Filip Ginter | Sampo Pyysalo | Jari Björne |
Juho Heimonen | Tapio Salakoski

BioInfer Relationship Annotation Manual

Turku Centre *for* Computer Science

TUCS Technical Report
No 806, August 2007

BioInfer Relationship Annotation Manual

Filip Ginter
Turku Centre for Computer Science (TUCS) and
Department of Information Technology, University of Turku
filip.ginter@it.utu.fi

Sampo Pyysalo
Turku Centre for Computer Science (TUCS) and
Department of Information Technology, University of Turku
sampo.pyysalo@it.utu.fi

Jari Björne
Department of Information Technology, University of Turku
jari.bjorne@utu.fi

Juho Heimonen
Department of Information Technology, University of Turku
juho.heimonen@utu.fi

Tapio Salakoski
Turku Centre for Computer Science (TUCS) and
Department of Information Technology, University of Turku
tapio.salakoski@it.utu.fi

TUCS Technical Report
No 806, August 2007

Abstract

This annotation manual supplements the paper *BioInfer: a corpus for information extraction in the biomedical domain (Pyysalo, Ginter, Heimonen, Björne, Boberg, Järvinen, Salakoski. BMC Bioinformatics, 8(50), 2007)* that describes the BioInfer corpus. The manual details the rules guiding the annotation of entities and their relationships, extending the more general description given in the paper. The rules are given with examples clarifying their use and organized according to topic. This manual can serve as a reference for the users of the corpus as well as for producing annotation following the BioInfer scheme.

Keywords: corpus annotation, protein-protein interactions, BioInfer

TUCS Laboratory
BioInformatics Laboratory

Contents

1 Preamble

This manual details the rules that determine how the BioInfer corpus annotation should be produced. BioInfer uses the Link Grammar scheme [3] as well as the Stanford dependency scheme [1] for syntactic annotation. This manual describes the rules for entity name annotation and relationship annotation. The manual is not intended as an introduction to the BioInfer corpus, for this see [2].

1.1 Terminology

A bioentity is a physical, biochemical object which has *properties* and is involved in *processes*. *Relationships* exist between bioentities, properties, processes, and other relationships.

Bioentities, their properties and processes are jointly referred to as *entities*. Annotated established names of bioentities of the gene, protein and RNA types are termed *basic named entities* and processes, properties and other physical entities pertaining to basic named entities are termed *extended named entities*. When discussing an extended named entity, the basic named entity that extended entity pertains to is referred to as the *core named entity*. Other annotated entities are termed *unnamed entities*.

Each entity has *a type* and *a text binding*. The type is assigned from *the entity type ontology*. The text binding identifies the part of the sentence text that specifies the entity. Entities for which there is no associated text present in the sentence are referred to as *anonymous entities*. Extended entity annotation allows *nesting*, where one entity is (textually and conceptually) contained within another.

Relationships are annotated with *relationship formulas*, which consist of *a predicate* and its *arguments*. The predicate is selected from *the relationship type ontology*. As an example, in the formula *BIND(actin*, *profilin)* the predicate is *BIND* and the arguments are *actin* and *profilin*. A relationship formula has a text binding similar to that of entities when applicable.

1.2 Conventions used in this manual

The full annotation is unnecessarily complex for demonstration purposes for most features discussed in this manual. Therefore, the annotation presented in the examples is simplied to focus on the feature currently being discussed. The examples given in this manual follow one of the following conventions in their presentation.

The full annotation including entities, their nesting, identitifiers and text bindings, as well as relationships, their defining predicates, and text bindings, is presented as in (i). The syntax $<_{e}$*text*$_{e}>$ indicates that *text* is the annotated text binding of the entity or formula *e*. The bracketing identifies the full nesting of the entities.

If the identities of the entities and formulas are clear from the context, the identifying letters may be omitted and the specifying text used directly (ii). The bracketing may be omitted in those examples that do not relate to nesting (iii). Similarly, the relationship formulas and the text binding annotation for predicates can be omitted when discussing only the nesting of extended named entities (iv). Finally, in those examples that are related only to the identification of basic named entities, the convention is to simply list the basic named entities (v). If the entity types are being discussed, the entity type annotation is shown as in (vi).

1.2.1 i. $<_{a}$alpha-catenin$_{a}>$ $<_{A}$inhibits$_{A}>$ $<_{b}<_{c}$beta-catenin$_{c}>$ signaling$_{b}>$
A:SUPPRESS(a, b)

ii. <alpha-catenin> <inhibits> <<beta-catenin> signaling>
inhibits:SUPPRESS(alpha-catenin, beta-catening signaling)

iii. alpha-catenin inhibits beta-catenin signaling
inhibits:SUPPRESS(alpha-catenin, beta-catening signaling)

iv. alpha-catenin inhibits beta-catenin signaling
<alpha-catenin>, <<beta-catenin> signaling>

v. alpha-catenin inhibits beta-catenin signaling
alpha-catenin, beta-catenin

vi. alpha-catenin inhibits beta-catenin signaling
alpha-catenin - Individual_protein
beta-catenin - Individual_protein
beta-catenin signaling - Function_property

Note that these are merely presentation conventions. In the actual annotation, full detail is always included.

2 Entity names

2.1 Basic named entities

Synopsis: The names of proteins, genes and RNA as well as the names of complexes, families or groups of these types are annotated as basic names, without regard to their inner structure. Short established names, such as *H4*, are annotated normally, and even a single letter can be annotated as a basic name if it is defined within the sentence as an abbreviation for a basic name.

2.1.1 Types of annotated basic named entities

Annotation: Established names of proteins, genes, RNAs, and their complexes, as well as established names of structurally or functionally well-defined families or groups of these types are annotated as basic names (i–vii). The smallest annotated bioentity names are thus units of quaternary structure – for example, the names of domains or substructures are not annotated (viii–ix). Established names of bioentities larger than those mentioned above are annotated if their type is found in the *substance* subtree of the entity type ontology. To determine whether a candidate name is to be considered an established name, typical use in literature and databases, such as Swiss-Prot, should be followed.

