Guidelines for Protein Name Tagging
Version 2.0

April 2004

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

NSF ITR-0205470

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.

1.0 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:

1.0.1 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.

1.0.2 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] www.mitre.org/technology/alembic-workbench/

1.0.3 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.

1.1 Annotation Format

References to protein named entities 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>

We shall use two types of tags: *<protein>… </protein>*, and *<long-form> … </long-form>*, the latter of which will allow the optional expansion of the protein named entity when protein name boundaries are not clear.

2 Basic Guidelines

2.0 Protein named entity

The protein named entity refers to protein names (full names, acronyms or other symbolic names) that may be used in the literature to describe proteins, or protein associated or related objects, such as domains, pathways, expression, gene, etc. We will tag the protein named entity portion of the text, such as <protein>NF-kappaB</protein> gene expression.

2.1 General protein names

Protein names could range from very specific to very general terms. General names may include "ribosomal proteins", "nuclear proteins", "protein kinase". Some one-word terms represent even more general protein names such as "enzyme", "receptor", "kinase", etc.

- Generic **one-word protein names** may or may not be tagged, depending on whether they are specified in the two dictionaries (see appendix).
  - Not to be tagged: e.g. protein, subunit, chain, enzyme, homolog, inhibitor, suppressor, repressor, activator, complex, antibody, molecule, Ig,
  - To be tagged: e.g. **receptor**, **kinase, protease**.
- General **multi-word protein names** are usually to be tagged: e.g., ribosomal protein, transmembrane protein, nuclear proteins, 26-KD myristylated protein. But do not tag names such as "conserved proteins", "hypothetical proteins", etc., which are too general without sensible meanings.
- Several proteins or enzymes are used as reagents of methodology (assay), not as the subject of research from the context, they will not be tagged, such as: luciferase (Luc, for promoter assay); green fluorescence protein (GFP, as fusion protein tag); beta-galactosidase (beta-Gal, promoter assay); Glutathione S-transferase (GST, as fusion protein tag); Taq polymerase; (reverse transcriptase) RT-PCR, RNase assay. However, if they were the subject of the research paper, they would be tagged as protein named entities.

2.2 Context-specific names

Protein named entities usually appear as full names when first mentioned in the paper, but subsequent often appear as "short forms", which may have to be understood within the context. For example, "core I precursor protein" or "core I protein" refer to "ubiquinol-cytochrome C reductase complex core protein I", "E2", "E1 alpha" or "E1 beta" refer to the subunits of "pyruvate dehydrogenase complex", which should not be tagged if appeared alone.

- Stand-alone context-specific names (without modifiers) should not to be tagged: e.g., beta subunit, E2, "E1 alpha" or "E1 beta".

2.3 Protein names with modifiers

2.3.1 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.

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.

A DICTIONARY of protein names from PIR-NREF database composite names will be used during the tagging as a reference.

2.3.2 Modifiers to be tagged as part of <Protein> tag

- Modifiers that are part of protein names should be tagged as <protein>
    - <protein>heat shock protein</protein>
    - <protein>atrial natriuretic factor</protein>
- Tag “protein”, “subunit”, “chain”, “precursor”, “isoform”, “subtype”, “trimer”, “dimer”, “complex”, “soluble form”, “long form”, when they are used together with protein names.
    - <protein>CREB protein </protein>; <protein> hDcp2 protein</protein>.
    - <protein>RAC-PK alpha subunit</protein>; <protein>alpha subunit of RAC-PK</protein>
    - <protein>PZY receptor subtypes</protein>
    - <protein>soluble form of the growth hormone receptor</protein>
    - <protein>Laminin-8 trimer </protein>
- Both modifiers and head nouns are protein named entities, tag them as whole <protein>
    - **Protein-named modifier**: <protein>transcription factor GATA-4</protein>; <protein>T1R3 G-protein-coupled receptor></protein>; <protein>CD28 surface receptor </protein>
    - But if the pre-modifier is one word falling in the list of “not to be tagged” (2.1), then they are not included in the protein tag, e.g., the protein <protein>GATA-4</protein>, the enzyme <protein>N-acetyl-D-glucosamine (GlcNAc) 2-epimerase</protein>.
- Several modifiers are also general terms that do not always refer to protein named entities, so don’t tag them unless it is a protein entity as indicated in the text, such as inhibitor, suppressor, repressor, activator, exporter, carrier.
    - Protein entity: <protein>ERK kinase activator</protein>, <protein>Diazepam binding inhibitor</protein>, etc.
    - Non-protein entity: inhibitor of <protein>JNK kinase</protein>
- **synonyms or homologs separated by “/” “**()”, generally tag them separately unless it is apparent from the text that they are abbreviations or acronyms, which will be tagged as one <protein>.
    - <protein>MGF</protein>/<protein>STAT5</protein>
    - <protein>IL-17BR<protein>/</protein>Evi27 </protein>

However, if they appear in the DICTIOANARY as one name, they will be tagged as one: e.g. <protein>Aly/REF</protein>, <protein>peripherin/rds</protein>. If the two synonyms are modifiers as in “CD94/NKG2 receptor”, tag them as one entity, e.g. <protein>CD94/NKG2 receptor</protein>.

2.3.3 Modifiers/head nouns to be tagged optionally as part of <Long-form> tag

The long-form tags only occur outside <protein>, and should encompass at least one protein tag, e.g. <long-form>…<protein>…</protein>…</long-form>.

