Heparan sulfate D-glucosaminyl 3-O-sulfotransferase-3A sulfates [~~N-unsubstituted~~] [glucosamine] residues.

Two [~~3-OST-3A-modified~~] [tetrasaccharides] were purified from the [~~3-O-(35)S-sulfated~~] [heparan sulfate] that was digested by heparin lyases.

2. Head words which are not part of the name.

While SSAT-1 mRNA was inducible by [polyamine] [~~analogues~~] in a variety of cell lines, SSAT-2 was not.

The results demonstrate that 3-OST-3A sulfates N-unsubstituted [glucosamine] [~~residues~~].

3. Characters appearing in the same token as the name but are not part thereof.

Indeed, MAP4K3 is required for phosphorylation of known mTOR targets such as S6K1 (S6 kinase 1), and overexpression of MAP4K3 promotes the [rapamycin][~~-sensitive~~] phosphorylation of these same targets.

These findings suggest that PI 3-kinase, through the Rho-GAP homology domain of p85, can couple to the effector domain of Cdc42Hs and that p85 may be a target for the [GTP][~~-bound~~] forms of Cdc42Hs and Rac1.

Linking to ChEBI Database

Each chemical compound annotation should be linked to an entry in the ChEBI database. Based on the compound name and the context surrounding it in the document, the annotator should be able to identify the corresponding CheBI entry.

Annotation of genes or gene products (GGPs)

Annotation scope

In general, only specific names of genes or gene products should be assigned the label of GGP. A mention is considered specific if it allows the annotator, after considering textual context, to identify the entity in a relevant database (e.g., UniProt). This does not mean, however, that only names or aliases found in databases are acceptable; any name considered by the annotator as referring to a specific gene or gene product should be annotated.

✔Include:

1. Gene, protein and RNA. No distinction is made between these types.

The expression of [DNase] as well as the [slo] and the [scpC] genes in [emm1]-genotype strains was enhanced under the [csrS] mutation.

2. Mentions of genes or proteins occurring as parts of other names.

> [Cdc25A] is a [tyrosine phosphatase] that is involved in the regulation of the G1/S phase transition by activating [cyclin E]/[Cdk2] and [cyclin A]/[Cdk2] complexes through removal of inhibitory phosphorylations.

3. Abbreviations, including unofficial ones. The full name and abbreviation should be annotated separately.

> PGN caused time-dependent activations of [IκB kinase αβ] ([IKKdβ]) and [p65] phosphorylation at Ser276, and these effects were inhibited by NS398 and KT5720.

> When cells were stimulated with [insulin-like growth factor-1] ([IGF-1]), an increased interaction between [hBVR] and [PKCδ] was detected by FRET-fluorescence lifetime imaging microscopy.

✖Exclude:

1. Names of gene/protein families or groups.

> Activation of [IRF3] requires signal-dependent phosphorylation, but little is known about the signaling pathway or [~~kinases~~] involved.

2. Names of gene/protein domains or regions.

> Proteins containing a so-called [~~GGDEF~~] domain are responsible for the synthesis of c-di-GMP, and those with a so-called [~~EAL~~] domain for its degradation.

3. Descriptive references, even if by reading them the annotator can identify an entity in a relevant database.

> [Proteinase 3], [~~Wegener's autoantigen~~]: from gene to antigen.

Annotation span

In general, only the minimal span of text containing the name of the GGP should be annotated and nothing else.

✔Include:

1. Complicated structures (e.g., prepositional phrases and coordinating conjunctions) only if they appear as part of a GGP name.

> Acetylcholine leads to [signal transducer and activator of transcription 1] ([STAT-1]) mediated oxidative/nitrosative stress in human bronchial epithelial cell line.

> Here we show that increased expression of [zinc-finger protein regulator of apoptosis and cell-cycle arrest] ([Zac1]) is implicated in apoptosis in F9 and P19 EC cells.

2. The common fragment(s) in the case of coordinated names and entity ranges. The number of annotated spans should correspond to the number of entities identified by the annotator, except in entity ranges where only the initial and final entities in the range should be counted.

> Immunoblot assays of homogenate from atria, ventricles, and septa of 14 nonfailing human hearts established expression of [[[[Na,K-ATPase alpha1], alpha2], alpha3], and beta1].

> Moreover, activity of [[MMP-10] to 12] co-localised with markers of classical activation in human atherosclerotic plaques in vivo.

3. Intervening text corresponding to embedded abbreviations/full forms/aliases within the name.

> [Toll-like receptor (TLR) 2] and [TLR4] are implicated in the recognition of various bacterial cell wall components, such as lipopolysaccharide (LPS).

> Mice with inactivated [glycogen synthase kinase (GSK)-3beta] die from hepatocyte apoptosis during development.

4. Names of GGPs appearing in complexes.

> … that is involved in the regulation of the G1/S phase transition by activating [cyclin E]/[Cdk2] and [cyclin A]/[Cdk2] complexes through removal of inhibitory phosphorylations.

✖Exclude:

1. Modifiers which are not part of the name.

> Throughout interphase, [~~human~~] [Myt1] localizes to the endoplasmic reticulum and Golgi complex.

> The [~~mouse~~] [UPase-2] cDNA was also identified and shown to be predominantly expressed in liver.

2. Head words which are not part of the name.

> In vitro analysis demonstrates that [p53] directly binds to a [TRRAP] [~~domain~~] previously shown to be an activator docking site.

