

Human Gene/Protein Mention Normalization Annotation Guidelines

Alex Morgan

May 2006

Overview

The task is to identify the mentions of distinct, named, human genes and gene products (transcripts, proteins) and to link each to a unique EntrezGene (EG) identifier when possible.

Guiding Principles

- 1) Mentions of distinct genes and their specific products that map to Entrez Gene identifiers will be annotated.
- 2) Genes and gene product references that constitute mentions must be explicitly named or enumerated.
- 3) Human genes are the only ones of current interest for the annotation task.
- 4) All specific, named, non-anaphoric, mentions will be collected as the minimum direct text excerpts that constitute the name.

Specific Examples

- * The mention of genetic diseases that may be caused by variations in a particular polypeptide mentioned in the context of the disease and not as a name for the protein do not count as mentions.
- * Polypeptide names that actually refer to other proteins do not constitute a mention of those other proteins. A mention of "insulin receptor" is not a sufficient mention of "insulin".
- * Mention of a polypeptide or RNA product is sufficient mention as long as it clearly maps to a single gene.
- * The mention must be for a human gene/polypeptide. If no source organism is explicitly mentioned, and assuming it is a human is a reasonable default, then that constitutes a mention.
- * Mentions of protein families or gene groups do not constitute sufficient mention to annotate for individual members of the group or family. The only exception to this is if the group members are clearly enumerated. Taking an example from yeast, "G2-specific (CLB1-CLB4) cyclins" is considered sufficient mention to generate annotations for CLB1, CLB2, CLB3, CLB4.
- * Mentions of alternate transcripts or post translational variants that map to a single gene (EG ID) all constitute different mentions of the same gene as do allelic variants.
- * If a gene name is ambiguous, but the text clearly refers to a particular gene either through an examination of the full text or through inference from document level annotations, then the mention should be annotated with the correct identifier.

* Mentions that do not use a specific name such as "the gene" or "the protein" and pronouns should not be annotated.

* References to domains named after a particular gene/protein do not constitute mentions.

* The excerpts that are extracted and linked to mentions should capture the minimal text span. The full excerpt "PKC isoforms alpha, delta, epsilon and zeta" is required to link with PKC zeta whereas "PKC isoforms alpha, delta" is sufficient for PKC delta.

- Mentions of fusion proteins (eg Bcr-Abl) do not count as a mention of the constituent genes. Abstracts that include mentions to fusion proteins of this sort should be marked as problems.

- If there is a clear mention of a human gene/protein, but it cannot be identified in EntrezGene, then that abstract should be marked for removal.

- A locus which can be directly associated with an EntrezGene identifier is a mention of that EG id.

- Only 'Current' EG identifiers should be used, not expired ones.