2.1.1 i. myosin heavy chain, desmin, tropomyosin, and a fraction of the actin were all cleaved
myosin heavy chain, desmin, tropomyosin, actin

ii. Immunofluorescence microscopy confirmed the presence of paxillin, talin, and vinculin
paxillin, talin, vinculin

iii. RAD51, RAD52, and RAD54 encode proteins that are critical to . . .
RAD51, RAD52, RAD54

iv. Herpes simplex virus type 1 (HSV-1) encodes a heterotrimeric helicase-primase
helicase-primase

v. The genes considered are those for 5S, 5.8S and 18S rRNA, . . .
5S rRNA, 5.8S rRNA, 18S rRNA

vi. the beta 1 integrin subunit
beta 1 integrin subunit

vii. Arp2 subunit
Arp2

viii. the death domains of FAS and RIP
FAS, RIP

ix. the proximal S2 domain
-

In (v), *5S rRNA*, *5.8S rRNA*, and *18S rRNA* are the names of specific RNA molecules while *5S*, *5.8S*, and *18S* are merely the descriptions of the sizes of those molecules (see Section 2.1.3). Note that in (vi) *beta 1 integrin subunit* is annotated as a whole because it is a unit of a quarternary structure. Here *subunit* is an essential part of the name while in (vii) *Arp2* alone is sufficient to identify the bioentity. Note also that *beta 1 integrin* is not annotated because it is nested within another basic named entity (see Section 2.1.2).

2.1.2 Inner structure of basic named entities

Annotation: Basic named entities do not nest, and are annotated without regard to their inner structure. This applies even if the structure contains a candidate name or a relationship that would be annotated if it were not contained within the name. Each basic named entity is annotated as a whole, and if several candidate names are nested, the broadest, enclosing name is chosen (i–iv).

2.1.2 i. actin-depolymerizing factor (ADF)
actin-depolymerizing factor, ADF
ii. actin-binding proteins (ABPs)
actin-binding proteins, APBs
iii. cyclin-dependent kinase inhibitor p21
cyclin-dependent kinase inhibitor, p21
iv. mitogen-activated protein kinase (MAPK) kinases
mitogen-activated protein kinase kinases, MAPK

In (i–ii), neither *actin* nor the relationships implied by *depolymerizing* and *binding* are annotated because *actin-depolymerizing factor* and *actin-binding proteins* are established names. Similarly, in (iii), neither of the candidate names *cyclin* or *kinase* nor the relationships implied by *dependent* and *inhibitor* are annotated as these occur within the established protein family name *cyclin-dependent kinase inhibitor*. Note that in (iv), *MAPK* is an abbreviation of *mitogen-activated protein kinase* which is not annotated as it is contained within a larger basic name.

2.1.3 Short basic names

Annotation: Typically, short basic names cannot easily be distinguished from identifiers that specify the members of families (i–ii). To determine whether a short candidate name should be annotated, the general rules, including consulting literature and databases, should be followed (Section 2.1.1). Single letters and numbers are normally not considered names (ii), but any abbreviation defined in the sentence is considered a name, even if it is only a single letter (iii–iv).

2.1.3 i. histone H4
histone, H4
ii. cyclin D
cyclin D
iii. nucleocapsid protein (NP) and phosphoprotein (P)
nucleocapsid protein, NP, phosphoprotein, P

iv. nucleocapsid protein (N) . . . recombinant N proteins
nucleocapsid protein, N, N

2.2 Extent of basic named entities

Synopsis: Basic named entities are limited to certain syntactic structures, mostly elementary noun phrases. In addition, only modifiers specifying the source of the referred bioentity within an organism are included. Basic names typically consist of a continuous sequence of words, however discontinuity is allowed in the annotation when necessary.

2.2.1 Syntactic structure

Annotation: A basic name is typically an elementary noun phrase consisting of the head noun and some of its premodifiers (i–ii) (see also Sections 2.2.2 and 2.2.3). When they occur as part of an established name, numeric postmodifiers are included (iii–v). Prepositional phrases, participle modifiers, and relative clauses are never included in basic named entities (vi–viii) (see also Section 2.4.1).

2.2.1 i. alpha-smooth muscle actin (alpha-SMA)
alpha-smooth muscle actin, alpha-SMA

ii. beta 1-syntrophin
beta 1-syntrophin

iii. the presence of thymosin beta 4 (T beta 4)
thymosin beta 4, T beta 4

iv. dipeptidyl peptidase IV
dipeptidyl peptidase IV

v. the soluble phospholipase C-gamma 1
phospholipase C-gamma 1

vi. brain profilin
brain profilin

vii. profilin from porcine brain
profilin

viii. smooth muscle talin prepared from chicken gizzard
smooth muscle talin

In (vi) *brain* is included because it occurs as a premodifier for *profilin* while in (vii) *brain* is excluded as it occurs in a prepositional phrase.

2.2.2 Discontinuous names

Annotation: Discontinuous basic names typically arise from coordinations (i–iii), in cases where e.g. an abbreviation or part of the name appears in parentheses (iv–vi) and in some cases where words that are not considered part of the name appear in a sequence of words that are (vii–viii). Discontinuous names are annotated as detailed in (i).

