Guidelines for event annotation

Tomoko Ohta, Jin-Dong Kim and Jun'ichi Tsujii

University of Tokyo

※ Last revision was made on 20 December, 2007

Contents

1. Terms and definitions 2
2. Scope of annotation 2
   2.1. General Guidelines 2
   2.2. What should be annotated 3
   2.3. What should not be annotated – General 3
   2.4. What should not be annotated - Inferences 5
3. A sentence and event annotation 7
   3.1. A sentence shown to annotators 7
   3.2. Event frames attached to a sentence 7
4. Elements of event annotation 8
5. Detail of each element of event description 9
6. Event annotation - Patterns and their meaning 26
7. Adding, editing and deleting of the name tag in the sentence. 30
8. GENIA ontology 32
   8.1. Term ontology 32
   8.2. Event ontology 36

1. Terms and definitions

(1) Bio-event

A change of the biological state, properties or location of a bio-molecule.

(2) Bio-molecule

Molecules that naturally occurs in living organisms, e.g. proteins, DNAs, RNAs, cells, etc., or their equivalents which are prepared for experimental purposes, e.g. cultured cells, specially treated proteins, etc.

2. Scope of annotation

2.1. General Guidelines

Event annotation in the current corpus follows the following general principles.

(1) Events to be annotated should alter properties or locations of biological entities.

Biological entities that participate in events to be annotated belong to the following classes defined in the GENIA term ontology (Section 8.1). Especially we focus on events which occur upon proteins, DNAs, RNAs and cells.

If you identify an event in text but nonetheless cannot find appropriate classes of the entities which participate in the event, you should report to the coordinator. The event that you identify may be out of the scope of the current annotation. For example, events related with diseases, symptom of diseases, effects of drugs, etc. are excluded from the current annotation.

It is also important that the biological entities the properties or locations of which are altered by the event should be explicitly described in text and indicated as such in your annotation.

(2) Events to be annotated are classified into 36 classes shown in Section 8.2.

The definitions of these classes are the same as those given in GO (http://www.geneontology.org/), except for the following 5 classes: Gene_expression, Artificial_process, Corellation, Regulation, and Positive_regulation, Negative_regulation. The definitions of these classes are given in Appendix I. Events which do not belong to these classes should not be annotated, and if you encountered such events, you should report to the coordinator for advice.

(3) A key expression should be identified.

A key expression, by which you judge the class of an event, should appear in text, and annotated as such. If you cannot identify such a key expression in text but nonetheless you think that an event of the class is described, you should report to the coordinator for advice. We enlist some of the examples in Section 2.3 and 2.4 which some annotators identified events in text but we decided not to be included.

2.2. What should be annotated

The scope of event annotation covers the following three levels of description of biological events. Other than individual events which are mostly mapped to upper level classes of Gene Ontology, regulatory events and dynamic relations between events or bio-molecules are annotated as Regulation (or its sub-classes) type of events.

(1) Individual event

A change of properties or states of a bio-molecule or bio-molecules.

Examples:

- Protein expression (IL2 gene expression)
- Formation of the complex (CD229-Grb2 complex formation)
- Virus infection (HIV infection)

(2) Regulatory event

A change of the frequency, rate or extent of an individual event.

Examples:

- Initiation of transcription (initiation of IL2 transcription)
- Promotion or inhibition of cell proliferation (T cell proliferation is initiated)

(3) Dynamic relation between events or bio-molecules

A causal relationship between events. Usually, this type of relationship is expressed as regulatory events which are described together with causal elements

Examples:

Promotion or inhibition of an individual event (expression of LMP1 induced COX-2 expression)Enzymatic reaction (PT catalyzed ADP-ribosylation)

2.3. What should not be annotated – General

There are descriptions in text, which look like events but describe rather relationships. These are not to be annotated in the current event annotation.

(1) Static relation between or bio-molecules

The following cases are NOT dynamic relations and NOT to be annotated by event annotation though they may encode biologically important information.

(2) Part-Whole Relationship and Structural relationships

Examples:

- connexin has four transmembrane domains
- Bcl-3 contains four ankyrin repeats
- B2 complex contains p50 and p55.
- B1 complex is composed of p50 and p55.

- Membership (is-a relation)

Examples:

- mafA, a member of the large-maf family
- NF kappa B is a transcription factor
-TCEd act as an inducuble proto-enhancer elemen

- Similarity of homologous

Examples:

- The structural similarity of SNI1 to Armadillo repeat protein…
- …the sequence requirements for PMA induction are distinct from those that mediate induction by virus or LPS.

(3) Miscellaneous

- Expressions which indicate the state after an event

- The prevalent complex is the heterodimer p50-p65.

* The word heterodimer may indicate an event of binding. However, since there are no key expressions in text to indicate the event. Only the state or existence of a protein (heterodimer) which suggests an event that results in existence of a protein with the stated property is not taken as an expression of event in this annotation.

- The term(s) that comparer(s) the activities etc.

- NOS was increased more markedly in the LY group than that in the OG

- The term(s) that indicate(s) the degrees.

- Additionally, significantly, poorly, highly, rarely, minmal, low, medium, high, X%, only slightly, etc.

* Only “exist” (positive) or “non-exist" (negative) should be specified, regardless the degrees.

2.4. What should not be annotated - Inferences

There are many cases where you may be able to infer events by using your biological knowledge. However, if an event lacks a key expression which indicates the event type in text, you should not annotate it as an event. The following are typical examples which should not be annotated, though you may be able to infer an event from text.

(1) When terms referencing bio-molecules indicate potential events

Ex. 1	*p50 is translated as a precursor of 105 kDa.*
Event annotation	Translation of a precursor of 105 kDa (○) Protein_processing of 105 kDa to produce p50 (×)

In the example sentence above, the translation event can be read directly from the textual key “translation”. Biologists also can infer that the Protein_processing event might happen from the term “precursor”. However, since there is no textual evidence which directly denote the Protein_processing event, it should not be annotated.

Ex. 2	*The prevalent complex is the heterodimer p50-p65*
Event annotation	Binding of p50 and p65 (×)

In the above example, the term “heterodimer” indicates a biological state of the two proteins p50 and p65 that interact each other. Biologists would infer that the state can be resulted by a binding event. However, since there is no textual evidence which directly denote the Binding event, it should not be annotated.

(2) Influence of functional change(s) of entity(-ies)

Ex. 3	*Inhibition of PI3K inhibits phosphorylation of Rb*
Event annotation	Negative_reg. of PI3K (○) Phosphorylation of Rb (○) Negative_reg. of (Phosphorylation of Rb) by (Negative_reg. of PI3K) (○)

	Positive_reg. of (Phosphorylatin of Rb) by PI3K (×)

In the above example, the last event may be inferenced by using a simple inference rule. However, you should not annotate such events which are not directly encoded in the given text.

(3) Linguistic expressions indicating similarity of bio-molecule's behavior.

Sometimes, a bio-molecule's property is implied in a linguistic expression indicating its similarity to another bio-molecule.

Ex. 4	*As observed with I kappa B alpha, nuclear RelA stimulates p100 mRNA and protein expression.*
Event annotation	Transcription of p100 (○) Expression of p100 (○) Positive_regulation of (Transcription of p100) by RelA (○) Positive_regulation of (Gene_expression of p100) by RelA (○) Positive_regulation of (Transcription of p100) by IkBa (×) Positive_regulation of (Gene_expression of p100) by IkBa (×)

In the above example sentence, the last two candidates (marked ×) of event annotation are infered by the text expression, "As observed with Ikappa B alpha"

(4) Prerequisite events

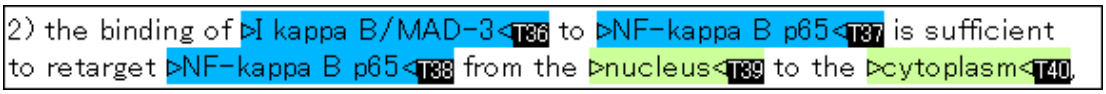
Ex. 5	*Nuclear expression of NF-kappa B occurs after its induced dissociation from its cytoplasmic inhibitor I kappa B alpha.*
Event annotation	Gene_expression of NF-kappa B (○) Dissociation of NF-kappa B and I kappa B alpha (○) Positive_reg. of (Gene_expression of NF-kappa B) by (Disassociation of NF-kappa B and I kappa B alpha) (○) Binding of NF-kappa B and I kappa B alpha (×)

From the dissociation event of NK-kappa B from I kappa B alpha in the above example, annotators may infer that there should be binding event of the two proteins prior to this event. However, since there are no explicit mentions of the binding event, you should not annotate it for this sentence.

