NSF ITR-0205470

TECHNICAL REPORT

Guidelines for Protein Name Tagging Version 1.0

May 2003

Protein Information Resource
Georgetown University Medical Center
and
Department of Linguistics
Georgetown University

Contact: Zhangzhi Hu zh9@georgetown.edu

1 Introduction

A human reading a biology paper is able to understand it using her knowledge of language as well as her knowledge of biology. To get a computer to do the same, it is helpful to prepare examples of text marked up with whatever information the human needed to extract from it. The resulting corpus of annotated examples can then be used to teach the computer to extract the same kind of information.

The goal of this document is to specify how to annotate one specific kind of information in biomedical texts, namely, references to *protein name objects*. The resulting annotated texts can then be used to automatically train a computer program to tag such references automatically. It is expected that the guidelines developed here will be used by a variety of groups interested in automatically identifying protein names in the biomedical literature, for example, to link texts and protein database entries, or to support further information extraction, e.g., about protein-protein interactions. By using common annotation guidelines, it becomes possible for groups to share annotated data, compare automatic annotation results, and in general advance the field of biological information extraction.

It is expected that these guidelines will be implemented using an annotation tool such as the Alembic Workbench[1]. A version of the Workbench augmented to tag protein names has been used here at Georgetown University.

Constraints on availability of full-text documents have resulted in many groups focusing just on tagging abstracts. It is worth noting that these guidelines are intended to be applicable to both abstracts as well as full-text documents.

1.1 Varieties of Protein Names

Protein names are characterized by their great variety. Like people, proteins can have the equivalent of nicknames as well as official or formal or full names. The same protein can be called by different names. It is extremely common to have a plethora of variations of spelling, capitalization, punctuation, spacing, etc., especially for the nicknames. Protein names usually fall into the following three types:

1.1.1 Single-word names

Simple protein names are single words with only lower-case letters (except when they begin a sentence): “trypsin”, “myosin”, “tropomyosin”, “insulin”, “hemoglobin”, “collagen”. Even these, however, are more properly understood to refer to a fairly specific class or type of protein that may be further differentiated by additional modifiers or specifiers.

[1] www.mitre.org/technology/alembic-workbench/

1.1.2 Symbolic names

Single-word names that mix uppercase and lowercase letters, numerical figures, and non-alphabetical characters. Commonly they are well-established or ad hoc abbreviations or acronyms (the equivalent of nicknames), gene symbols, or arbitrary designations.

1.1.3 Complex names

Complex names are either single words that include Roman and Arabic numbers, Greek letters (or their spelled out names), and non-alphabetical characters (e.g., hyphen, slash, parentheses) or multiple word phrases of mixed characters. These names can include simple protein names, nicknames, common English words (even including "and" and "of"), and words that describe some general or specific property or activity of the protein.

1.2 Annotation Format

References to protein objects are annotated by inserting a special SGML (Standard Generalized Markup Language) tag around the text string. At the start of the expression, <protein> is inserted directly into the text, and at the end of the expression, </protein> is inserted (the same tag, but with a backslash). For example:

<protein>myosin</protein>

As we shall see, there are also two other types of tags: *<compound-protein>… </compound-protein>*, and *<acronym>* … *</acronym>*.,

2 Basic Guidelines

The guidelines for protein annotation are motivated by a number of basic kinds of references that annotators need to be aware of.

2.1 Specific vs. general protein names

Protein name objects could range from specific terms to general ones. General names may include “ribosomal proteins”, “trans-membrane proteins”, “nuclear proteins”, “protein kinase”, “nuclear hormone receptors” etc. Some one-word terms represent even more general protein objects such as “protein”, “peptides”, “enzyme”, “receptor”, “kinase”, etc. As long as they are used to describe protein objects, but not other non-protein objects (e.g. “the receptor gene promoter”) in the paper, these general terms should be tagged regardless of singular or plural forms.