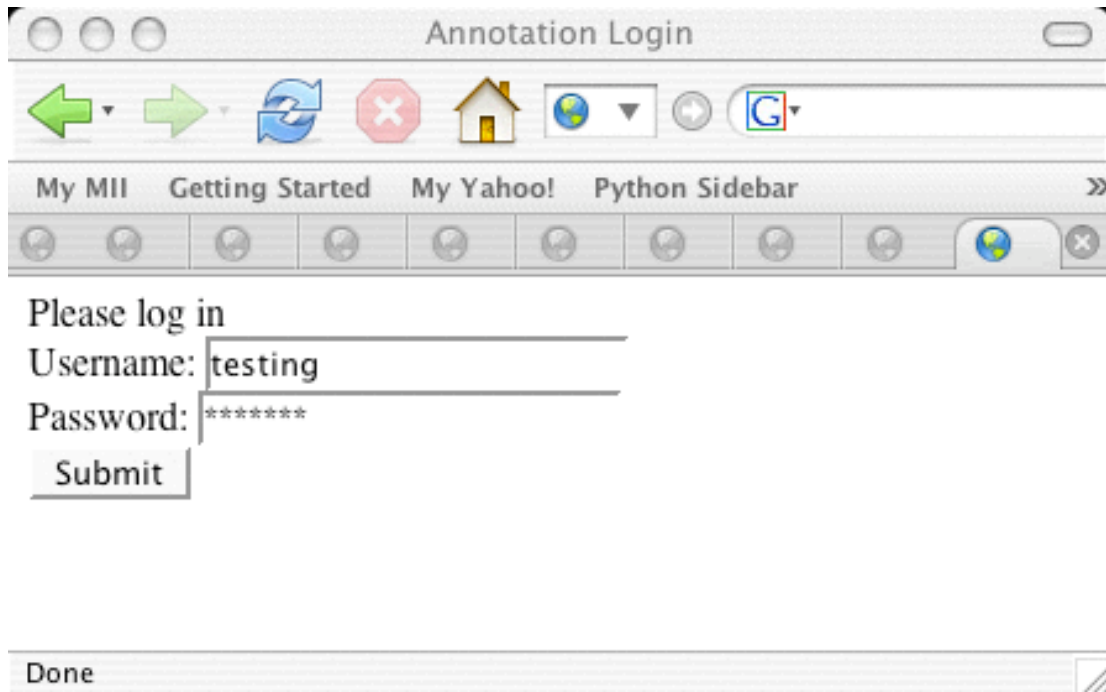
Logging Into the Annotation System

The process of annotation starts by connecting to the server.

Default address:

`http://smiley.mitre.org/~amorgan/annotation`

To begin annotation use the username and password provided to you. If you want to examine the system, please use the guest account (username: `testing`, password: `testing`).



The screenshot shows a web browser window titled "Annotation Login". The address bar contains the URL `http://smiley.mitre.org/~amorgan/annotation`. The browser's toolbar includes navigation buttons (back, forward, stop, home, search) and a search engine dropdown set to Google. Below the toolbar is a tab bar with tabs labeled "My MII", "Getting Started", "My Yahoo!", and "Python Sidebar". The main content area displays a login form with the text "Please log in". The form has two input fields: "Username:" with the value "testing" and "Password:" with the value "*****". A "Submit" button is located below the password field. At the bottom of the browser window, a status bar shows the word "Done".

Getting an Abstract

Once you've successfully logged into the system, you should be quickly directed to a list of PubMed identifiers. This is part of the queue of abstracts you have been assigned. As you annotate abstracts, this list will be continually updated. If there are no abstracts listed, you are not currently assigned to annotate.

Your username should be listed at the top. If this isn't correct, you will need to logout and login once again.

At this point you can click on any number on the list, and you can begin annotating.



List of Abstracts to Annotate

Annotator: alex

[6688123](#)
[11102480](#)
[2172835](#)
[7624774](#)
[2558868](#)
[8663294](#)
[8579597](#)
[8579598](#)
[8890164](#)
[2674117](#)

Once you have selected an abstract, you will be presented with the annotation page for that abstract. At the top of the page is your username again, a link to the PubMed entry for the abstract, and then the title and text of the abstract itself. If there are any existing EBI curations for this abstract they will follow along with a list of suggested synonyms. Your annotations will go in the table at the bottom.

[illegible]

The annotation task is to identify mentions in the text of human genes and their products and provide the corresponding unique identifiers from EntrezGene.

<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene>

First read the title and text carefully to find any mentions of human genes or their products. The EBI annotations may be used as a guide for assistance.

In the far left column is a space for a UniProt identifier, but this is optional, the EntrezGene identifier follows with finally any comments about that entry.

Human Gene/Protein Normalization Annotation to EntrezGene Identifiers

http://smiley.mitre.org/~amorgan/pop/genedict

My Mail Getting Started My Yahoo! Python Sidebar NCBI HomePage DAVID SF: BioCreAtivE G063 Wiki: Patterns ...

In... Hum... Index... Hum... Hum... Hum... Hum... Hum... Hum... Index...

Annotator: alex

8663294 : ICE-LAP6, a novel member of the ICE/Ced-3 gene family, is activated by the cytotoxic T cell protease granzyme B.

Members of the ICE/Ced-3 gene family are likely effector components of the cell death machinery. Here, we characterize a novel member of this family designated ICE-LAP6. By phylogenetic analysis, ICE-LAP6 is classified into the Ced-3 subfamily which includes Ced-3, Yama/CPP32/apopain, Mch2, and ICE-LAP3/Mch3/CMH-1. Interestingly, ICE-LAP6 contains an active site QACGG pentapeptide, rather than the QACRG pentapeptide shared by other family members. Overexpression of ICE-LAP6 induces apoptosis in MCF7 breast carcinoma cells. More importantly, ICE-LAP6 is proteolytically processed into an active cysteine protease by granzyme B, an important component of cytotoxic T cell-mediated apoptosis. Once activated, ICE-LAP6 is able to cleave the death substrate poly(ADP-ribose) polymerase into signature apoptotic fragments.