3. A sentence and event annotation

Event annotation is made sentence by sentence. Annotators are supposed to read every sentence in a given text. Whenever an annotator recognizes an event mention in a sentence, (s)he create an event frame right after the sentence and fill out the slots in the event frame.

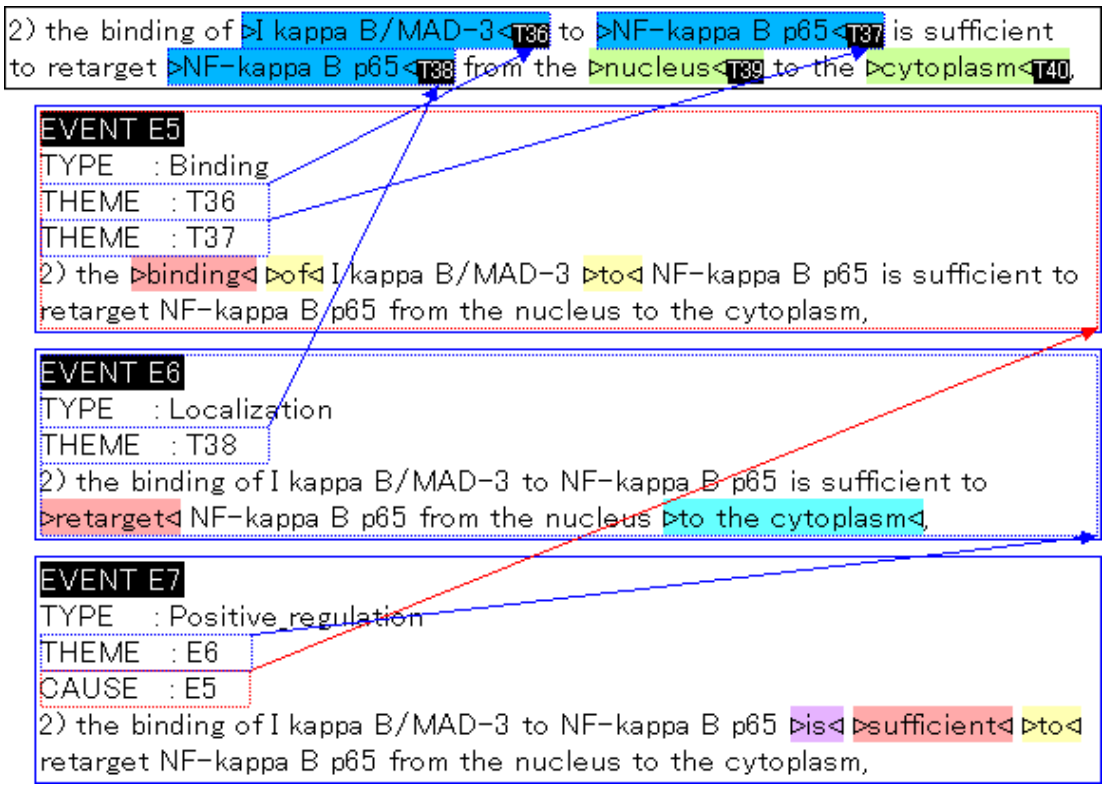
3.1. A sentence shown to annotators



As shown in the above example, each sentence to be shown to annotators has terms already annotated. Highlighted text spans in blue mean proteins and those in green mean cellular locations. Since each instance of term annotation has its own Id, annotators can refer to them when necessary for event annotation.

3.2. Event frames attached to a sentence

Whenever an annotator finds event mentions in a sentence, (s)he has to create event frames right after the sentence. Since each event frame is meant to describe one event, event frames need to be created as many as the number of event mentions found. The order of event frames does not matter.



4. Elements of event annotation

An event is described by an event frame which is composed of following elements.

(1) event

The *event* element defines an event frame. Every event element has to have its own Id which is unique over the document.

(2) type

The *type* element is meant to describe the type of an event. Its mandatory attribute class has to be assigned with an event class defined in the GENIA event ontology (see section 6.2)

(3) theme

A theme of an event means a bio-entity whose properties are changed by the event. Its mandatory attribute idref is supposed to point to the Id of an annotated terms or sometimes another event.

(4) cause

A *cause* of an event means a bio-entity which affect the way of occurrence of an event. Its mandatory attribute idref is supposed to point to the Id of an annotated terms or sometimes another event.

(5) clue

The *clue* element is the place where fragments of text are marked-up to indicate textual expressions which are responsible for the event mention. In order to do so, the corresponding sentence without term annotation is copied into the clue element. Annotators have to mark-up text spans which are supposed to give the following types of information of the event.

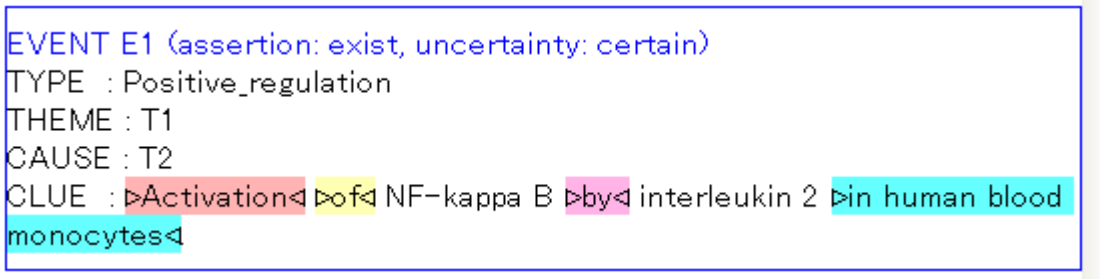
- clueType: a text span which gives clue for the event type.
- linkTheme: a text span which links the clueType with the theme of the event.
- linkCause: a text span which links the cluetype with the cause of the event
- corefTheme: a text span which is a coreferential expression of the theme reference.

- corefCause: a text span which is a coreferential expression of the cause reference.
- clueLoc: a text span which gives clue for the location of the event.
- clueTime: a text span which gives clue for the temporal information of the event.
- clueExperiment: a text span which gives clue for the context of the event in terms of experimental method.

(6) comment

The *comment* element, which is optional, is prepared for communication between annotators and coordinators. COMMENT is the free description for the EVENT, have some operation rules, which is shown in the 3.6.

< Example of the EVENT description>



>1st line, EVENT ID and 3 properties in the grouping symbols.
>2nd line, TYPE
>3rd line, THEME
>4th line, CAUSE
>5th line, CLUE.
>After the CLUE, COMMENT will be shown when needed.

5. Detail of each element of event description

5.1. Attributes of the *event* element

The event element may have the following attributes.

(1) assertion

At “assertion”, “exist” or “non_exist” is chosen. “exist” means that the event is positive, and “non_exist” means that the event is negative. “exist” is selected by default. “assertion” should be chosen by the context clue.

(2) id

The specific ID number in the sentence, whose initial is “E” is given at “id”.

(3) uncertainty

The level of certainty of the event description is chosen from 3 alternatives, certain, probable, and doubtful, depending on the degree. “certain” is selected by default.

<Example of the “property” of the EVENT>

```
EVENT E1 (assertion: exist, uncertainty: certain)
TYPE  : Positive_regulation
THEME : T1
CAUSE : T2
CLUE  : ⊳Activation⊲ ⊳of⊲ NF-kappa B ⊳by⊲ interleukin 2 ⊳in human blood
monocytes⊲
```

> On the 1st line, there are 3 properties; id is “E1”, assertion is “exit” and uncertainty is “certain”. “assertion” and “uncertainty” is shown in the grouping symbols.