2.2 Stand-alone names vs. context-dependent names

Protein objects are usually referred to with full names when they first appear in the paper, but subsequent references often use “short forms”, which have to be understood in the context of the earlier reference(s). In general, these references should all be tagged as long as they are referring to protein objects. For example, the full name “ubiquinol-cytochrome C reductase complex core protein I” and context-specific terms such as “core I precursor protein” and “core I protein” (PMID-1712295) should all be tagged. Similarly, context-specific terms such as “this protein”, “the enzyme” or “this subunit”, or symbolic names like “E2”, “E1 alpha” or “E1 beta” (subunits of “pyruvate dehydrogenase complex”) (PMID-2007123) should be tagged only if they refer to protein objects (but here, “this, the” should not be included).

2.3 Ambiguity between genes, proteins and mutant genotypes

Often, the distinction between genes, proteins and mutant genotypes is not always clear when acronyms or symbolic names are used. They should be tagged if and only if it is clear that they are referring to protein objects. When in doubt, do not tag it. Following are several such cases:

2.3.1 Symbolic names with case differences

“HypA” and “HypB” refer to proteins, but “hypA” and “hypB” refer to genes. But “hypA protein” is equivalent to “HypA” or “protein HypA”, and thus should be tagged. (PMID-8326860). Also, “multidrug exporter QacC” refers to a protein, “qacC” to gene (PMID-8494372); also “trpD polypeptide” or “trpC protein” vs. “trpD and trpC genes” (PMID-6355484). But in another case, “UBP1” refers to gene and Ubp1 refers to protein (PMID-1429680). In short, “hypA” by itself will not be tagged, whereas “protein HypA” or just “HypA” will be tagged, as will “hypA protein”.

2.3.2 Symbolic names with suffix

Papers describing the yeast gene/protein often contain names such as “SUN2”, “Sun2p” and “sun2”. “SUN2” refers to the protein coding sequence, “Sun2p” refers to SUN2 encoded protein product, while “sun2” refers to the genotype of the yeast strain resulted from genetic manipulation of the gene “SUN2” (deletion or mutation). Only “Sun2p” should be tagged, as it refers to protein objects (PMID-8668124). Similarly, given “PHO81” vs. “Pho81p” (PMID-8709965), only “Pho81p” should be tagged.

2.3.3 Symbolic names ambiguous between genes and proteins

Yeast hypothetical proteins are often named by open reading frames (ORF) nomenclature, e.g. yeast ORFs “YDR332w”, “YDR036c”, “YDR027c”... they may refer to proteins or genes. (PMID-10077188, PMID-11447599). Often times these are the only identifiable names in the abstract. They may NOT be tagged unless there is a clear distinction in the paper.

2.3.4 Yeast genotypic strains named after genes

Such strains are named after genes but without explicitly referring to the GENE or GENE products. Since they don’t refer to proteins, they are NOT to be tagged. For example, yeast strains “ptr1”, “ptr2”... (PMID-1782673) or “rad1”, “rad2” ... “rad16” (PMID-80746), may NOT be tagged as protein objects.

2.4 Cases referring to more than one protein object

When more than one protein is mentioned, separated by conjunctions (‘and’, ‘or’) or by comma/ hyphen, they can be tagged as a whole. For example, “G- or F'-actin”(PMID-1847147), “hypA, B, F, C, D and E” (PMID-8326860). Similarly, “RAC-PK alpha and beta” are two protein subunits, and “RAC-PK alpha, beta, and gamma” are three subunits (PMID-7488143); “MoaA-MoaE” (A through E); “moaA, B, C and E products” (PMID-8361352).
A ***compound protein*** tag will be used for these cases, e.g., *<compound-protein>moaA, B, C and E products</compound-protein>. Note that no attempt should be made to separately tag the parts of the compound protein name*.
Note: do not confuse ‘complex names’ in Section 1.1.3 with ‘compound protein names’ or ‘protein complex’ in Section 2.7 – they are all very different things.

2.5 Modifiers

2.5.1 Grammatical structure of terms

From a linguistic standpoint, a term potentially referring to a protein object can be considered as being a noun phrase (NP) (in other words, ‘a phrase headed by a noun’), made up of particular grammatical components.
In particular, from this standpoint, we can think of the NP as itself being made up of a NP followed by an optional prepositional phrase (PP) or acronym:

- The component NP consists of optional *pre-modifiers* (made up of determiners, adjectives, nouns, proper name symbols – we will collectively refer to this as a single pre-modifier), and a non-optional *head* (nouns or proper name symbols, possibly more than one word long).
- The PP is made up of a preposition (e.g., “for”, “in”, etc., and also “of”) followed by a NP.