2.2.2 i. $<_{ab}$TRAIL receptors$_{ab}>$ $<_{a}1_{a}>$ and $<_{b}2_{b}>$
TRAIL receptors 1, TRAIL receptors 2
ii. myosin heavy chain and light chains
myosin heavy chain, myosin light chains
iii. The two cardiac myosin heavy chain isoforms, alpha and beta
cardiac myosin heavy chain alpha, cardiac myosin heavy chain beta
iv. DNA polymerase (Pol) delta
DNA polymerase delta, Pol
v. alpha-smooth muscle (SM) actin
alpha-smooth muscle actin
vi. cytokeratins (7 and 20)
cytokeratins 7, cytokeratins 20
vii. CD44 isoform 6
CD44 6
viii. profilin mutant (H119E)
profilin H119E

In (v) *SM* is not annotated as it is an abbreviation of *smooth muscle*, which is not an annotated named entity. In (vii) the word *isoform* is not part of the name but rather specifies the type (Section 2.2.3). In other words, the occurrence of the number *6* does not influence the decision whether *isoform* is included in the basic named entity.

2.2.3 Included modifiers

Annotation: The words that specify the source of the bioentity within an organism, such as a tissue or an organ, are included in basic named entities when these words occur in an elementary noun phrase (i–ii) (see Section 2.2.1). In contrast, organism names are not included (iii) (see also Section 2.4.3). Descriptive modifiers specifying weight, type, form, location, modification, and other similar properties are not parts of basic names either (iv–x).

2.2.3 i. skeletal muscle actin
skeletal muscle actin
ii. lymphocyte talin
lymphocyte talin
iii. yeast cofilin
cofilin

iv. the 265-kDa co-activator protein CREB-binding-protein
CREB-binding-protein

v. Fluorescently-labeled fimbrin
fimbrin

vi. endogenous alpha-catenin
alpha-catenin

vii. LIMK2 mutant with replacement of threonine 505 by valine
LIMK2

viii. cofilin isoforms
cofilin

ix. Immunoreactive E-cadherin, alpha-catenin, beta-catenin, and gamma-catenin proteins
E-cadherin, alpha-catenin, beta-catenin, gamma-catenin

x. nucleocapsid protein and phosphoprotein gene sequences
nucleocapsid protein, phosphoprotein

Note that the word *proteins* in (ix) only defines the type of the entities *E-cadherin*, *alpha-catenin*, *beta-catenin*, and *gamma-catenin* while the word *protein* in (x) is a part of the name *nucleocapsid protein*: by itself, *nucleocapsid* is not a basic name while *E-cadherin*, *alpha-catenin*, *beta-catenin*, and *gamma-catenin* are (Section 2.1.1).

2.3 Extended named entities

Synopsis: Entities that participate in an annotated relationship and pertain to a basic named entity are annotated as extended named entities. In the following, we use the term *core named entity* to refer to the basic named entity that an extended named entity pertains to.
Annotation: When a participant in an annotated relationship is not a basic named entity, it is annotated as an extended named entity that nests the core named entity. Extended named entities are always annotated with full nesting, each enclosed extended named entity separately annotated, with the core named entity as the innermost level.

2.3.1 i. TGFbeta regulates clusterin gene expression
<TGFbeta>, <<<clusterin> gene> expression>

ii. alpha-catenin expression did prevent accumulation of beta-catenin
<<alpha-catenin> expression>, <accumulation of <beta-catenin>>

2.4 Extent of extended named entities

Synopsis: Extended named entities are typically noun phrases with phrase complements. Each extended named entity consists of the words that specify its type and the words of the entity nested within. Organism names are included, if allowed by the other rules, in those extended named entities that directly nest a basic named entity.

2.4.1 Syntactic structure

Annotation: An extended named entity always contains the head noun of its corresponding noun phrase as well as the nested entity. Primarily, an extended named entity consists of a head noun and either a premodifier or prepositional phrase that contains the nested entity (i–ii). The preposition of a prepositional phrase is included in the annotation. Extended named entities can also contain participle modifiers (iii).

2.4.1 i. abundance of alpha 5 integrins
<abundance of <alpha 5 integrins>>

ii. RAD51 overexpression
<<RAD51> overexpression>

iii. aggregation induced by thrombin
<aggregation induced by <thrombin>>

Note that the prepositions *of* in (i) and *by* in (iii) are annotated.

2.4.2 Included modifiers

Annotation: At each level of nesting in extended named entities, the words that directly contribute to the type of the extended named entity at that level are included (i–ii). Further, if the type is specified by a name (even if not an annotated name) or an established multi-word phrase that would be assigned a different type than its components, it is included as a whole (iii). Other words are omitted.

Some adjectives, such as *monomeric* and *polymeric*, imply the type in certain expressions. However, as the type is only indirectly defined by these words they are excluded from any extended entity (iv–v).

2.4.2 i. alpha-catenin is not affected by the differential calreticulin expression
<alpha-catenin>, <<calreticulin> expression>

ii. high concentrations of Acanthamoeba profilin inhibit the elongation rate of muscle actin filaments
<concentrations of <Acanthamoeba <profilin>>>,
<rate of <<muscle actin> filaments>>

iii. glutamic acid of actin
<glutamic acid of <actin>>

iv. monomeric actin
<actin>
v. the protein actin
<protein <actin>>

In (i) and (ii), the words *differential*, *high* and *elongation* are not part of the extended named entities as the type of the entities can be resolved without these words. In (iii) *glutamic* is included as omitting it would split the term *glutamic acid*, changing the type of the extended named entity.

In (iv) and (v), both *monomeric* and *protein* indicate that *actin* is a protein. However, in (iv), *monomeric* specifies the type of *actin* only indirectly and *monomeric actin* is thus not annotated while, in (v), *protein actin* and *actin* are annotated as different entities since the word *protein* directly specifies the type of *actin*.

2.4.3 Organism names

Annotation: When an organism name is annotated, it is included in an extended named entity that directly nests the core named entity and consists of the organism name and the basic name (i). In addition, the words specifying the type of the core named entity (and hence also, by definition, that of the extended named entity) are included if present (ii). The annotated extended named entity is nested in all other extended named entities that share the same core named entity (iii).