5.2. type

- EVENT ontology has the arborize structure which was made based on the GO, Gene Ontology.
- In principal, the phenomenon, which is not shown in the provided event ontology, cannot be a target.
- If there is the term can be an EVENT but cannot be classified in the appropriate class in the EVENT ontology, it should be UNCLASIFIED and “SOO”, short of ontology, should be put at the comment section, cf. 3.6.

Examples: dissociate, oncogenesis, (tumor) invasion, metastasis, infection with bacteria, etc.

5.3. theme

We have defined a convention for assigning IDs to annotation instances.

Examples

Idref: T1… The name tag which is prepared in the sentence.

Idref: A10 … The name tag which is made newly as needed.

Idref: E12 … The EVENT ID which is already described.

Idref: S7 … The ID of the sentence, which should be specified as the THEME.

Following the Idref tag, Idref1, Idref2 and Idref3… are used for setting the scattering terms in the sentence as a series of the term.

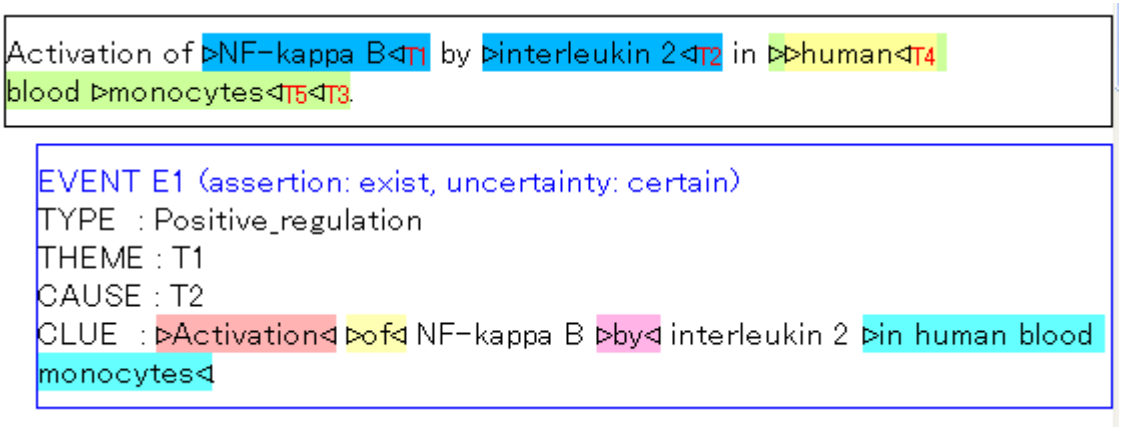
Example

In the sentence of "NFAT-1 or NF kappa B binding sites……", when "NFAT-1 binding site" should be chosen as the THEME, the name tag of "NFAT-1" should be set in the "Idref" and that of "binding sites" should be set in the "Idref1".

5.4. cause

- The setting of the cause is the same procedure of that of theme.

<Example of the setting of THEME and CAUSE>



> The colored parts, which are in the sentence in the inside of the top box, are the prepared name tags.

> The reddish marks shown on the bottom-right corner of the colored tags are the IDs of the tag.

> These IDs can be a used in the THEME and CAUSE to describe the EVENT in the next box.

<Notes>

- The specified range of THEME and CAUSE

1) The target of the EVENT has wide range, i.e. individual protein, the sequences of the protein or gene, protein family, cellular activity, phenomenon of the disease and the concept of mechanism or pathway can be the target. Since there is no priority between them, the focus of the context and shortest term should be chosen.

Examples

"NFkB" should be chosen from the terms of "transcription factor NFkB".
"HIV-enhancer" should be chosen from the terms of "HIV-enhancer activity".
"PMA" should be chosen from the terms of "ability of PMA".

2) In the description like xxx mRNA, xxx protein or xxx gene, the target should be set with mRNA, protein and gene, respectively.

Example

TG induced increases in IL-2R alpha protein as well as IL-2R alpha mRNA

3) The modifier can be included in the target term, if the term has no meaning without it.

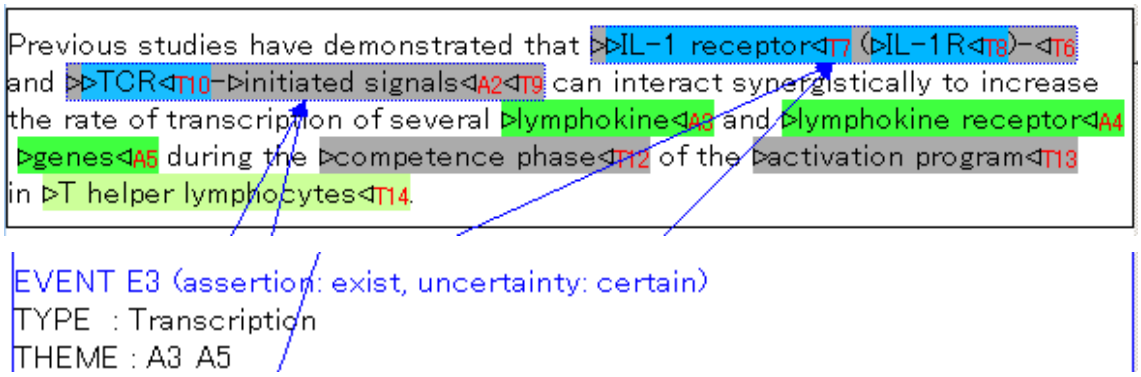
Example; peptide carrying the nuclear localization sequence of NF-kappa B

When the terms following "carrying" are omitted, there is no meaning. So, the whole terms from "carrying" to "NF-kappa B" should be chosen.

4) When the term is written as "full name (abbreviation)" or "abbreviation (full name)", the abbreviation should be chosen. For example, in the terms of "IL-1 receptor (IL-1R)", "IL-1R" should be chosen.

- In order to choose the scattering terms as a series of the term for the target, Idref1 and followings followed by Idref can be used.

1) < Example of the term-binding>



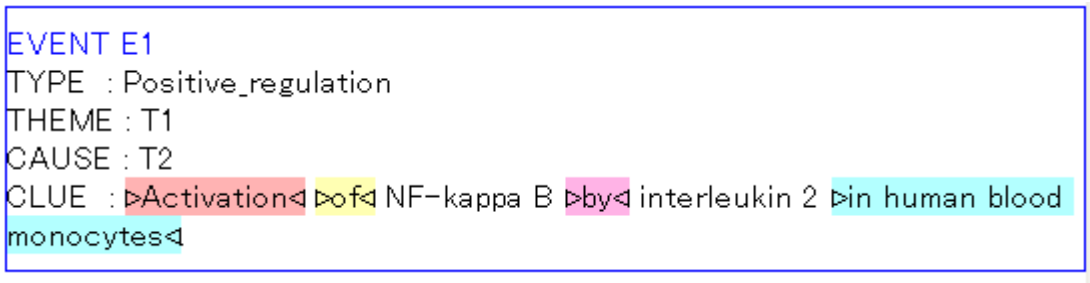
> Idref identifies with A3 and Idref1 identifies with A5 in the exampple. And they are chosen as the THEME. That means “lymphokine genes”.

1) If needed, the part of the term can be tagged as the target.
Example; “homo- and heterodimer”
“heterodimer” can be divided into “hetero” and “dimer”, and each section is tagged. “homo-“ & “dimmer” and/or “hetero” & “dimer” can be set.

5.5. clue

- The *clue* element is the place where the location information can be recorded. The information is classified and the tagged in the sentence. Tags show the object, cause etc. of the event in the sentence. clueType, clueExperiment, clueTime, clueLoc, linkCause, linkTheme, corefCause, corefTheme are the tags used in the clue element.