UniProt	Entrez	Names
P55211	842	APAF-3; ICE-LAP6; Q53Y70; CASP-9; CASPASE-9c; ICE-like apoptotic protease 6; CASP9; Caspase 9, apoptosis-related cysteine protease; Apoptotic protease Mch-6; Caspase-9 precursor; CASP9_HUMAN; caspase 9, apoptosis-related cysteine peptidase; Caspase-9; caspase 9, apoptosis-related cysteine protease; APAF3; Apoptotic protease activating factor 3; MCH6; Q53Y70_HUMAN

UniProt	Entrez	Excerpt	Comment
P55211	842	ICE-LAP6	
P42574		Yama CPP32 apopain	
P55212	839	Mch2	
P55210	840	ICE-LAP3 Mch3 CMH-1	

Done

Many of the abstracts will contain some previous annotations that were done to the UniProt identifiers only and then mapped automatically to EntrezGene. It is important to still check these carefully and also to make sure all the mentions from the title are included.

It is important to cut and paste any mention from the text into the **Excerpt** column, with one mention per line. Only distinct mentions (not the same phrase) should be included.

Human Gene/Protein Normalization Annotation to EntrezGene Identifiers

http://smiley.mitre.org/~amorgan/pop/genedict

My MII Getting Started My Yahoo! Python Sidebar NCBI HomePage DAVID SF: BioCreAtivE G063 Wiki: Patterns ...

Annotator: alex

8663294 : ICE-LAP6, a novel member of the ICE/Ced-3 gene family, is activated by the cytotoxic T cell protease granzyme B.

Members of the ICE/Ced-3 gene family are likely effector components of the cell death machinery. Here, we characterize a novel member of this family designated ICE-LAP6. By phylogenetic analysis, ICE-LAP6 is classified into the Ced-3 subfamily which includes Ced-3, Yama/CPP32/apopain, Mch2, and ICE-LAP3/Mch3/CMH-1. Interestingly, ICE-LAP6 contains an active site QACGG pentapeptide, rather than the QACRG pentapeptide shared by other family members. Overexpression of ICE-LAP6 induces apoptosis in MCF7 breast carcinoma cells. More importantly, ICE-LAP6 is proteolytically processed into an active cysteine protease by granzyme B, an important component of cytotoxic T cell-mediated apoptosis. Once activated, ICE-LAP6 cleaves the death substrate poly(ADP-ribose) polymerase into signature apoptotic fragments.

Hints from EBI annotations

UniProt	Entrez	Names
P55211	842	APAF-3; ICE-LAP6; Q53Y70; CASP-9; CASPASE-9c; ICE-like apoptotic protease 6; CASP9; Caspase 9, apoptosis-related cysteine protease; Apoptotic protease Mch-6; Caspase-9 precursor; CASP9_HUMAN; caspase 9, apoptosis-related cysteine peptidase; Caspase-9; caspase 9, apoptosis-related cysteine protease; APAF3; Apoptotic protease activating factor 3; MCH6; Q53Y70_HUMAN

UniProt	Entrez	Excerpt	Comment
P55211	842	ICE-LAP6	
P42574		Yama CPP32 apopain	
		Mch2	
		ICE-LAP3 Mch3 CMH-1	

Blank space in need of EntrezGene ID

Existing annotation

Done

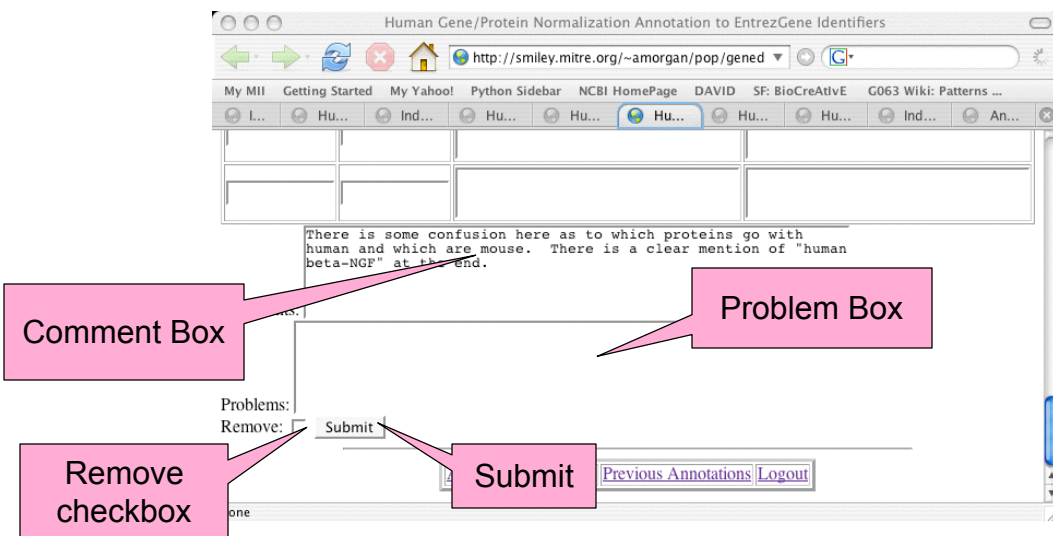
If the existing annotations (with links instead of open entries for ID's), simply remove everything from the excerpt and comment fields and these entries will be removed. Only annotations with entries in the excerpt field will be loaded into the database, and removing all the excerpt text implicitly removes that annotations.