An organism name is annotated if it appears either as a noun pre-modifier or in a prepositional phrase for a basic name (iv–v). In addition, the annotated extended named entity, or an extended named entity that nests it, must be a participant in a stated relationship (Section 2.3).

2.4.3 i. Acanthamoeba profilin
<Acanthamoeba <profilin>>
ii. RVS161 products of Saccharomyces cerevisiae
<<RVS161> products of Saccharomyces cerevisiae>
iii. Acanthamoeba actin polymerization
<<Acanthamoeba <actin>> polymerization>
iv. Herpes simplex virus type 1 (HSV-1) encodes a heterotrimeric helicase-primase
<helicase-primase>
v. actin and profilin were purified from Amoeba
actin, profilin

2.4.4 “Gene product”

Annotation: The structure of the phrases of the type *RAD51 gene product* can be interpreted as either <<*RAD51*> *gene product*> or <<*RAD51 gene*> *product*>. In the annotation, the former interpretation is followed. Under this interpretation the multiword expression *gene product* is annotated as in (i) (see Section 2.4.2).

2.4.4 i. RAD51 gene product
<<*RAD51*> *gene product*>

3 Relationships

3.1 Negation

Synopsis: The predicate *NOT* is used to annotate any explicit statements of the non-existence of a relationship. Different levels of confidence are not considered (see Section 3.6).
Annotation: Simple negations (i), negative statements (ii–iii), statements of independence (iv) or other statements that imply negation are annotated with the predicate *NOT*. The statements regarding the level of confidence in the non-existence of the relationship are not considered (Section 3.6).

The predicate *NOT* takes as its argument the negated relationship. Typically, the predicate is bound to simple negators such as *no*, *not*, or *without*. However, the statement of non-existence may also be more complex as in (ii–iii). Both the predicate of the positive relationship and the *NOT* predicate may have the same text binding (iv).

3.1.1 i. Abundance of actin is not affected by calreticulin expression.
not:NOT(affected by:AFFECT(abundance of actin, calreticulin expression))
ii. N-WASP mutant unable to interact with profilin
unable to:NOT(interact with:BIND(N-WASP mutant, profilin))
iii. Actin mutant is defective in binding fimbrin.
defective in:NOT(binding:BIND(actin mutant, fimbrin))
iv. Phosphorylation of beta-catenin is independent of dissociation of alpha-catenin from E-cadherin.
independent of:NOT(independent of:AFFECT(phosphorylation of beta-catenin, dissociation of from:UNBIND(alpha-catenin, E-cadherin)))

3.2 Equality

Synopsis: The annotation includes two main kinds of equality: synonymy and coreference. These terms are understood broadly. For example, synonymy includes abbreviation definitions and coreference includes pronouns and anaphora. Synonymy is annotated with the predicate *EQUAL* and coreference with the predicate *COREFER*.

3.2.1 Synonymy

Annotation: Synonymy is always annotated — even if the entities do not take part in any other relationships. Synonymy is annotated with the predicate *EQUAL* such that the first argument is the referred entity and the second argument the referring entity. Only the referred entity is used in other relationships. If the synonymy is stated using a phrase like *also called* or *also known as*, this phrase is used as the text binding for the predicate (ii, iv). Commonly, there is no text binding for the *EQUAL* predicate (i).

3.2.1 i. MORT1 (FADD)
EQUAL(MORT1, FADD)

ii. MORT1 (also called FADD)
also called:EQUAL(MORT1, FADD)

iii. components of the pathway such as HCS77/WSC1/SLG1
EQUAL(HCS77, WSC1) EQUAL(HCS77, SLG1)

iv. CREB binding protein referred to as CBP
referred to as:EQUAL(CREB binding protein, CBP)

v. a proline-rich protein (End5p or verprolin)
EQUAL(End5p, verprolin)

vi. act1-157, an actin mutant
EQUAL(act1-157, actin mutant)

vii. mitogen-activated protein kinase (MAPK) kinases
mitogen-activated protein kinase kinases, MAPK

Note that in rare cases, such as (vii), the abbreviation relationship is omitted because the full name (here *mitogen-activated protein kinase*) is not annotated (see Section 2.1.2).

3.2.2 Coreference

Annotation: Coreference is annotated only when it is needed for capturing another relationship. Relationships stated using the referring entity (e.g. a pronoun) are annotated using the referring entity (as opposed to synonymy, where always the referred entity is used). Coreference is not annotated if the entities are closely syntactically bound (see Section 3.7 for definition).

3.2.2 i. Addition of profilin to polymerized actin causes it to depolymerize.
COREFER(it, actin)
causes:CAUSE(addition of profilin, depolymerize:DEPOLYMERIZE(it))

ii. Verprolin establishes and maintains its location independent of the actin cytoskeleton.
COREFER(its, verprolin)
independent of:NOT(independent of:AFFECT(actin cytoskeleton, its location))

iii. Studies demonstrated the gamma-catenin distribution to be remarkably similar to that of beta-catenin.
COREFER(that, distribution)
similar to:SIMILAR(gamma-catenin distribution, that of beta-catenin)

iv. a heterotrimeric helicase-primase, the subunits of which are encoded by the UL5, UL8 and UL52 genes
COREFER(which, helicase-primase)
encoded by:ENCODE(UL5 genes, subunits of which)
encoded by:ENCODE(UL8 genes, subunits of which)
encoded by:ENCODE(UL52 genes, subunits of which)

v. A substitution found in talin does not destroy the capacity of the protein to bind actin.
COREFER(protein, talin)
bind:BIND(protein, actin)

vi. Segments of PRT1 and TIF35 were found to be responsible for $<_{ab}$their$_{ab}>$ binding to TIF34.
COREFER(a, PRT1)
COREFER(b, TIF35)
responsible for:CAUSE(segments of PRT1, binding to:BIND(a, TIF34))
responsible for:CAUSE(segments of TIF35, binding to:BIND(b, TIF34))

3.3 Membership

Synopsis: Family membership is stated through the predicate *MEMBER*.
Annotation: Both implicit (i–iv) and explicit (v–viii) statements of gene/protein family memberships are annotated. Implicit statements are typically in the form of apposition while explicit statements are commonly expressed with phrases like *identified as* and *such as*.