- Acronyms are dealt with separately in Section 2.8.

Here are some examples:

1. In the NP “26-KD myrislated protein”, “26-KD myrislated” is the pre-modifier and “protein” is the head.
2. In the NP “pyruvate ferredoxin oxidoreductase”, “pyruvate ferredoxin” is the pre-modifier, and “oxidoreductase” is the head.
3. In the NP “the soluble form of the growth hormone receptor”, “the soluble form” is a NP, with pre-modifier “the soluble” and head “form”, and “of the growth hormone receptor” is a PP with preposition “of” and NP “the growth hormone receptor” which itself has “the growth hormone” as pre-modifier and “receptor” as head.
4. The NP “halorhodopsin for the shark”, has “halorhodopsin” as NP and head with no pre-modifiers, and “for the shark” as PP with “for” as the preposition, and “the shark” as NP with pre-modifier “the” and head “shark”.

This recursive analysis (NP within NPs) is rather a complex structure to have to think in terms of when carrying out annotations. Instead, to keep things simple, given a top-level NP, we will consider *all the words preceding the first head* as the pre-modifier, and *all the words following the first head* as a **post-modifier**. Thus, in this simpler scheme – the one we will use for the guidelines – we have:

1. (as before) In the NP “26-KD myrislated protein”, “26-KD myrislated” is the pre-modifier and “protein” is the head, and there is no post-modifier.
2. (as before) In the NP “pyruvate ferredoxin oxidoreductase”, “pyruvate ferredoxin” is the pre-modifier, and “oxidoreductase” is the head, and there is no post-modifier.
3. In the NP “the soluble form of the growth hormone receptor”, the pre-modifier is “the soluble”, the head is “form”, and “of the growth hormone receptor” is the post-modifier.
4. In the NP “halorhodopsin for the shark”, there is no pre-modifier, “halorhodopsin” is the head, and “for the shark” is the post-modifier.
5. In the NP, “alpha and beta chains of the pyruvate dehydrogenase (lipoamide) component (E1) of the pyruvate dehydrogenase multienzyme complex”, “alpha and beta” is the pre-modifier, “chains” is the head, and “of the pyruvate dehydrogenase (lipoamide) component (E1) of the pyruvate dehydrogenase multienzyme complex” is the post-modifier.

2.5.2 Pre-modifiers with protein object heads

When pre-modifiers are present, it can sometimes be difficult to determine whether or not to include the pre-modifier as part of the protein tag. For example, “dimeric POR”, “26-KD myrislated protein (PMID-1409649)”, “homo-octameric enzymes”, etc. In general, if the pre-modifier is not part of a formal name or is not used as an acronym, the pre-modifier should NOT be tagged. Therefore, only “POR”, “protein” or “enzyme” in the above examples should be tagged. However, in “atrial natriuretic factor (ANF)” and “brain natriuretic peptide (BNP)”, “atrial” or “brain” is part of the formal name, and so they should be tagged as whole. Thus, we have *<protein> atrial natriuretic factor</protein>*.

The pre-modifier rule is also applied to species names. For example, while "halorhodopsin" (abbreviated as HR) refers to a protein object, the reference to species in "halobium HR" and "shark HR" should NOT be tagged. However, if the species name is expressed by means of acronyms or symbolic names, e.g., human growth hormone (hGH) or bovine growth hormone (bGH), then the species name in the acronym should be included in the tag. Thus we have: *halobium<protein>HR</protein>* and *human<protein>growth hormone</protein> <acronym>(hGH)</acronym>*.