<Example of the *clue* element>



> Red ; clueType
> Cream ; clueTheme
> Purple ; linkCause
> Blue ; clueLoc

- The tags used in the clue element

- clueType

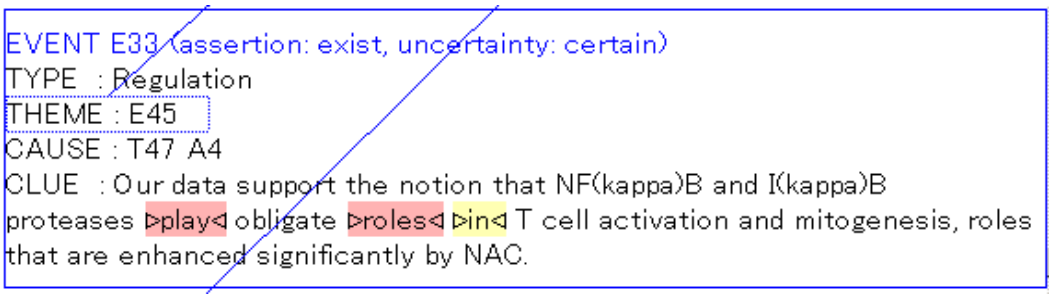
"clueType" shows the EVENT itself and it is essential. In the box of "Example of the CLUE", it is shown as reddish term. "clueType" should be the minimal terms or phrases. Multiple dispersed terms can be a set of the "clueType".

Examples

One word; transcriotion, actuvated, adhesion, proliferated etc.

Phrase; form complex, play role, have effect etc.

<Example of the "clueType">



> The reddish terms, play and roles are the "clueType".

- If the "clueType" means the negative, the negative expression should be included in the it.
  Examples; not activated, neither inhibited

- As a general rule, the "clueType" is the essential. Exceptionally, when the "clueType" can not be found, "No clueType" should be written in the COMMENT elemsnt. Cf. 3.6.

- clueExperiment

"clueExperiment" shows the experimental technique which examines the possibility of the EVENT. The name of the experiment, experimental conditions or other experimental procedures should be included in the "clueExperiment"

Examples; by ELISA, Northern blot analysis, in a gelshift assay, by gel-shift assay using an oligonucleotide probe corresponding to xxx, etc.

<Example of the “clueExperimen”>

EVENT E37 (assertion: exist, uncertainty: certain)
TYPE : Binding
THEME : T42
THEME : T45
CLUE : As shown ▷by a methylation interference analysis and oligonucleotide competition experiments◁, purified NF-kappa B ▷binds◁ ▷at◁ positions -82 to -91 (GGGAACTACC) of the GM-CSF promoter sequence with an affinity similar to that with which it binds to the biologically functional kappa B motif in the beta interferon promoter (GGGAAATTCC).

> The peppermint-green phrase is the “clueExperiment”.

- clueTime

“clueTime” shows when the EVENT will happen or happened. The time after the treatment, cell stage or phase of the diseases etc.

Examples; by 2 hours, within 24 h, during monocytic differentiation, during acute infection

<Example of the “clueTime”>

EVENT E6 (assertion: exist, uncertainty: certain)
TYPE : Cellular_process
THEME : T10
CLUE : Endothelial cells stimulated by LPS express E-selectin, which plays an important role in mediating neutrophil ▷adhesion◁ ▷during inflammation◁

> The bluish-violet phrase is the “clueTime”.

- clueLoc

“clueLoc” shows where the EVENT happen. What kind of animal, which organ, what cell or which part of the cell etc. can be the “clueLoc” and it should be colored in blue.

Example; in B lymphoid cells, in the cytoplasm, surface, nuclear etc.

However, in vitro, in vivo and in situ etc. do not belong to this category.

- linkCause

“linkCause” shows the word that links CAUSE to EVENT. It is purple word in the box of <Example of CLUE>. Frequently, “linkCause” is the preposition, by, through, with, in response to treatment with and only if etc.

Examples

Activation of NFkB by IL2

Mediated thorough NF-kapa B binding sites

It failed to respond to these mitogenic stimuli if both sequences were absent.

- linkTheme

"linkTheme" shows the word that links THEME to EVENT. It is cream word in the box of <Example of CAUSE>. Frequently, "linkTheme" is the preposition, of, in, to and on etc.

Examples

transcription of NFkB

increase in thiol

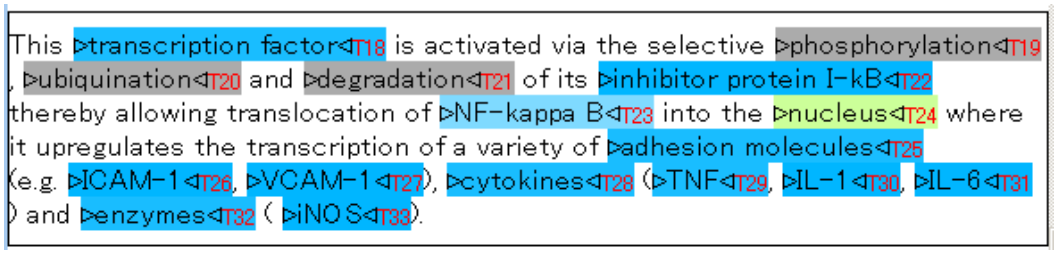
- corefCause

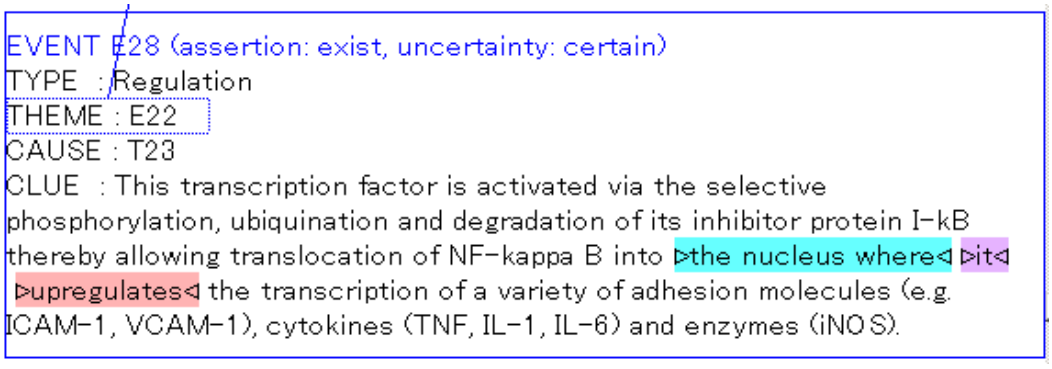
"corefCause" shows the origin of the coreference of the CAUSE.

Example

When "this sequence" is equal to "NFAT-binding", "this sequence" should be the "corefCause" and the tag of the "NFAT-binding sites" should be the CAUSE to show the relation between "this sequence" and "NFAT-binding".

<Example of the "corefCause>





> it = NF-kappa B, corefCause = it, CAUSE = T23 (the ID of "NF-kappa B")

- corefTheme

"corefTheme" shows the origin of the coreference of the THEME. "corefTheme" is set same as "corfCause".

Example

When "this activation" is equal to the EVENT described above, "this activation" should be the "corefTheme" and the tag of the previous EVENT should be the THEME to show the relation between "this activation" and the previous EVENT.

<Note for the individual tag>

- “clueType” is essential but other tags should be set when needed.
- These tags will not necessarily be one word or one phrase. A number of terms can be a set as each tag.

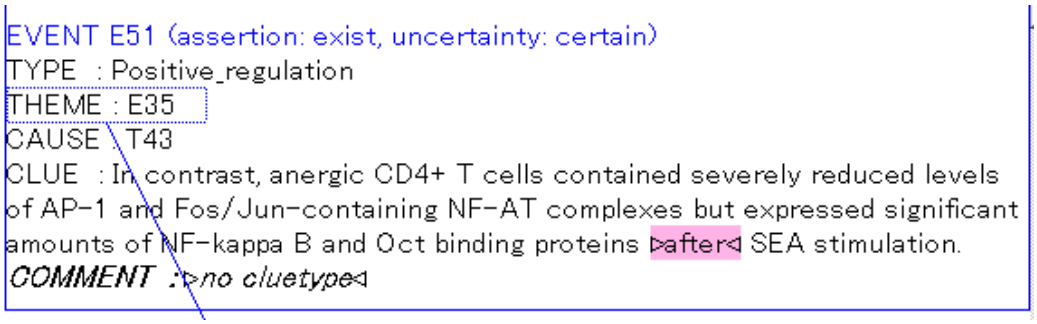
- About “clueType”

- Usually it is the verb or gerund, but the word that indicates the EVENT should be “clueType” depends on the contextual meaning.
  Examples
  The performer of the inhibitor, activator or messenger etc.
  The expression of the down stream of the regulation.
  In some cases, The expressions to show the conditions as sensitive or tolerant can be the “clueType”

- is, was, are, were and be, which are used to make a passive voice, cannot be included in the “clueType”

- Not clueType; As a general rule, “clueType” is necessary, but there is not the appropriate term can be an EVENT, “no clueType” should be put in the COMMENT element.

