Annotation guidelines for Named Entity Recognition in the FlySLIP project

Andreas Vlachos, Nikiforos Karamanis, Ruth Seal, Ian Lewin, Chihiro Yamada, Caroline Gasperin, Ted Briscoe

April 20, 2006

1 Introduction

The guidelines presented in this document were used to annotate gene-names in 82 abstracts of articles curated by FlyBase curators. The inter-annotator agreement on gene-names was 91%. They are based on the guidelines from the ACE project. The basic idea is that each gene-name (gn) is surrounded by a mention tag, which covers the noun phrase which contains the gene name. Two mention tags were used in this annotation, the gene-mention (gm) and the other-mention (om). When the noun phrase containing a gene name is referring to the gene entity itself, then the gm tag is used. Otherwise, when the noun phrase refers to a biomedical entity other than the gene itself the om tag is used. In cases where two or more gene names are contained in the same noun phrase, one mention tag is assigned to all of them. These points are presented more elaborately in the sections that follow. Additional examples are presented in the Appendix.

2 Tagging gene names

We annotate all the gene names that appear in the abstracts. Non-Drosophila genes, family gene names, reporter genes, transposable elements, gene names used in naming any genetic products are all annotated as well. Modifying or modified entities are excluded from the gn tag, they are usually tagged as part of the mention (see section Tagging mentions). Examples:

(1) the <gn>faf</gn> gene
(2) the <gn>Toll</gn> protein
(3) the Drosophila <gn>Toll</gn> gene family
(4) the <gn>Adh-related</gn> gene
(5) the <gn>string</gn>-<gn>LacZ</gn> reporter genes
(6) the homeotic gene <gn>Sex combs reduced</gn> (<gn>Scr</gn>)
(7) the <gn>N-ethylmaleimide-sensitive fusion protein</gn> (<gn>NSF</gn>)
(8) <gn>male-specific lethal-1</gn>, <gn>-2</gn> and <gn>-3</gn> genes, ...
(9) <gn>suppressor of fork</gn> appears to...
Comments: (5): Note that both gene names are tagged, even though the resulting name refers to a transgenic construct. (4),(7): Note that words such as “protein” and “related” can be part of a gene name. (8): While “-2” and “-3” are not gene names on their own, they are tagged as such because they stand in for “male-specific lethal-2” and “male-specific lethal-3”. (9): In this case, there are two gene names and we tag only the outer one.

3 Tagging mentions

The mention tags (gene-mention (gm) and other-mention (om)) cover the shortest complete noun phrase (NP) that contains the tagged gene-name. Intuitively, the gm tag should cover a text portion that could be replaced by the expression “the gene”. Similarly, the om tag should cover a text that could be replaced by “it” or “they”/“them”. More formal explanations and examples follow. First we discuss the extent of the mentions and then the semantic typing (distinguishing between gene-mentions and other-mentions).

3.1 The extent of the mentions

In this section we discuss the extent of the mentions (m), which can be either gene-mentions (gm) or other-mentions (om). The mentions cover the shortest complete noun phrase (NP) that contains the tagged gene-name and they cannot be overlapping. Linguistically, we tag the baseNP chunk that contains the gene-name. According to Ramshaw & Marcus (Text chunking using Transformation-Based Learning, 1995):

"The goal of a baseNP chunk is to identify essentially the initial portions of non-recursive noun phrases up to the head, including determiners but not including postmodifying prepositional elements or clauses."

It is worth noting that the mentions in the ACE guidelines are different because they are full nominal phrases including any postmodifiers. In our annotation, in agreement with Ramshaw and Marcus, we take anything from the determiner or leftmost modifier up to the head, which usually is the last noun of the phrase. We also include anything up to the rightmost non-clausal modifier of the head. Examples:

(1) <m>the <gn>faf</gn> gene</m>
(2) <m>the gene <gn>faf</gn></m>
(3) <m>the four muscle <gn>actin</gn> genes</m> of Drosophila virilis
(4) ... a novel Drosophila gene, <m><gn>Nk6</gn></m> ... 
(5) <m>the homeotic gene <gn>Sex combs reduced</gn></m> 
   (<<m><gn>Scr</gn></m>)
(6) <m><gn>Hexokinase</gn> coding <gn>DM1</gn> and <gn>DM2</gn> sequences</m> were obtained ...

Comments: (2): The gene-name “faf” is treated as a non-clausal modifier of the head “gene”. (4)-(5): Parentheticals and text adjuncts following the head are tagged as separate mentions. (6): The head of the conjunction is “sequences”, therefore the mention covers the two coordinated gene names (“DM1”, “DM2”) and the gene-name “hexokinase” in a modifier position.
3.2 Semantic typing

The discrimination between \texttt{gm} and \texttt{om} is based on whether the NP being tagged is referring to the gene entity or not, respectively. In many cases, the head noun of the noun phrase can be used to determine the type of the mention. Examples:

(1) \texttt{gm}\textit{the} <\texttt{gn}>faf</\texttt{gn}> gene</\texttt{gm}>
(2) \texttt{om}\textit{the} <\texttt{gn}>Reaper</\texttt{gn}> protein</\texttt{om}>
(3) \texttt{gm}\textit{the four muscle} <\texttt{gn}>actin</\texttt{gn}> genes</\texttt{gm}> of Drosophila virilis
(4) \texttt{gm}\textit{the} <\texttt{gn}>LacZ</\texttt{gn}> reporter gene</\texttt{gm}>
(5) \texttt{om}\textit{the} <\texttt{gn}>string</\texttt{gn}>--<\texttt{gn}>LacZ</\texttt{gn}> reporter genes</\texttt{om}>
(6) \... \textit{a novel Drosophila gene}, <\texttt{gn}>Nk6</\texttt{gn}>/</\texttt{gm}> ... 
(7) \texttt{om}\textit{the} <\texttt{gn}>Ras</\texttt{gn}> signal transduction pathway</\texttt{om}>
(8) \texttt{gm}\textit{a deleted} <\texttt{gn}>hobo element</\texttt{gn}></\texttt{gm}>