- **sources:** For organisms, tag them as part of <long-form> unless they are specified as part of the protein name (as indicated by acronym) or specified in the protein name DICTIONARY (then tag them as part of <protein>):
    - <long-form>human <protein>IGF-II</protein></long-form>
    - <protein>human growth hormone (hGH)</protein>
    - If species is tagged as part of name in the abstract, it will be tagged the same way throughout the same abstract when species appears with the name.

  Words like “mammalian”, “rodent”, etc. are considered as species names too. However, for tissue and cell, tag them as part of protein just like pre-modifiers.
    - <protein>cell membrane estrogen receptor</protein>
    - <protein>mitochondrial aconitase</protein>

- **list of names or subunits sharing common core terms:** The protein tag inside the list should include the core term and the immediate adjacent part of the list.
    - <long-form>G- or <protein>F-actin</protein></long-form>;
    - <long-form>< protein>hypA</protein>, B, C, and D</long-form>;
    - <long-form><protein>MoaA</ protein>-<protein>MoaE</protein></long-form>;
    - <long-form>alpha(1)A/< protein>alpha(1)B-AR</protein></long-form>
    - <long-form><protein>heterogeneous nuclear ribonucleaoprotein (hnRNP) A1</protein>/A2</long-form>;
    - <long-form><protein>CSN subunits 4</protein>, 5, 6</long-form>;
    - <long-form><protein>RAC-PK alpha</protein> and beta subunits<long-form>;
    - <long-form><protein>RAC-PK alpha </protein> and beta</long-form>;
    - <long-form>alpha and <protein> beta chains of the pyruvate dehydrogenase (lipoamide) component (E1)</protein></long-form>;
    - <long-form>multiprotein complex of <protein >Mre11</protein> and <protein>NBS1</protein ><long-form>

- **multiple long-forms,** only tag the outermost: <long-form>heteromer of the taste-specific <protein>T1R1</protein> and <protein>T1R3 G-protein-coupled receptor</protein></long-form>

2.3.4 Modifiers/head nouns not to be tagged

- **Protein properties**: terms associated with various protein properties, expression, transcription, activation, activity, binding affinity, etc. and other type of entities such as genes.
    - <protein>IGF-II</protein> binding activity; <protein>RAR</protein> transcription;
- **family**: <protein>elastase I</protein>family

- **homolog**: <protein>elastase I</protein>homolog; homolog of the <protein>human growth hormone</protein>
- **gene-related entities**: gene, promoter, enhancer, coding region, ORF, mRNA, cDNA, locus, 5’ region/3’ region, operon, phenotype, etc.
    - <protein>elastase I</protein>gene expression
    - enhancer of <protein>elastase I</protein>
    - <protein>ptr2</protein> strains;
- **domain or site**: peri-<protein>kappa B</protein> site; <protein>CD4</protein> epitopes
- **pathway or cascade**: <protein>MAP kinase</protein> pathway
- **fusion protein**: <protein>IL-18Ralpha</protein>-Fc protein

2.3.5 Nested protein names

- If the nested protein names refer to only **one protein entity**, tag only the outermost name
    - <protein>Ca2(+)-activated actin-binding proteins</protein>
    - <protein>MAP kinase kinase</protein>
- If **more than one protein objects**, tag all
    - <protein>kinase</protein> of the <protein>ERK</protein> family
    - <long-form><protein>NF-kappa B</protein>/<protein>CD28</protein>-responsive_complex</long-form>

2.4 Protein names with abbreviation or acronyms

- when abbreviation or acronym **co-occur** with full protein name, place <protein> tag outside delimiters encompassing the acronym (e.g. parenthesis should be included).
    - <protein>Activin receptor-like kinase (ALK)-1 </protein> and <protein>ALK-2 </protein>
    - <protein>2-ketoacid oxidoreductase (ORs) </protein>
    - <protein>cyclophilin (CyP)-type peptidyl prolyl cis-trans isomerase (PPIase) </protein>
    - <protein>G-protein coupled receptor (GPCR) kinase(GRK)</protein>
    - If the source name is used as part of formal name (appearing in the acronym) in the abstract, they will be tagged as part of protein tag in all occurrences. <protein>human growth hormone (hGH)</protein>;

But sometimes, the acronym appears with other things in the parenthesis following a full name, e.g. RB-associated KRAB (RBAK, Hs.7222), then tag the full name and the acronym separately as <protein>RB-associated KRAB </protein> (<protein>RBAK</protein>, Hs.7222).

3. APPENDIX

3.0 Sample Tagged Documents

Examples will be added later.

3.1 Common Word Dictionary

Single words to be tagged:

receptor
neuropeptide
antiporter
transporter
symporter
photoreceptor
exchanger
complement
oncoprotein
antizyme

Also words in mc_e will be tagged (see pir.georgetown.edu/ /~huz/iProLink.html).

Single words not to be tagged:

acceptor
activator
adapter
adaptor
antibody
biglycan
binder
carrier
chain
channel
coactivator
coatomer
coenzyme
complex
component
cotransporter
dipeptide
domain

effector
enhancer
enzyme
facilitator
factor
fragment
glycopeptide
heterodimer
holoenzyme
homolog
homologue
inducer
inhibitor
initiator
integrator
interactor
isoenzyme
isoform
isolog
isotype
isozyme
mediator
modifier
modulator
motif
oligopeptide
ortholog
partner
pentapeptide
peptide
peptidoglycan
polypeptide
polysaccharide
precursor
proactivator
product
proenzyme
propeptide
protein
proteoglycan
proteoglycans
proteolipid
pump
regulator
repressor
responder
sequence