<Example of the “No clueType”>



- When the “clueType” in the other EVENT, clueType in Ex should be put in the COMMENT element.

- About “clueTime”

- Most of the expressions described by the “when” or “after” do not belong to the “clueTime” but “EVENT”, sometimes it belongs to “CAUSE”.
  Example
  human B lymphocytes undergo long-term proliferation when activated through CD40 RelB, accumulated after B cell receptor stimulation.

- About “clueLoc”

- “clueLoc” should be chosen by the meaning or context.
  Example; cells stably expressing a transdominant Rev protein
  “cells” is the subject, but it is not CAUSE or THEME. It is “clueLoc”, when the “gene expression” is the EVENT .
  Example; The monoclonal antibody (mAb) J393 induces apoptosis in Jurkat T-cells
  “ in Jurlat T-cells” is the adjunct, but it is not “clueLoc”. It is THEME, when the EVENT is the “Cellular_process > apoptosis”.

- The words that show the locations for THEME or CAUSE cannot be the location for the EVENT, that is they are out of target.
  Example; binding to three regulatory sequences in the IL-2 promoter region
  “in the IL-s promoter region” is the information of the location, but it is the place of the three regulatory sequences, and it is not the “clueLoc”.

- Some locations can be the “clueLoc” at the same time. But when the places can be the locations for different EVENTs respectively, the EVENT and the “clueLoc” should be separated as exception.
  Example; in monocytes and macrophages
  E: clueLoc = in monocytes and macrophages
  Exception; in monocytes and macrophages respectively
  E1: clueLoc = in nmonocytes
  E2:clueLoc = in macrophages

- About “linkCause”

- "mediate" is often included a "clueType" of the EVENT but sometimes can be a "linkCause".
  Example; monocyte-mediated killing

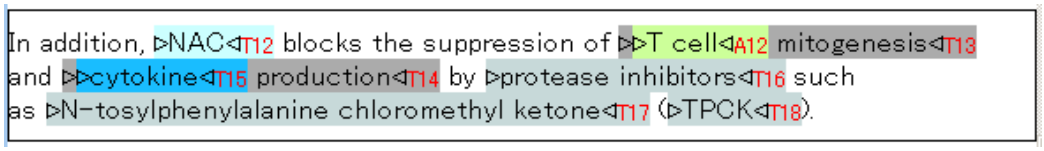
- About "corefCause" and "corefTheme"

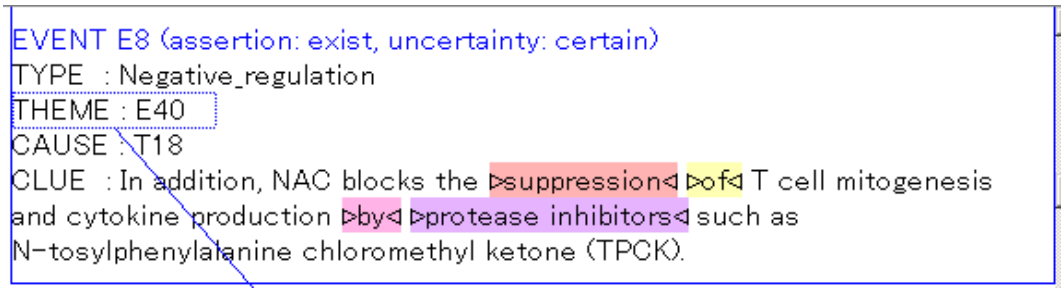
- In principal, "corefCause" and "corefTheme" should be the word, which is equal to the pronoun or pronoun phrase as one-to-one relation.
  Example; this transcription factor = NF kappa B
    corefTheme = this transcription factor
    THEME = NF kappa B

- When the multiple words are included in pronoun phrase, which is using "both" or "these", for example "both these lymphokine genes", and these words are equal to the multiple objects, the EVENT should be made as same number as the references.
  Example; both these lymphokine genes = IL2 and IL4
    E1; corefTheme = both these lymphokine genes, THEME = IL2
    E2; corefTheme = both these lymphokine genes, THEME = IL4

- If the references indicate a part of the targets using "including", "such as" or namely", those should be "corefCause" or "corefTheme".
  Example; Several other transcliption factors including NF-IL-6, AP-1
    E1; corefTheme = several other transcription factors, THEME = NF-IL-6
    E2; corefTheme = several other transcription factors, THEME = AP-1
  Example; cytokines such as TNF-alpha
    E; corefTheme = cytokines, THEME = TNF-alpha

  <Example of the "corefCause" corresponding to multiple objects>





> In the phrase of “protease inhibitor, such as TPCK”, “protease inhibitor” can be a “corefCause” and CAUSE should be “T18”tagged “TPCK”.

- The term or phrase, which is tagged by “corefCause” or “corefTheme”, don’t have to have the name tag.

- These anaphoric relations should be set with only CAUSE and THEME. Even if the “clueLoc” or “clueType” is showed by the pronoun (phrase), they should not be related the word or phrase.

  Example

  in these cell = monocytes and macrophage, clueLoc = in these cells

- In these anaphoric relations of CAUSE and THEME, the EVENT or sentence can be the object.

  Example; this stimulation may occur through increased kappa B binding activity

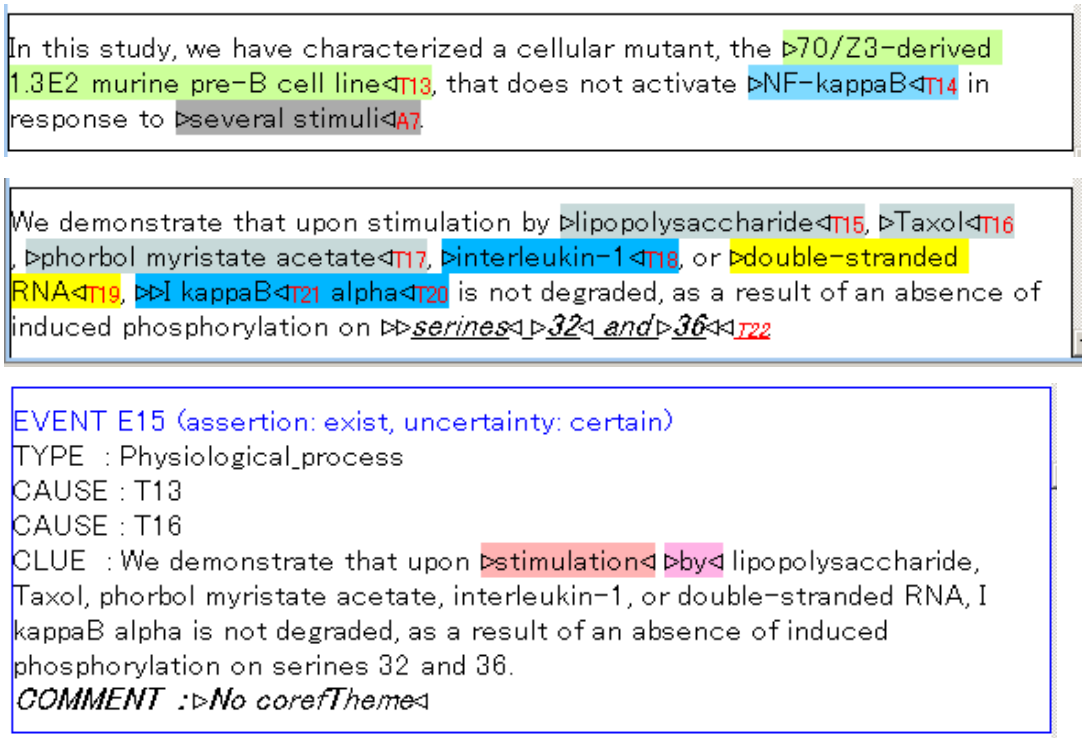
  When “corefTheme” of the reddish phrase in the above sentence shows the previous EVENT, its id should be chosen as the THEME, that is “corefTheme” = this stimulation and THEME = E16.