> The recombinant [nim1] [~~protein~~] autophosphorylates on both tyrosine and serine residues and can phosphorylate the isolated [wee1] [~~protein~~] directly in a cell-free system.

3. Characters appearing in the same token as the name but are not part thereof.

> The [nim1][~~-catalyzed~~] phosphorylation of the [wee1] protein occurs in its C-terminal region and leads to a substantial drop in its activity as a [cdc2][~~-specific~~] tyrosine kinase.

> We created a [spy0129] deletion mutant in strain SF370 ([~~SF370Δ~~][spy0129]) to determine if genes contained within the cluster were directly involved in adherence to pharyngeal cells.

Linking to UniProtKB

Each GGP annotation should be linked to an entry in the UniProt knowledgebase. Based on the entity name and the context surrounding it in the document, the annotator should be able to identify the corresponding UniProt entry.

Annotation of metabolic process triggers

Annotation Scope

Each metabolic process is expressed in text by a trigger, i.e., a span of text that explicitly signifies the biochemical alteration. Of particular interest for this annotation task are the types of metabolic processes specified in the CTD Interaction Types ontology. Examples of triggers are verbs (e.g., acetylates), verb nominalisations (e.g., hydroxylation) and adjectives (e.g., phosphorylated).

✔Include:

1. Any form of expression (e.g., past, present, future, progressive, nominalised), that explicitly signifies the biochemical alteration of a molecule's structure, even if only potential or hypothetical.

Structural characterization of eicosanoids [synthesized] in these preparations revealed an abundance of 15-lipoxygenase metabolites including 15-HETE when arachidonate was used as substrate and 5(S),15(S)-dihydroxy-6,8,11,13(E,E,Z,Z)-eicosatetraenoic acid when 5(S)-HPETE was used as substrate.
Purified recombinant enzymes UCK2 and dCK, but not UCK1, could [phosphorylate] 2'-MeC in vitro.

✖Exclude:

1. Expressions of static or fixed relations which do not explicitly signify any change or alteration

We further found that phosphatidylinositol 4-phosphate, which [~~contains a monoester phosphate~~] attached to its myo-inositol headgroup, also supported activity of factor VIIa.

2. Expressions which signify alteration but only implicitly, where the annotator would have to infer the (possible) existence of the change.

Chemists can [~~obtain~~] taxol from fungi in the bark of the Pacific yew tree.

Annotation Span

✔Include:

1. Characters (corresponding to chemical prefixes) appearing in the same token as the trigger, only if they provide more information on the kind of metabolic process being described.

CYP3A7, expressed in the human fetal liver and normally silenced after birth, plays a major role in the [16alpha-hydroxylation] of dehydroepiandrosterone (DHEA), DHEA sulfate (DHEAS), and estrone.
The organophosphate-induced alterations in K(d) represented changes in binding affinity of thioflavin t, with [diethylphosphorylation] of Ser203 increasing K(d), and [dimethylphosphorylation] of Ser203 decreasing K(d).

✖Exclude:

1. Auxiliary verbs.

Recombinant expression of rat Nats in Escherichia coli showed that MDA [~~was~~] [acetylated] by both recombinant rat Nat1 and Nat2.

IRP2 [~~is~~] [ubiquitinated] and [degraded] by the proteasome in iron-replete cells but is relatively stable in iron-depleted cells.

2. Prepositions connecting the trigger to metabolic process participants.

Most substrates tested were [converted] [~~to~~] detoxification products, except in the case of benzo[a]pyrene-7,8-dihydrodiol, which was [converted] [~~into~~] the very potent carcinogenic metabolite 7,8-dihydrodiol-trans-9,10-epoxide.

Similarly, estradiol 3-methyl ether is [demethylated] [~~by~~] CYP2C19 into estradiol, a CYP3A4 substrate for [ortho-hydroxylation].

3. Cues of negation, speculation or manner in which the metabolic process occurs.

[Dephosphorylated] Rb2/p130 exhibits decreased [ubiquitination] and thus is [~~not~~] [degraded] by the proteasome.

Purified recombinant enzymes UCK2 and dCK, but [~~not~~] UCK1, [~~could~~] [phosphorylate] 2′-MeC in vitro.

MAO A [~~preferentially~~] [oxidizes] serotonin (5-hydroxytryptamine, 5-HT) and norepinephrine (NE), whereas MAO B [~~preferentially~~] [oxidizes] phenylethylamine (PEA).

4. Characters appearing in the same token as the trigger which correspond to the names of participants involved in the metabolic process.

Previous studies in our laboratory showed that among cDNA-expressed human cytochrome P450 (P450) supersomes, CYP2C19 was the most active in [~~methoxychlor-~~][O-demethylation].

[~~CBDP-~~][phosphorylated] cholinesterases are nonreactivatable due to ultra fast aging.

5. Characters appearing in the same token as the trigger but are not part thereof nor giving more information on the type of metabolic process occurring.

These data support a critical role for Runx2 [acetylation][~~/~~][deacetylation] during osteogenic differentiation in MSCs in vitro.

Linking to CTD Interaction Types

Each trigger annotation should be linked to an entry in the CTD Interaction Types ontology. The entries in the said ontology are arranged in a hierarchy. The most specific (i.e., lowest in the hierarchy) metabolic process type that best matches the trigger should be chosen by the annotator.