Comments: (2): Proteins, alleles, mutants, mRNAs, etc. are tagged as \texttt{om}. (3): Gene families are tagged as \texttt{gm}. (4)-(5) LacZ is a reporter gene and therefore tagged as \texttt{gm}, but string-LacZ is a transgenic construct, therefore tagged as \texttt{om}. (6): Note the extent of the \texttt{gm} tag. (7) Transduction pathways, expressions, activity are all tagged as \texttt{om}. (8): Transposable elements are tagged as \texttt{gm}.

In the examples above, the noun phrase tagged provides enough evidence to determine the type of the mention. In many cases though, a gene name is used without the authors stating whether they refer to the gene or a product of it. In such cases, the context must be used to determine the type of the mention. Some useful clues follow (mainly aimed to non-biologists):

a Genes tend to be passive. Active terms such as binding, cleaving, localizing, interacting physically tend to refer to gene products, rather than genes.

b The terms expression and transcription usually relate to a gene, while the term translation relates to a gene product.

c Genes are measured in kb or bp.

d Proteins are measured in kDa.

e Capitalization can be a useful clue, when in the same document a gene-name appears capitalized and lowercase, then usually the lowercase form refers to the gene and the capitalized one to the protein.

f Proteins have an amino acid sequence, genes have a DNA/nucleotide sequence.

g The terms peptide, domain, carboxy and amino termini relate to proteins, rather than genes.
4 Tokenization & Conversion to IOB

The IOB format is more convenient for training and evaluating machine learning approaches. We tokenized the abstracts using the RASP tokenizer and fixed a few blatant mistakes. In cases where a token was partially annotated as gm, gn or om, we annotate the whole token as such. For example:

... in <om>a <gn>Duf</gn>-dependent manner</om> ...

becomes:

<table>
<thead>
<tr>
<th></th>
<th>in</th>
<th>a</th>
<th>Duf-dependent</th>
<th>manner</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>O</td>
<td>B-GN</td>
<td>I-OM</td>
<td>O</td>
</tr>
</tbody>
</table>

5 Appendix

5.1 Gene names

Common English modifiers such as “related”, “like”, etc. can be part of gene-names, but not always. When such ambiguity appears, the context is usually helpful.

<gm><gn>Adhr</gn></gm> (<gm><gn>Adh-related</gn> gene</gm>)
<gm><gn>the <gn>Toll-like receptor</gn></gm> (<gm><gn>tlr</gn></gm>)

But:

<om>a Drosophila <gn>Iroquois</gn>-related homeobox transcription factor</om>

Plural versions of gene names are considered to be gene families and they are tagged as gn:

<gm><gn>Rhodopsins</gn></gm> ...

5.2 Extent of mentions

Gene names followed by their abbreviations are contained in two mentions

The previously reported gene, <gm><gn>Lethal hybrid rescue</gn></gm> (<gm><gn>Lhr</gn></gm>) ...

When annotating coordinations, one should tag the shortest complete noun phrase:

<om>the <gn>dunce</gn> and <gn>rutabaga</gn> mutants</om>
<gm><gn>dunce</gn></gm> and <gm><gn>rutabaga</gn></gm>

In the case of possessives we tag as gene-name the name without the “’s”:

<om><gn>Fra</gn>’s ectodomain</om>

1http://www.informatics.susx.ac.uk/research/nlp/rasp/
5.3 Semantic typing

Although capitalization of the gene-name is often an indication that the gene-name is used to refer to the protein, this is not always the case:

<gm>The homeotic genes <gn>abdominal A</gn> and <gn>Abdominal B</gn></gm>

Noun phrases headed with the terms “locus” and “region” are annotated as other mentions:

<om>the <gn>string</gn> locus</om>  
<om>the <gn>apterous</gn> region</om>

Noun phrases referring to gene families are tagged as gm:

<gm>the <gn>p53</gn> gene family</gm>

Alleles and mutants are tagged as om:

<om>the <gn>HD</gn> allele</om>  
<om>the <gn>lgl</gn> mutant</om>

Transgenic constructs are annotated as other mentions:

<om> a <gn>glass-responsive</gn> <gn>gfp</gn> fusion gene </om>

In order to determine the mention type, one could look for terms that relate to genes (classifying the mention as gm) or gene products (classifying the mention as om). Examples:

... ectopic expression of <gm><gn>hth</gn></gm> .....  
... transcription of <gm><gn>string</gn></gm> .....  
... a synthetic leucine-rich repeat peptide (LRP32) representing one of the repeats found in <om>Drosophila <gn>chaoptin</gn></om> .....  
... <om><gn>Rols7</gn></om> localizes in a <om><gn>Duf-dependent</gn> manner</om> .....  
... the carboxy terminus of <om><gn>p85</gn></om> .....  
... <om>The recombinant <gn>p85</gn> interacts directly with both <om>the <gn>TATA box-binding subunit</gn> (<om><gn>TFIID tau</gn></om> or <om><gn>TBP</gn></om>)</om> .....  

Sometimes authors refer to a mutant using exactly the same name as the gene. The context of the sentence must be used to spot these cases.

... Two mutants that fail to exit cellular quiescence at larval hatching (<om><gn>milou</gn></om> and <om><gn>EIF4(1006)</gn></om>) ...