2.5.3 Pre-modifiers with non-protein object heads

A protein name can be used to describe non-protein objects such as genes (or cDNA/mRNA), promoters, sequence domains (a region of a protein sequence, e.g. protein kinase domain), or even subcellular structures. For example, "elastase I promoter/ enhancer" (PMID- 3649277), "the insulin gene", or "actin filaments".
These pre-modifiers and heads will NOT be tagged when used to describe these non-protein objects.
However, when they are used to describe certain characteristics or properties associated with the protein, the pre-modifiers should be tagged. For example, "the MAP kinase activity", or "the LH receptor binding affinity", "RAR activation/expression", "TGF receptor cascade/pathway" should all be tagged, e.g., *the <protein>MAP kinase</protein> activity.*

2.5.4 Post-modifiers beginning with "of".

Protein names containing post-modifiers beginning with "of" can be used to describe a subset of protein objects, different protein objects, or characteristics or properties associated with the protein. We discuss these in turn.

2.5.4.1 Ontological Relations

Certain post-modifiers express an ontological relation, as in "soluble form of the growth hormone receptor", which is not a specific name, but indicates a specific form of the receptor protein. Whether or not to tag them as a whole as a protein name depends on the relation expressed by "of" in the context of the pre-modifier and head on one hand, and the post-modifier on the other. There are three types of such relations:

- *Part of* (subunits or chains of a complex). For example, "subunit of NADH dehydrogenase (complex I)", "alpha and beta chains of the pyruvate dehydrogenase (lipoamide) component (E1) of the pyruvate dehydrogenase multienzyme complex", will all be tagged as a whole, e.g., *<protein> subunit of NADH dehydrogenase (complex I)</protein>*. However, if the part refers to a subregion of a protein or a polypeptide such as "c-terminal tail of the hLHR" or "the leucine rich repeat region of the hLHR", we only tag the head "hLHR".
- *Kind of* (type/form of a protein). For example, "17 KD form of TNF" and "long form of the human prolactin receptor" refer to variant forms of "TNF" or "prolactin receptor"; therefore they should be tagged as a whole.
- *Member of* (a family). For example, in "kinase of the ERK family", only the member "kinase" will be tagged. However, in "a member of cytokine receptor superfamily", "member" would not be tagged, since "a member" is a general

description and does not name a specific protein object, nor would "cytokine receptor superfamily" be tagged, since it refers to a family of proteins.

2.5.4.2 Relations between different protein objects

In examples such as "activator of ERK kinase" or "inhibitor of JNK kinase", the "activator" or "inhibitor" is a different object from the kinases, and should be tagged if it refers to a protein object but not any other non-protein objects (lipids, chemicals, etc.). The kinases should be tagged separately, e.g. *<protein>activator</protein> of <protein>ERK kinase</protein>*. However, if the head and modifier together is a formal name, then they should be tagged as a whole, such as "tissue inhibitor of metalloproteinases" (called TIMP rather than metalloproteinase), or "signal transducer and activator of transcription" (called STAT, a transcription factor). We should distinguish the above from the following cases, "the gene of (or for) rat LHR" or "the promoter of human IGF-II", in which the head is non-protein object (genes or promoters), therefore, only the protein objects in the post-modifier part should be tagged ("LHR" or "IGF-II").

2.5.4.3 Protein properties

Certain post-modifiers express properties associated with proteins, as in "transcriptional properties of RARs", "dimerization of ER, or "the expression of CREB protein". Here only the references to protein name objects "RARs", "ER", or "CREB protein" will be tagged.

2.6 Embedded protein names

Multi-word protein names may embed other protein names within, e.g. actin in "actin-binding proteins" or "Ca2(+)-activated actin-binding protein" (PMID-1847147). Similarly, "ABC protein-mediated exporters" (PMID-8918458) or "MAP kinase kinase kinase". *The outer protein name, not the embedded one, should be tagged as a whole.* Thus, one would have *<protein>actin-binding proteins</protein>*, or *<protein>MAP kinase kinase kinase</protein>.*

2.7 Protein object refers to protein complex

Protein objects may refer to a protein complex (sometimes called 'system') that contains more than one subunit or chain. For example, "periplasmic binding-protein-dependent transport system" (PMID-7997183) or "pyruvate dehydrogenase complex" (PMID-2200674), they should be tagged as a whole, e.g., *<protein> periplasmic binding-protein-dependent transport (PBT) system</protein>.*