  Example; NAC not only blocks the effect of TPCK

  When the “corefTheme” of the reddish phrase in the above sentence shows the some matters in the previous sentences, the sentence number can be chosen as the THEME, that is “corefTheme” = the effect of TPCK and THEME = S6. This means that the all EVENTs in the sentence are chosen as the THEME.

- When there is no part of “coref” but there are apparently cause or theme of the EVENT in the sentence, “No corefTheme” or “No corefCause” should be put in the COMMENT element, cf 3.6, and the THEME or CAUSE in another sentence should be chosen.

  <Example of “No corefTheme”>



> In the box of EVENT E15, T13 is shown as the THEME, but the T13 is not the same sentence of which E15 is described. Like this situation, if apparently, the word or phrase can be the THEME of E15, but in the box of EVENT E15, “No corefTheme” should be put in the COMMENT element.

5.6. comment

In the current operational rule, there are ten COMMENTs are provided. They should be written with capital letters.

- FAQ （ FAQ = Frequently asked question）
- NER: Tn MOD （NER = Named Entity Recognition, MOD = modified）
- NER: Tn DEL （NER = Named Entity Recognition, DEL = deleted）
- TPS: An （TPS = Term with Phrase Structure）
- SOO (SOO = Shortage Of Ontology)
- NO CLUETYPE
- CLUETYPE IN En
- NO COREFTHEME
- NO COREFCAUSE
- CAUTION

● FAQ

Free note for frequently-asked question etc.

Ex. comment: FAQ: type = transcription or gene_expression ?

● NER: Tn MOD

When the name tag is edited, the tag ID should be shown as Tn at the COMMENT element.

Ex. comment: NER T3 MOD

● NER: Tn DEL

When the name tag is deleted, the tag ID should be shown as Tn at the COMMENT element.

Ex. comment: NER T14 DEL

● TPS: An

When the phrase which includes the preposition is edited, TPS: An should be shown at the COMMENT element. . Cf. 4

Ex. comment: TPS: A10

● SOO

It means there is no appropriate class for the EVENT, cf.3.2.

Ex. comment: SOO

● NO CLUETYPE

It means there is no appropriate clueType, cf.3.4.

Ex. comment: NO CLUETYPE

● CLUETYPE IN En

It means there is no appropriate clueType but it is in another EVENT, the EVENT id should be shown in the COMMENT element. Cf.3.4.

Ex. comment: CLUETYPE IN E4

● NO COREFTHEME

It means there is no appropriate corefTheme, cf.3.4.

Ex. comment: NO COREFTHEME

● NO COREFCAUSE

It means there is no appropriate corefCause.

Ex. comment: NO COREFCAUSE

● CAUTION

Other than those above and the review will be needed. Free comment and it can be written in Japanese.

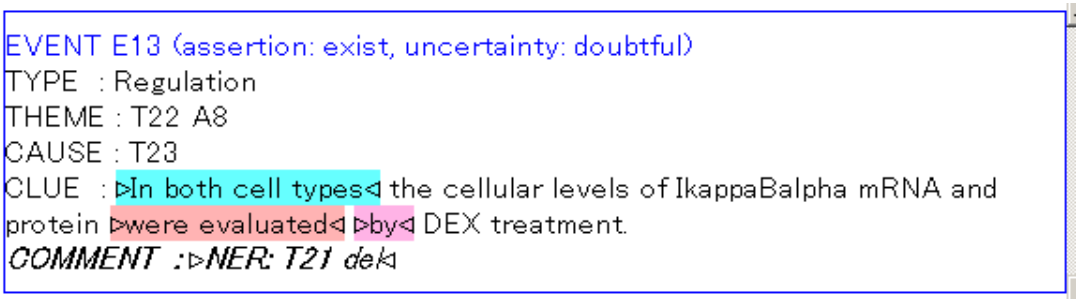
Ex. comment: CAUTION: migrate = binding として処理

- COMMENT is for the EVENT not for the sentence, so NER should be shown at the EVENT that the NER is related to.

- One EVENT can have one COMMENT element. If the EVENT has some comments, they should be separated by comma and listed in the same COMMENT element.

Ex. comment: NER T4 mod, TPS

<Example of COMMENT>



>The last line, which was written in italics is the COMMENT element.

6. Event annotation - Patterns and their meaning

● CAUSE (x1) – THEME (x1)

Regulation (porsitive_regulation, negative_regulation etc.) is prescribed by one CAUSE and one THEME.

Example: activation of NFkB by IL-2

THEME: NFkB

CAUSE: IL-2

● THEME (x1) – THEME (x1)

In the EVENT of binding or correlation, it is impossible to describe which term(s) is CAUSE, so both of them should be THEME.

Example: binding activity of NFkB to its consensus sequence

THEME: NFkB

THEME: its consensus sequence

>Additionally, “cell to cell adhesion”, “chemical treatment of the cell”, virus infection of the cell” etc. should be described with same rule of above, THEME – THEME.

● CAUSE (x2) – THEME (x1)

When two entities are necessary for one EVENT, both entities can be CAUSE, respectively. This is the “AND” type.

Example: Both protein A and B activate C.
THEME: C
CAUSE: A
CAUSE: B

>“synergistically” is also included in “AND” type.

● (CAUSE (x1) – THEME (x1)) x2

When one protein activate more than one protein, each activated protein should be described as individual EVENT, respectively. This is the “OR” type.

Example: Protein A activates protein B and C.
E1: THEME: B、CAUSE: A
E2: THEME: C、CAUSE: A

● CAUSE (x1) – THEME (x2)

When the sentence has the sentence-pattern of “subject, verb, object and object”, and THEMEs are separated before and after the EVENT, two THEMEs should be created and first THEME should be treated as the precursor, before EVENT, and second THEME should be treated as the mature, after EVENT.
Example: monocytes differentiated into macrophage
CAUSE: -
THEME: monocytes
THEME: macrophage

● CAUSE (x0) – THEME (x1) , No-CAUSE EVENT

Example: activation of xxx、regulation of yyy, zzz synthesis
CAUSE: -
THEME: xxx

● CAUSE (x1) – THEME (x0), No-THEME EVENT

Example: xxx-mediated inhibition、the inhibitory effect of yyy
CAUSE: xxx
THEME: -

● CAUSE (x0) – THEME (x0)

The EVEND ,which has either THEME nor CAUSE, is not EVENT. These EVENTs should be treated with “name”, sem = other_name.

Example: “inhibition” in the sentence of “xxx-mediated inhibition”.

● When the EVENTs are in the sentence which has the nesting structure

Other EVENTs can be selected as THEME or CAUSE. When the EVENT TYPE is different, each type needs EVENT. When the EVENT TYPEs are same but are in different stages, it should be shown in the EVENT nesting structure.

Example: LTB4 induces synthesis of IL-6

E1: synthesis of IL-6

E2: LTB4 induces E1

＜Other notes＞

- In the abstract, if the sentence(s) is (/are) not the main or focal points of the article but the quoted or background, it (/they) should be the target(s) of EVENT.

  Example:

  “The proteolytic degradation of the post-translationally modified I-kappa B is known to be mediated by the 26S proteasome complex.”

  This shows the background of the abstract, the EVENT should be searched in this sentence.

- In complete sentence as title or the terms which can be treated as EVENT partially can be EVENT, if possible.

  Example:

  Frequently, the title is not completed as “Oxidoreductive regulation of nuclear factor kappa B.” But the “regulation” can be an EVENT in this example.

Example: NF-kappaB activation in Jurkat are probably distinct from the CD28 costimulation pathway.

In this sentence, “distinct” is the main word but it cannot be EVENT. However, “NF-kappaB activation in Jurkat” and “CD28 costimulation pathway” can be independent EVENTs.

- The EVENT between highly abstracted terms should be EVENTs.

Example: “transcription” in the phrase of “NF-kappa B activated transcription through the P sequence”

Example: “signal transducing components “ in the phrase of “signal transducing components in T cells can activate transcription of the GM-CSF gene”.