The family may remain unnamed if there are several membership statements related to it (ix). In this case, the family memberships relate the members together even though the family itself is unnamed.

A membership is annotated with the predicate *MEMBER*. The first argument in the predicate is the family name, the second argument the member name. As shown in (vii), the member can be not only an individual gene/protein but also another family.

3.3.1 i. pp125 focal adhesion kinase
MEMBER(focal adhesion kinase, pp125)

ii. activation of BCK1 (MEKK)
MEMBER(MEKK, BCK1)

iii. <cyclin-dependent kinase inhibitor> <<p21> expression>
MEMBER(cyclin-dependent kinase inhibitor, p21)

iv. syntaxin 3, a t-SNARE localized to the apical plasma membrane
MEMBER(t-SNARE, syntaxin 3)

v. profilin is identified as an actin-binding protein
identified as:MEMBER(actin-binding protein, profilin)

vi. Actin-binding proteins such as profilin
such as:MEMBER(actin-binding proteins, profilin)

vii. Death receptors belong to the TNF receptor family
belong to:MEMBER(TNF receptor family, death receptors)

viii. Actin-depolymerizing factor (ADF) and cofilin define a family of actin-binding proteins
define family:MEMBER(family of actin-binding proteins, actin-depolymerizing factor)
define family:MEMBER(family of actin-binding proteins, cofilin)
EQUAL(actin-depolymerizing factor, ADF)

ix. cdc12p is a member of a family of proteins including BNI1 and fus1
member of:MEMBER(family, cdc12p)
including:MEMBER(family, BNI1)
including:MEMBER(family, fus1)

3.3.1 MEMBER vs. EQUAL

In certain cases, the statements of synonymy and family membership are syntactically indistinguishable (i–iii). Even in these cases, the correct relationship is annotated, making reference to external knowledge as necessary.

3.3.2 i. Overexpression of cyclin-dependent kinase inhibitor (p27Kip1)
MEMBER(cyclin-dependent kinase inhibitor, p27Kip1)
ii. cyclin-dependent kinase inhibitor (CKI)
EQUAL(cyclin-dependent kinase inhibitor, CKI)
iii. the p27 cyclin-dependent kinase inhibitor (CKI)
EQUAL(cyclin-dependent kinase inhibitor, CKI)
MEMBER(cyclin-dependent kinase inhibitor, p27)

3.4 Implicit reference

Synopsis: The standard entity nesting mechanism cannot be used if the core named entity is only implicitly referred to. In these cases, the predicate *REL-ENT* is used to indicate the reference.
Annotation: There are sentences in which relationships involve extended named entities that do not nest the core named entity but implicitly refer to it. The core named entity is attached to the extended named entity with the predicate *REL-ENT* and the annotated relationship carries the same meaning as nesting does. For instance, in (i) the annotated reference corresponds to the nesting <<*alpha-catenin*> *residues*>. This relationship is annotated only when the extended named entity is involed in another relationship (as is the case with the regular nesting mechanism, see Section 2.3).

In *REL-ENT*, the first argument is always the extended named entity and the second the core named entity.

3.4.1 i. Analysis of alpha-catenin revealed that residues 48-163 are able to bind to beta-catenin.
REL-ENT(residues, alpha-catenin)
bind to:BIND(residues, beta-catenin)
ii. The stretch of residues 633-642 of the myosin heavy chain is part of the actin-binding site.
REL-ENT(site, myosin heavy chain)
binding:BIND(site, actin)
iii. Effects of substitutions in the actin-binding site on the activity of profilin
REL-ENT(site, profilin)
binding:BIND(site, actin)

iv. Addition of profilin to actin filaments causes depolymerization.
REL-ENT(depolymerization, actin filaments)
causes:CAUSE(addition of profilin, depolymerization)

v. Ser-133 phosphorylation enhances CREB activity.
REL-ENT(phosphorylation, CREB)
enhances:ACTIVATE(phosphorylation, CREB activity)

vi. The binding surfaces for segment 1 and profilin overlap on actin.
REL-ENT(surfaces, actin)
binding:BIND(surfaces, profilin)

Note that in (v), *Ser-133* specifies a position where the phosphorylation occurs rather than a named entity.

3.5 Anonymous entities

Synopsis: Anonymous entities are used when a relationship between two named entities cannot be expressed without a third entity that is not stated in the text.
Annotation: Anonymous entities are primarily used in the annotation of complexes (see Section 3.9.1). Another use is to annotate entities that are omitted by passive constructions (i–ii). Anonymous entities are annotated only when they are needed to express a relationship between two named entities.

Anonymous entities are indicated using the predicate *ANONYMOUS*.

3.5.1 i. The severing activity of cofilin is activated, leading to the generation of actin filaments.
ANONYMOUS(X)
leading to:CAUSE(activated:ACTIVATE(X, activity of cofilin), generation of actin filaments)

ii. p130 was phosphorylated, despite the expression of p21.
ANONYMOUS(X)
despite:NOT(despite:PREVENT(expression of p21, phosphorylated:PHOSPHORYLATE(X, p130)))

iii. Phosphorylation of p130 occurred despite the expression of p21.
despite:NOT(despite:PREVENT(expression of p21, phosphorylation of p130))

Note that (ii) and (iii) contain the same information but the latter can be annotated without an anomynous entity.