2.8 Full name and acronym (or abbreviation)

2.8.1 Acronyms in parentheses

When a full name is followed by an acronym in parentheses, or an acronym is followed by a full name (in parentheses or separated by comma), they will be tagged using two tags – a protein tag for the full name and an ***acronym*** tag for the acronym. For example, the following will all fall under this case: "CCAAT/enhancer binding protein (C/EBP)", "cAMP responsive element binding protein (CREB)"; or "HAT, histone acetyltransferase", or "GR-LACS, also known as gonadotropin-regulated long chain acyl-

CoA synthetase", etc. The first example above would be tagged as *<protein>CCAAT/enhancer binding protein</protein> <acronym>(C/EBP)</acronym>.*

2.8.2 Acronyms not in parentheses

Acronyms may also appear without parentheses, e.g., "the neural cell adhesion molecule NCAM". Here too, a protein tag and an acronym tag are both used. This example would therefore be tagged as *the <protein> neural cell adhesion molecule</protein> <acronym>NCAM</acronym>.*

2.8.3 Partial acronyms

Acronyms may designate only a part of the full name, e.g. "2-ketoacid oxidoreductase (ORs)", "pyruvate ferredoxin oxidoreductase (POR)". The full name and the acronym will both be tagged (the full name as a protein, the acronym as an acronym) regardless of how they correspond to each other.

2.8.4 Non-acronym parentheticals

When a full name is followed in parentheses by a symbolic name that is not an acronym, then tag the full name and symbolic name separately. For example, "sucrose phosphorylase" will be tagged as *<protein>sucrose phosphorylase</protein> <protein>(gtfA)</protein>* (PMID- 1537846).

2.8.5 Embedded acronyms

Notice that acronyms can be embedded in the middle of another protein name, rather than occurring at its beginning or end. When an acronym is embedded in the middle, it is treated as part of the protein name. This rule, which follows from the treatment of embedded protein names (see Section 2.6), overrides any other rules in this section. For example, in "cyclophilin (CyP)-type peptidyl prolyl cis-trans isomerase (PPIase)", "CyP" is an embedded acronym. Thus, we have *<protein>cyclophilin (CyP)-type peptidyl prolyl cis-trans isomerase</protein> <acronym>(PPIase)</acronym>.* Also in "G-protein coupled receptor (GPCR) kinase (GRK)", "GPCR" is embedded acronym, it should be tagged as *<protein>G-protein coupled receptor (GPCR) kinase</protein> <acronym>(GRK)</acronym>.*

2.8.6 Acronyms without the full name

An acronym can occur by itself, without the full form occurring, e.g., "hGH". In such cases, the acronym should be tagged using the ***protein*** tag, and NOT the acronym tag.

2.8.7 Abbreviations versus acronyms

No distinction is made between abbreviations and acronyms – they are both given the ***acronym*** tag.

2.9 Long protein names containing descriptive words

"Low-affinity cationic amino acid transporter-2" is a specific protein name (PMID-8385111). However, "low affinity, high capacity transporter of cationic amino acids" describes such a transporter, thus only “transporter” should be tagged based on the context.

3 Meta-Guidelines

This section describes some ‘background’ guidelines that are derived from the above guidelines.

3.1 What Should be Tagged

Only clear references to protein objects should be tagged, including protein complexes and sets of protein objects. References to genes, gene promoters, mutant genotypes, etc., should NOT be tagged.

3.2 Uncertainty

There will be many cases where it may not be clear what to tag. For example, it may not be clear whether a reference is to a protein object or a gene. *When in doubt, do not tag it!* That way, the computer has fewer examples to train from, but they will be of higher quality.

3.3 Case

References to protein objects should be tagged, irrespective whether they are in lowercase, uppercase, or mixed case. Note, however, that case may sometimes, but not always, distinguish proteins from genes, e.g., “HypA” and “HypB” refer to proteins, but “hypA” and “hypB” refer to genes. Having said that, it is important to be guided by what the term refers to, rather than case.

3.4 Embedded Tags

A **protein** tag or a **compound-protein** tag will internally contain no other tags.
For example, <protein>Ca2(+)-activated actin-binding protein</protein> and <compound-protein>moaA, B, C and E products</compound-protein>.

APPENDIX Sample Tagged Documents