Example: “DNA” in the phrase of “NF-kappa B DNA binding activity”.

Example: “the signal transduction pathways” in the phrase of “the signal transduction pathways which may be affected in T cells by constitutive expression of the nef gene”.

Although, above examples are highly abstracted, they can be EVENTs. Therefore, the Event should be applied the term and described as EVENT.

- The EVENT doesn’t have to be picked up from the expression mentioned below as examples.

Examples

xxx/yyy complex	(binding)
xxx (+) cells	(gene_expression)
In the CD40-negative variant	(gene expression non_exist)
yyy agonist	(positive_regulation)
yyy antagonist	(negative_regulation)

- EVENT which is included in the term.

Example: NF-kappa B dependent cis-acting sequence

The term, which is modified by the adjective phrase, is tagged as the series. However , “dependent” in this series of terms should be the EVENT by itself.

- EVENT which is in the clueLoc

Examples:

"in cells constitutively expressing the HTLV-I Tax protein".

"The anti-CD3- and anti-CD28-treated Rel-/- Tcells".

These examples are "clueLoc"s in the main EVENT. However, in this phrase, reddich terms can be EVENTs as "expression" in the first example and "artificial_process" in the second eample.

7. Adding, editing and deleting of the name tag in the sentence.

As a general rule, name tags have already given in the sentences. If the name tag is not given appropriately to describe the EVENT, the tags can be added, edited or deleted according to needs.

- Adding the name tag

The properties of the name are "Id", "lex" and "sem", but only "Id" and "sem" have to be set.

- Id: Newly "Id" should start from "A" and be specific in the abstract. But it has not be sequential number.

Examples: A12, A100

- sem: "sem" specifies DNA, RNA, protein, other compounds, cell or other phenomenon etc. from the ontology file same as "EVENT class".

Examples: Protein_complex, Protein_family_or_group, DNA_domain_or_region etc.

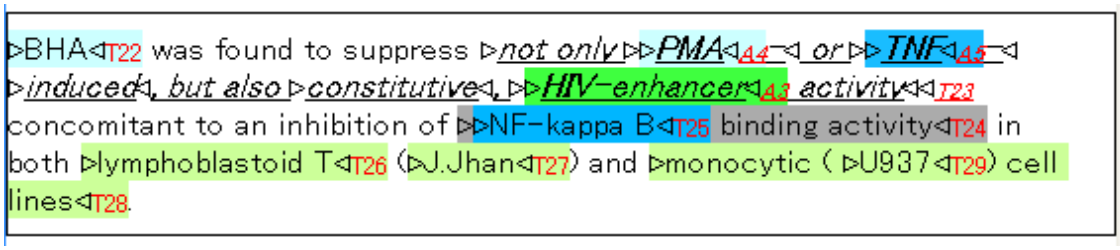
- When the provided name tag or the "sem" of it is changed, it should be noted in the COMMENT element according to the rule, NER: Tn mod.

- When the provided name tag or "sem" of it deleted, it should be noted in the COMMENT element according to the rule, NER: Tn del.

● “cons” and “ f rag” in the “coordination” don’t have to be added or edited.

About the details regarding the name tag, refer to the “GENIA Corpus Encoding Overview - Draft: 5 November 2005 - 2.Term annota t ion.”

<Example of adding the name tag in the sentence>



> A3, A4 and A5 in the above box are added tags.

> The italics means the part of which tagged “cons” and “frag” in the “coordination”.

> The color of tag is decided according to the class as follows.

“Other compounds” is colored aqua-blue.

“Protein” is colored blue.

“DNA” is colored green.

“Phenomenon” is colored gray.

“Cell” is colored light-green.

8. GENIA ontology

8.1. Term ontology

<Examples of the classes of term ontology>

- ○ Other_name: xxx binding activity, Redox status, oxidative stress, etc.
- ○ Cell_line: U937 cells, Hela cells, NIH 3T3 cells, etc.
- ○ Body_part: skin, lung, immune system, etc.
- ○ Cell_component: nucleus (nuclear, nuclei), cytoplasm, cell membrane, etc.
- ○ Cell_type: monocytes, macrophages, T cells, etc.
- ○ Multi_cell: mice, human, patients, HIV-infected individuals, etc.
- ○ Virus: HIV1, virus particles, human CMV, etc.
- ○ Tissue: endothelium, serum, fibrosarcoma, etc.
- ○ Atom: Ca(2+), calcium, Pb, iron, etc.
- ○ Inorganic: reactive oxygen, nitric oxide, ROIs, NOS, H2O2, etc.
- ○ Amino_acid_monomer: tyrosine, serine residues, N-acetyl-cysteine, etc.
- ○ Peptide: glutathione, glutashione disulfide, etc.
- ○ Protein_complex: proteasome, Ca(2+)-ATPase, NFkB, C-Rel complexes, etc.
- ○ Protein_domain_or_region: C-terminal region, N-terminal domain, catalytic domain, etc.
- ○ Protein_family_or_group: TNF, cytokine, Tat mutants, anti-CD3 antibodies, etc.
- ○ Protein_molecule: CD30, IFN gamma, I kappa B alpha, etc.
- ○ Protein_subunit: p65 RelA, p50 subunit, IL2 receptor alpha-chain, etc.
- ○ Carbohydrate: sugar moiety, O-linked monosaccharide glyncans, etc.
- ○ Lipid: LPS, phosphatidyl inosytol, LDL, high density lipoprotein, etc.
- ○ DNA_domain_or_region: IL-2 gene, IL-2 promoter region, enhancer, NfkB site, etc.
- ○ DNA_family_or_group: IFN genes, cellular genes , etc.
- ○ RNA_molecule: IL-2 mRNA, IFN gamma mRNA, etc.
- ○ RNA_family_or_group: TNF mRNA, virus RNA, etc.
- ○ Other_organic_compound: PMA, FK506, okadaic acid, Cholera toxin, etc.

<Note of the setting of term ontology>

- The class should be set lower class of term ontology.
- “xxx gene” means a part of DNA and it should be “DNA_domain_or_region”.
- “xxx mRNA” and “xxx protein” are independent molecules and they should be “mRNA_molecule” and “pritein_molecule”, respectively.
- “protein_ETC”, “DNA_ETC” and “RNA_ETC” have not be used. Almost all can be categorized as other classes.
- When the category is not clear, it should be searched by the “Xconq search” or “MeSH”.

<Other note about the name tag>

- As a general rule, highly abstracted term does not have the name tag. When it is set in the EVENT, it should be given the new tag.

  Examples
  proteins, DNA
  corresponding sequence
  other DNA-binding proteins
  additional protein
  different proximal mechanisms
  one base pair change, etc.

- As a general rule, the phrase, which is included the preposition, is not be recognized as the phrase to set the name tag. When the it is set in the EVENT, new tag should be set and “TPS” should be described in the COMMENT element.

  Examples
  analog of DHEA,
  antisense oligonucleotides to p65
  peptide carrying the nuclear localization sequence of NF-kappa B p50, etc.

Followings are exceptions, “TPS” has not be described in the COMMENT element.

>NFAT = nuclear factor of activated T cells

>STAT = Signal Transducers and Activators of Transcription

The name, which include the preposition in it, should be set the name tag.

- When the word selected as CAUSE or THEME, it can be given the name tag even if it is adjective.

  Example: mitogenic activation

- When the "coordination" should be set, each part should be given the term tag and "sem"s should be unified.

  Example: NFAT-1 and NF kappa B binding sites

  The "sem" of NFAT-1 should be "DNA_domain_or_tegion"

  The "sem" of the binding site should be "DNA_domain_or_tegion"

- When the "sem"s cannot be unified, make exceptions to the rule.

  Example: IL2 mRNA and protein

  The "sem" of IL2 should be "RNA_molecule"

  The "sem" of mRNA should be "RNA_molecule"

  The "sem" of "protein" is "protein_molecule".

  In this example, the "sem"s cannot be unified as "RNA_molecule".

- Generally, it is not necessary to delete the name tag, but there are several exceptions.