New annotations can always be entered into the empty boxes below existing annotations. It is optional to include new UniProt identifiers with new annotations. It may not even make sense in many instances.

For existing annotations, it is important to add any missing EntrezGene identifiers. This is the key part of the annotation.

At the bottom of the annotation page, there is a box for general comments (in addition to particular comments going with a specific entry). It may be useful to check this box for comments on existing annotations. There is also a box to include specific problems about the abstract and the annotation process. If there are particular problems that make the abstract impossible to annotate, there is also a check box to mark the abstract for removal from consideration. Comments in the problem box should explain the reason for the removal marking.

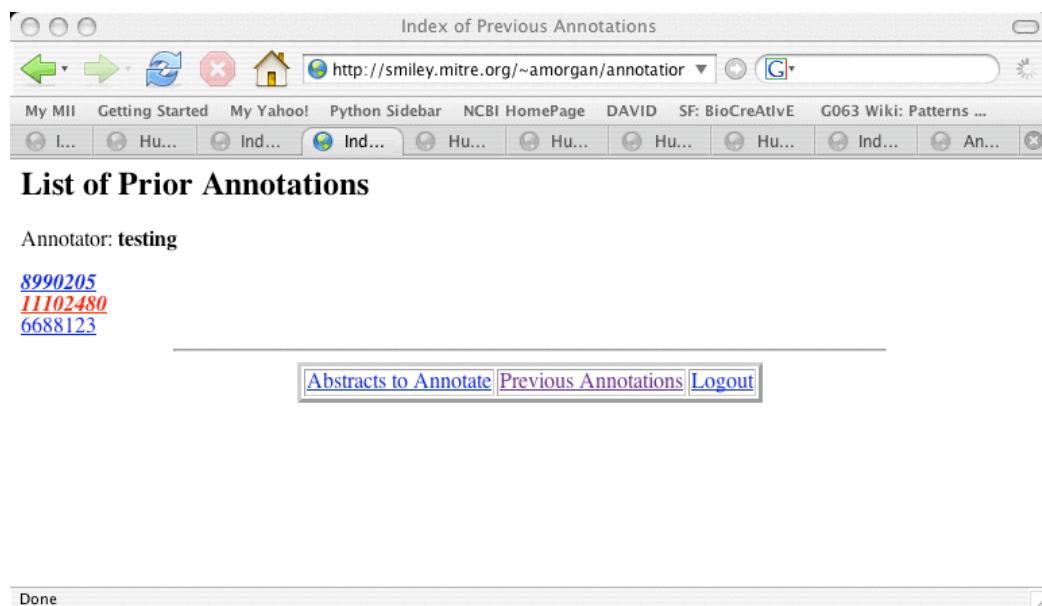
Once the annotation is complete, please hit the **Submit** button.



At the bottom of most pages there are a set of links to be able to return to the list of abstracts to annotate, to view your previous annotations or to logout of the annotation framework.



The prior annotations page should list all the abstracts you annotated in the current round of annotations. Ones with problems comments are in bold italics, and those marked for removal are colored red.



Annotation Tips

- Be sure to check to make sure that heteromeric proteins are not linked to identifiers for one particular subunit. Often subunit genes have the the name of the multi unit protein listed as a synonym (eg NF-kB), which can cause some confusion.
- To the right of the title of PubMed entries, there is "Links" offering that can provide links to database entries (such as Entrez Gene) that are associated with the paper. This can help aid in curation.
- Sometimes gene and protein names are split apart or fused differently in abstracts and the EntrezGene synonym lists. Be sure to check alternates if unable to find a match. (eg Foo Bar -> FooBar, FBar -> F Bar).

Quick List

[illegible]