3.6 Statements of confidence

Synopsis: The reported relationships can be stated in the text with different levels of confidence or conclusiveness. These levels are not captured in the annotation and all relationships are annotated regardless of the strength with which they are asserted.

Annotation: Hedging terms such as *we suggest that*, *we studied*, and *may be responsible for* are often used to express the uncertainty involved in the conclusions of the study. In these cases, the statements in the sentence do not assert a relationship. However, from a practical IE point of view, even non-affirmative relationships are of interest and are thus annotated.

There is currently no annotation mechanism to capture the various levels of confidence expressed in the statements. All statements are annotated regardless of their level of confidence. Negative statements are annotated as in Section 3.1. Other statements are annotated as positive relationships.

3.6.1 i. The effect of profilin on the thermal stability of actin was studied.
effect of on:AFFECT(profilin, stability of actin)

In (i), the sentence does not claim that *profilin* has an effect on *the thermal stability of actin*. However, it contains a positive statement, and that statement is annotated as a positive relationship.

3.7 Syntactic binding

Synopsis: The *syntactic binding* rule allows to omit coreference relationships that are trivially implied by the syntactic structure. When two entities are syntactically bound, their *COREFER* relationship is not annotated. Consequently, if a named entity and a common noun referring to the entity are syntactically bound, their relationship is not annotated and the named entity is used directly in relationships.
Annotation: Two entities are considered *syntactically bound* if they are related by an elementary syntactic construction (apposition and statements with copula) that specifies a *COREFER* relationship between them. In this case, the *COREFER* relationship is considered elementary and is not annotated. Consequently, if one of the two entities is unnamed, it is omitted from the annotation (i–v). If both of the entities are named, all their relationships other than *COREFER* must be annotated (vi).

In the most typical case, there is a simple apposition dependency between the syntactically bound entities (i–ii). The entities may also be separated, for example, by a simple connecting phrase (iii), a chain of coreferences (iv), or a copula expression (v).

In the examples below, the annotation without ("without s.b.") and with ("with s.b.") syntactic binding is shown in order to illustrate the effect of the syntactic binding on the annotation. All examples are annotated in the corpus according to the "with s.b." alternative.

3.7.1 i. *the vinculin-binding protein, talin*
without s.b.: *COREFER(protein, talin)*
binding:BIND(protein, vinculin)
with s.b.: *binding:BIND(talin, vinculin)*

ii. *talin, a known substrate of calpain*

without s.b.: *COREFER(substrate, talin)*
substrate of:AFFECT(calpain, substrate)
with s.b.: *substrate of:AFFECT(calpain, talin)*

iii. *IgG binds to two platelet components (identified as vinculin and talin).*

without s.b.: *identified as:COREFER(components, vinculin)*
identified as:COREFER(components, talin)
binds to:BIND(IgG, components)
with s.b.: *binds to:BIND(IgG, vinculin)*
binds to:BIND(IgG, talin)

iv. *Arp2/3 consists of seven polypeptides; two actin-related* $<_{a}$*proteins*$_{a}>$*, Arp2 and Arp3; and five novel* $<_{b}$*proteins*$_{b}>$*, p40, p35, p19, p18, and p14.*

without s.b.: *consists of:CONTAIN(Arp2/3, polypeptides)*
COREFER(polypeptides, a)
COREFER(polypeptides, b)
COREFER(a, Arp2)
COREFER(a, Arp3)
COREFER(b, p40)
. . .
with s.b.: *consists of:CONTAIN(Arp2/3, Arp2)*
consists of:CONTAIN(Arp2/3, Arp3)
consists of:CONTAIN(Arp2/3, p40)
. . .

v. *CD26 is an antigen known to bind adenosine deaminase.*

without s.b.: *is:COREFER(antigen, CD26)*
bind:BIND(antigen, adenosine deaminase)
with s.b.: *bind:BIND(CD26, adenosine deaminase)*

vi. *The monomeric actin-binding protein, profilin, is a key regulator of actin-filament dynamics.*

without s.b.: *MEMBER(actin-binding protein, profilin)*
COREFER(actin-binding protein, profilin)
regulator of:CONTROL(actin-binding protein, actin-filament dynamics)
with s.b.: *MEMBER(actin-binding protein, profilin)*
regulator of:CONTROL(profilin, actin-filament dynamics)

In (i), *protein* and *talin* are syntactically bound with a single apposition dependency. Similarly, in (ii) *talin* and *substrate* are syntactically bound. In (iii) the syntactic binding occurs through a simple phrase (*identified as*) while in (iv) *polypeptides* is related to *Arp2*, *Arp3*, and *p40* (etc.) through a chain of two appositions and in (vi) *CD26* and *antigen* are related through a copula expression. In (vi), the syntactically bound entities (*actin-binding protein* and *profilin*) are both named and thus their relationship (*MEMBER*) is annotated even though the syntactic binding is utilised in the other relationship (*CONTROL*). Note that, in this example, *actin-binding protein* is both the family name and a coreference for *profilin*. Thus the relationships in this sentence cannot be correctly annotated without syntactic binding.

3.8 Inference of relationships

Synopsis: Given that all annotated relationships must follow the general relationship rules, atomic relationships are annotated always and inferred relationships only when all the atomic relationships needed for the inference are not annotated.
Annotation: We define an *atomic relationship* as the simplest relationship among any entities (annotated or not) in the sentence. An *inferred relationship* is obtained by a biologically meaningful inference from atomic relationships.

An atomic relationship is explicitly annotated if it conforms to the general rules regarding relationships. Thus, for instance, the atomic relationships in (i) are not annotated, because *region* is not an annotated entity (not having a name at the inner-most level of nesting).

An inferred relationship is annotated unless all the atomic relationships necessary for the inference are already annotated. In addition, the general requirements regarding relationships must be fulfilled. In (i), the two atomic relationships are not annotated and thus the inferred relationship must be annotated. In (ii), the atomic relationships are annotated and therefore the inferred relationships are not.