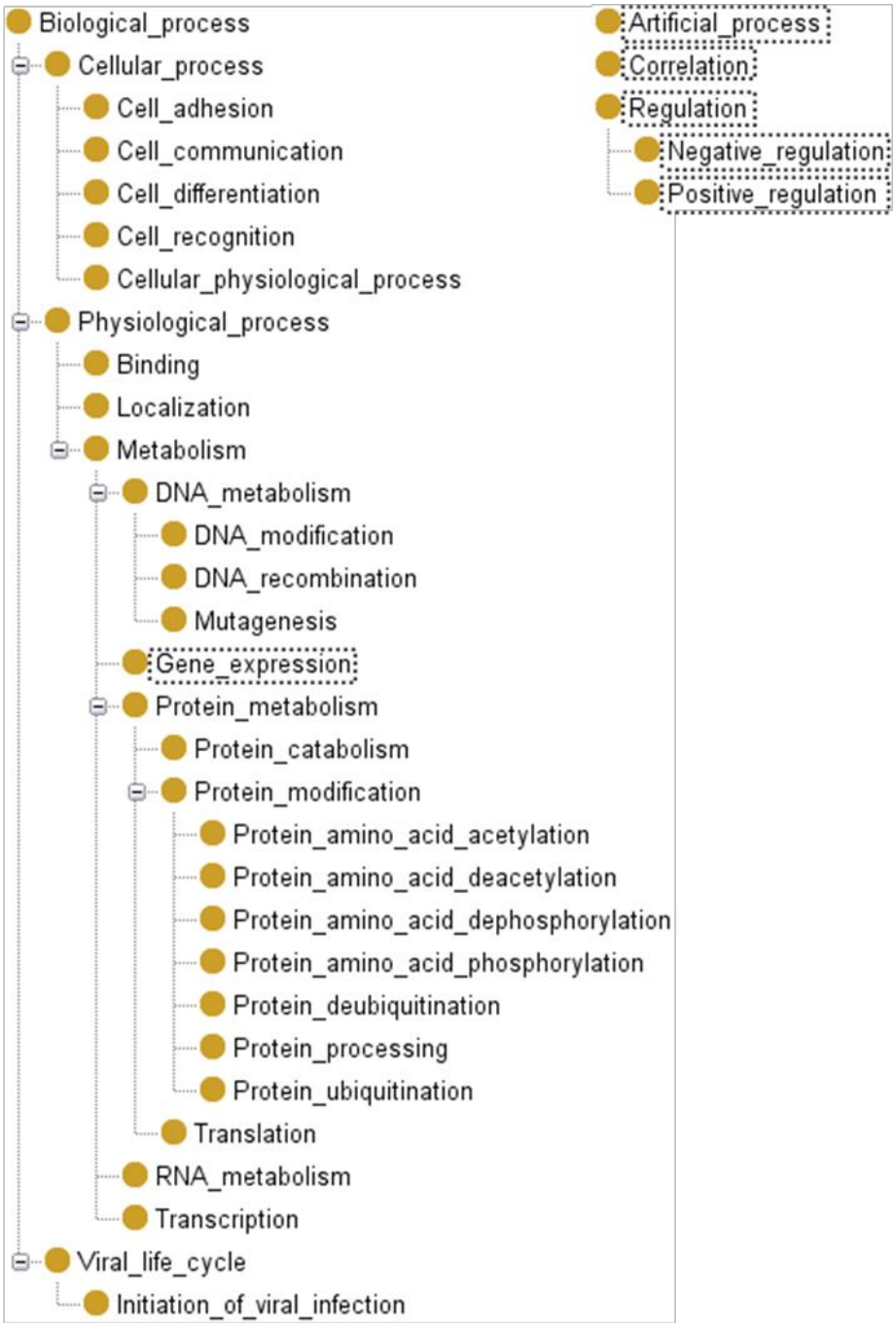
- In the part of "coordination", "cons" and "frag" are shown instead of term, name tag. They cannot be insert in the term and set as the term. When these should be set as the term, "cons" and "frag" should be deleted and whole phrase should be treated as term.

  Example: hybrid receptor composed of the extracellular and transmembrane regions of the human type II interleukin-1 fused to the cytoplasmic domain of the human type I IL-1R

  >There were "cons" and "frag" in the reddish phrase, but it should be term so they were deleted.

- When the name tag can be set inside of the "cons" and "frag" of the coordination. And in this case, "cons" and "frag" con't have to be deleted.

8.2. Event ontology



< Examples of the major categories of the EVENT ontology>

○ Artificial_process: Experimental process as “treatment”, “incubation”, “exposure” , etc.

○ Biological_process: “development”, “chemotaxis”, “response to stress”, etc.

● Cell adhesion: “cell adhesion” (inc. “cell-cell” and “cell-substance”)

● Cell recognition: “cell recognition” (inc. “cell-cell” and “cell-substance”)

● Cellular physiological process: “cell motility”, “cell cycle”, “cell migration”, “cell proliferation”, “cell killing”, “cell growth””cell cycle”, “cell homeostasis”, ”regulation of membrane potential”, etc.

○ Cell differentiation: “cell maturation”, etc.

○ Cellular_communication: “signal transduction”, etc.

○ Physiological_process: “immune response” (antigen presentation etc.), “cell activation”, “homeostasis”, “Ig class switch”, “excretion” etc.

○ Localization: “secretion”, “transport”, “translocation”, “release”, “disposition”, etc.

○ DNA_modification: “DNA methylation” etc.

○ DNA_recombination: “gene rearrangement” etc.

○ DNA metabolism: “DNA metabolism”, “mutagenesis”, “DNA replication”, etc.

○ Gene_expression: “xxx protein expression”, “xxx gene expression”, “xxx production”, “xxx synthesis” etc.

○ Protein_biosynthesis: (This class was set as a intermediate node, so it does not have to be used.)

○ Translation: “translation” and “expression” that is used as a specific term to describe as “protein expression”. This “expression” should be distinguished from “mRNA expression”.

○ Protein_catabolism: “protein degradation” etc.

○ Protein_modification: “protein processing” should be included in this class.

○ Protein metabolism: “protein maturation”

○ RNA metabolism: “RNA processing”, “RNA polyadenylation”, “RNA capping”, “RNA splicing”, etc.

○ Transcription: “transcription”, “mRNA expression”, etc.

○ Negative_regulation: “inhibit”, “suppress”, “down regulate”, “attenuation”, etc.

○ Positive_regulation: “initiate”, “promote”, “activate”, “enhance”, “essential”, etc.

○ Regulation: “regulate”, “modulate”, “modulate”, “control”, “response”, “alter”, “change”, etc.

○ Initiation_of_viral_infection: “virus infection”

○ Viral_genome_expression: “virus gene expression”

○ Viral_life_cycle: “viral replication”, “viral integration” and the transcription caused by the virus, etc.

○ Correlation: “correlate”, “involve”, “associate”, “synergize”, etc.

○ Binding: “bind”, “recognize”, “engage”, “interact”, “crosslink”, “form complex”, “conjugation”, “aggregation”, etc. This class can be used not only between the molecules but also the molecule and the cell. But the cell-to-cell adhesion should be classified in “Cellular_process”)

○ Catalysis: “catalyze” etc. But when this process is seen from the side of the substrate, it is a “metabolism”.)

<Notes>

- The class should be chosen from the lower layer of the ontology tree.

- The examples shown in the previous pages are the absolutely examples and the exceptional cases might be there. The class should be chosen contextually appropriately.

  Examples

  “require” can be classified in the “positive_regulation”, “regulation” or “correlation”, etc., according as the context.

  “target” can be classified in the “bind”, “regulation”, etc., according as the context.

  “detect” can be the term means the “expresson”.

  “secretion” can be the term means “expression”.

- The term that has 2 different EVENT categories, can be classified in the either class.

  Examples; “positive_regulation” can be the TYPE of this term in these two examples.

  transactivation = “transcription” + “positive_reguration”

  upregulates transcriptionally = “transcription” + “positive_regulation”

- The examples of “SOO”, short of ontology

Examples

The process what is related to the process of the disease, oncogenesis and pathogenesis etc., can be SOO.

“dissociation” falls into the counter-concept of “bind”, but it cannot be included the “non-exist” and “negative_regulation”.