Families, complexes, and fusion proteins can be annotated as unnamed entities. Thus in (iii) *family* is an annotated entity and the four atomic relationships are annotated. As a result, no inferred relationship remains to be annotated.

3.8.1 i. *Profilin has an actin-binding region.*
atomic: *has:SUBSTRUCTURE(profilin, region)*
atomic: *binding:BIND(region, actin)*
inferred: *binding:BIND(profilin, actin)*

ii. *Actin-binding proteins such as profilin and gelsolin regulate actin.*
atomic: *such as:MEMBER(actin-binding proteins, profilin)*
atomic: *such as:MEMBER(actin-binding proteins, gelsolin)*
atomic: *regulate:REGULATE(actin-binding proteins, actin)*
inferred: *regulate:REGULATE(profilin, actin)*
inferred: *regulate:REGULATE(gelsolin, actin)*

iii. *Cdc12p is a member of a family that includes BNI1 and fus1 and is involved in actin-mediated processes.*
atomic: *member of:MEMBER(family, cdc12p)*
atomic: *includes:MEMBER(family, BNI1)*
atomic: *includes:MEMBER(family, fus1)*
atomic: *involved in:PARTICIPATE(family, actin-mediated processes)*
inferred: *involved in:PARTICIPATE(cdc12p, actin-mediated processes)*
inferred: *involved in:PARTICIPATE(BNI1, actin-mediated processes)*
inferred: *involved in:PARTICIPATE(fus1, actin-mediated processes)*

3.9 Common special cases

Synopsis: The statements of protein complexes are mostly of the type *complex of A and B* but also the types *A forms a complex with B* and *A is complexed with B* are commonly used.

The constructions of the type *X related to Y via/through Z* or *X related to Y via/through relation to Z* are annotated with the predicates *MEDIATE* and *RELATE*.

The word *like* is commonly used in two different types of relationships: to state similarity and to state a relationship by analogy. The alternative appropriate in the current context is used.

3.9.1 Complexes

Annotation: Whenever a complex is stated as a noun phrase, the whole phrase is the text binding for the complex entity (i–ii). If there is no noun to bind the complex entity to, an anonymous entity is used as shown in (iii–iv). The complex entity may be unnamed, that is not to have a named entity at its innermost level of nesting (v).

3.9.1 i. complex of actin and profilin
complex of:CONTAIN(complex of actin and profilin, actin)
complex of:CONTAIN(complex of actin and profilin, profilin)

ii. complex containing actin and profilin
complex:CONTAIN(complex containing actin and profilin, actin)
complex:CONTAIN(complex containing actin and profilin, profilin)

iii. Cadherin is complexed with alpha-catenin and beta-catenin.
ANONYMOUS(X)
complexed with:CONTAIN(X, alpha-catenin)
complexed with:CONTAIN(X, beta-catenin)
complexed with:CONTAIN(X, cadherin)

iv. an assembly of UL5 and UL52
ANONYMOUS(X)
assembly of:CONTAIN(X, UL5)
assembly of:CONTAIN(X, UL52)

v. Munc18-2 forms a complex with syntaxin 3.
forms complex with:CONTAIN(complex, Munc18-2)
forms complex with:CONTAIN(complex, syntaxin 3)

Note that in (v) the relevant noun phrase is *complex* rather than *complex with syntaxin 3*.

3.9.2 "via" and "through"

Annotation: In the constructions of the type *X related to Y via/through Z* or *X related to Y via/through relation to Z*, *Z* mediates the relationship between *X* and *Y*. This mediation is annotated with the predicate *MEDIATE* (i–iii). Further, when there is no direct interaction between *X* and *Z*, the formula *RELATE(X, Z)* is used (ii–iii). Text binding for both *MEDIATE* and *RELATE* is the preposition *via*/*through*.

3.9.2 i. CREB effects transcription of NOSI via interaction with CBP
via:MEDIATE(interaction with:INTERACT(CREB, CBP),
effects:INITIATE(CREB, transcription of NOSI))

ii. beta2 integrins are linked to the actin cytoskeleton via talin
via:MEDIATE(talin, linked to:RELATE(beta2 integrins, actin cytoskeleton))
via:RELATE(beta2 integrins, talin)

iii. regulation of actin dynamics through phosphorylation of cofilin by
LIM-kinase
through:MEDIATE(phosphorylation of by:PHOSPHORYLATE(LIM-kinase,
cofilin), regulation of actin dynamics)
through:RELATE(LIM-kinase, regulation of actin dynamics)

3.9.3 "like"

Annotation: When *like* describes a similarity, the predicate *SIMILAR* is used with *like* as its text binding (i). When *like* is used to express analogy, *like* is added to the text binding of the predicate having the analogous entity as its argument (ii).

3.9.3 i. PRP5 (a DEAD box helicase-like protein)
like:SIMILAR(PRP5, DEAD box helicase)

ii. Like beta 1-syntrophin, alpha 1-syntrophin interacts with dystrophin.
interacts with:INTERACT(alpha 1-syntrophin, dystrophin)
like interacts with:INTERACT(beta 1-syntrophin, dystrophin)

4 Entity types

4.1 Basic named entities not involved in relationships

Synopsis: Basic named entities not involved in relationships are annotated as *gene/protein/RNA*.
Annotation: As the corpus focuses on relationships, basic named entities that are not involved in any annotated relationships and are not the core named entity for any extended named entity do not need to be annotated for their exact types. These basic named entities are annotated as *gene/protein/RNA*.

4.1.1 i. Actin expression was reduced by 25-50%, relative to that of myosin heavy chain
actin - *gene/protein/RNA*
myosin heavy chain - *gene/protein/RNA*

4.2 Basic named entities with ambiguous type

Synopsis: A basic named entity that is involved in a relationship may be in a context that does not exactly specify its type. Naming conventions in the sentence and in the corpus are examined in order to determine the type.
Annotation: There are many naming conventions on how to keep gene and protein names distinguishable. If a systematic naming convention is seen in the sentence, the ambiguous basic named entities are typed according to that convention. For example, in (i) *RAD* clearly refers to a protein while *rad* refers to a gene.

If no systematic naming convention is seen and there is no other evidence that supports a certain type, the usage of the name (and the related names) in the corpus is examined. If there is a clear consensus for either gene or protein, that type is used. If both types are abundantly present, the type that fits the context best is chosen. For example, MSH is generally used as a gene name in the corpus. Therefore the entities in (ii) are typed as genes.

In all other cases, the type is determined by reference to external knowledge.

4.2.1 i. A core activity associated with the N terminus of the yeast RAD52 protein is revealed by RAD51 overexpression suppression of C-terminal rad52 truncation alleles.
RAD52 - *Gene/protein/RNA*
RAD51 - *Individual_protein*
rad52 - *Gene*
suppression of:DOWNREGULATE(RAD51 overexpression, rad52 alleles)

ii. MSH2 functions in mismatch repair in conjunction with MSH3 or MSH6
MSH2 - *Gene*
MSH3 - *Gene*
MSH6 - *Gene*
functions in conjunction with:RELATE(MSH2, MSH3)
functions in conjunction with:RELATE(MSH2, MSH6)

4.3 Special relationships between nested and nesting entities

Synopsis: Special entity types such as *Substrate*, *Antibody* and *Analog* specify both the type of the nesting entity and its relationship to the nested entity.

Annotation: Primarily, an extended named entity nests an entity that is closely related to it. In case of physical entities, the nesting entity is equivalent to (i), derivable from (ii), or substructure or superstructure of (iii) the nested entity. In case of processes and properties, the nested entity is a participant in the process (iv) or the possessor of the property (v), respectively.

In some cases, the relationship between a nesting entity that corresponds to a bioentity and its nested entity is more indirect. This relationship is usually an agent–patient relationship (vi–vii) or a similarity of properties (ix). These nesting entities do not have to be of genes, RNAs or proteins (see Section 2.1.1). For example in (ix), the actin analog may be an organic molecule with properties similar to actin.

Specific entity types have been defined in the entity type ontology to capture these relationships and to take into account the fact that the presence of the nesting entity does not imply the presence of the nested entity. The words indicating these types are always included in the extended named entities (see Section 2.4.2).

4.3.1 i. actin protein
actin - *Individual_protein*
actin protein - *Individual_protein*

ii. product of the gene armadillo
gene armadillo - *Gene*
product of gene armadillo - *Individual_protein*

iii. region of beta-catenin
beta-catenin - *Individual_protein*
region of beta-catenin - *Substructure_of_protein*

iv. phosphorylation of talin
talin - *Individual_protein*
phosphorylation of talin - *PHOSPHORYLATION*

v. rate of actin polymerization
actin polymerization - *POLYMERIZATION*
rate of actin polymerization - *Dynamics_property*

vi. p120cas gene encodes a protein tyrosine kinase substrate
p120cas gene - *Gene*
protein tyrosine kinase - *Individual_protein*
protein tyrosine kinase substrate - *Substrate*
encodes:ENCODE(p120cas gene, protein tyrosine kinase substrate)
substrate:AFFECT(protein tyrosine kinase, protein tyrosine kinase substrate)

vii. alpha-spectrin antibody
alpha-spectrin - *Individual_protein*
alpha-spectrin antibody - *Antibody*

viii. Antibody TD77 binds talin.
TD77 - *Individual_protein*
antibody TD77 - *Individual_protein*
talin - *Individual_protein*
binds:BIND(antibody TD77, talin)

ix. actin analogs with a high affinity for profilin
actin - *Individual_protein*
actin analogs - *Analog*
profilin - *Individual_protein*
affinity for:BIND(actin analogs, profilin)

Note the contrast between (vii) and (viii): *alpha-spectrin antibody* is related to *alpha-spectrin* through a binding relationship whereas *antibody TD77* is *TD77* (*antibody TD77* is not an antibody against *TD77*). Note also that the types *Substrate* (vi), *Antibody* (vii), and *Analog* (ix) do not imply the presence of the nested entity.

4.4 Process vs. property

Synopsis: Processes and properties have separate subtrees in the entity type ontology.
Annotation: Phrases that indicate an action, or a process, by or on an entity are annotated using types from the process subtree (i). Phrases that describe a property of an entity but do not state an occurring process in the context of the sentence are annotated using types within the property subtree (ii).

4.4.1 i. phosphorylation by LIMK
LIMK - *Individual_protein*
phosphorylation by LIMK - *Process*

ii. phosphorylability by LIMK
LIMK - *Individual_protein*
phosphorylability by LIMK - *Property*

References

[1] Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. Generating typed dependency parses from phrase structure parses. In N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk, and D. Tapias, editors, *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), Genoa, Italy*, pages 449–454, 2006.

[2] Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Björne, Jorma Boberg, Jouni Järvinen, and Tapio Salakoski. BioInfer: A corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8(50), 2007.

[3] Daniel D. Sleator and Davy Temperley. Parsing english with a link grammar. Technical Report CMU-CS-91-196, Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA, October 1991.

Turku
Centre *for*
Computer
Science

Lemminkäisenkatu 14 A, 20520 Turku, Finland | www.tucs.fi

University of Turku

- Department of Information Technology
- Department of Mathematics

Åbo Akademi University

- Department of Computer Science
- Institute for Advanced Management Systems Research

Turku School of Economics and Business Administration

- Institute of Information Systems Sciences

ISBN 978-952-12-1850-7
ISSN 1239-1891