BioCreative III Interactive Task: curation bottlenecks and solutions

Phoebe Roberts
15 September 2010
Agenda

• Acknowledgments
• Recommendations
• Interactive Task gene normalization bottlenecks
• Possible solutions
• Possible metrics
• Questions for UAG panel

If we assume that system won’t get it perfect, how can system make it easier to deal with errors and still aid curation?
Acknowledgments

All participating systems!

*Without your participation, User Advisory Group would have nothing to respond to*

- User Advisory Group
  - Lois Maltais (MGI)
  - Pascale Gaudet (DictyBase)
  - Andrew Chatr-Aryamontri (MiNT)
  - Livia Perfetto (MiNT)
  - Donghui Li (TAIR)
  - Phoebe Roberts (Pfizer)
  - Many others not in attendance today

- IAT organizers
  - Cecilia Arighi
  - Zhiyong Lu

- BioCreative organizers
Recommendations

- Allow curator to highlight all gene mentions at once
  - Relationships among gene mentions helps normalization
- Interactive systems a must!
  - Allow curator to delete wrong gene mentions
  - Allow curator to add missed gene mentions
    - Include lookup function to find options
- Alert curator to level of ambiguity within species and between
- Present curator with clues for ambiguity resolution
  - Description
  - Synonyms
  - Chromosomal location
  - Sub-cellular localization
  - Interacting partners
    - Less helpful:
      - Other mentions in article
      - Titles of articles with same gene mention
- Allow curator to make decision *in situ* (bring answers to them, don’t make them go to answers)
- Allow curator to see decisions and change them
DIFFICULT GENE NORMALIZATION
EXAMPLES
How ambiguous is the gene mention & what information helps resolve ambiguity?

**AIP1** mediates TNF-α-induced ASK1 activation by facilitating dissociation of ASK1 from its inhibitor 14-3-3

**ANSWER:** gene description from iHOP:

**VERY AMBIGUOUS 8 HUMAN GENES!**

**ANSWER:** gene description
Another ambiguity resolution problem

Figure 1 from paper
Species ambiguity: fair or foul?

Use case determines need for species specificity

• Sometimes species is required for curating:
  – model organism database info
  – Protein-protein interactions with definitive molecules

• Sometimes species interferes:
  – Authors are making point about evolutionary conservation of finding
    • Target of thalidomide
  – Most proteins not sufficiently characterized to warrant species filtering
    • E.g. at Pfizer, our protein dictionary is mammalian, not human, to increase recall of biological process and function
Earth-shattering conclusion derived from results across species

Half a century ago, thalidomide was widely prescribed to pregnant women as a sedative but was found to be teratogenic, causing multiple birth defects. Today, thalidomide is still used in the treatment of leprosy and multiple myeloma, although how it causes limb malformation and other central defects is unknown. Here, we identified cereblon (CRBN) as a thalidomide-binding protein. CRBN forms an E3 ubiquitin ligase complex with damaged DNA-binding protein 1 (DDB1) and Cul4A that is important for limb outgrowth and expression of the growth factor Fgf8 in zebrafish and chicks. Thalidomide initiates its teratogenic effects by binding to CRBN and inhibiting the associated ubiquitin ligase activity. This study reveals a basis for thalidomide teratogenicity and may contribute to the development of new thalidomide derivatives.

Conclusion from results: non-specific
Difficult gene mentions

Synonym not found

• Synonym is not found in databases searched (FN)
  – AtHscB (PMC2764847)
• Synonym is not found in any databases (FN)
  – Arabidopsis examples (PMC2764847)
• Species prefix obfuscates synonym (FN)
  – AtHscB (PMC2764847)

Ambiguity

• Synonym is a common English word (FN not in dictionary/FP many hits)
  – WASp
• Synonym maps to more than one identifier (FN for missed mapping/FP for wrong mapping)
  – AIP1
• Species not clearly specified (FP)
  – AIP1/ALIX
• Species deliberately not specified (FP)
  – One of the AIP1 papers
• Synonym is adjective that modifies a non-gene (FP)
  – SufD-like protein (PMC2764847)
• Synonym refers to a protein family or an enzymatic activity (FP)
  – ATPases (PMCID 2275796)
  – Not appropriate to map to an identifier, BUT still some utility from annotating it
SOLUTIONS
Difficult gene mentions and solutions

How can a curator more easily resolve a...

**Synonym not found**
- New synonym is not found in any databases (FN)
  - Increase breadth of databases searched by tool
- Synonym is not found in all databases (FN)
  - Ability to add a synonym and reprocess highlighting
- Species prefix obfuscates synonym (FN)
  - Ability to add synonym or species-specific rules for string matching

**Ambiguity**
- Synonym is a common English word (FN not in dictionary /FP many hits)
  - Ability to add or remove a synonym and reprocess highlighting
- Synonym maps to more than one identifier (FN for missed ID/FP for wrong ID)
  - Present matches simultaneously with clues like other synonyms and interacting partners
- Species not clearly specified
  - Be able to navigate to other sections of the paper, other papers
- Species deliberately not specified (FP????)
  - Navigate to references
- Synonym is adjective that modifies a non-gene (FP)
  - Ability to remove from list
- Synonym refers to a protein family or an enzymatic activity (FP)
  - Ability to removed from list
Methods for dealing with multiple hits

Abstract

Background

The ALG2-interacting protein X (ALIX)/AIP1 is an adaptor protein with multiple functions in intracellular protein trafficking and egress by facilitating the transport of Gag to enveloped viruses. The ubiquitin E3-ligase POSH induces the ubiquitination of ALIX both in vivo and in vitro. This ubiquitination does not affect its regulatory function. As it is well established that ALIX regulates HIV-1 infectivity, we demonstrated that wild type POSH, but not the C-terminal domain mutant HIV-1, HIV-1AIP1, induced the ubiquitination of ALIX.

Results

In this study we identified ALIX as a POSH substrate. The ubiquitin E3-ligase POSH induces the ubiquitination of ALIX both in vivo and in vitro. This ubiquitination does not affect its regulatory function. As it is well established that ALIX regulates HIV-1 infectivity, we demonstrated that wild type POSH, but not the C-terminal domain mutant HIV-1, HIV-1AIP1, induced the ubiquitination of ALIX.

View options, check each, don’t have to commit, commitment is propagated
Synonym lookup functions to aid GN

IscU/Isu, which is regulated by HscB/Jac1 by binding to IscU/Isu to assist [Fe-S] delivery to the chaperone [12], [44]. Yeast Jac1, Ssq1 and Isu have been confirmed to be mitochondrial proteins [12].

Here we demonstrate that Arabidopsis contains a functional AthscA1/AthscB/AthiscU1 protein cluster involved in [Fe-S] protein biogenesis. In contrast to yeast, the AthscA1/AthscB/AthiscU1 protein cluster is localized to both mitochondria and the cytosol of Arabidopsis suggesting a dual action between these two spatially separate compartments.

Results

AthscB can rescue yeast Jac1 knockout mutant

A full-length cDNA (750 bp) encoding the AT5g06410 open reading frame was cloned into pET22b to E. coli HscB and respectively (Figure). The fragment contains the HPD motif, a predicted 59 amino acids according to the CDD in Arabidopsis has At5g06410 AthscB.

To confirm that AthscB protein, we performed a lethal knockout transformed Delta (URA3) marked plate. The AthscB and positive dropout media SDY-Leu (minus L-leucine) or on SDY-Trp (minus L-trypothan), respectively. Once scored the wild-type Jac1 cDNA

Reflect - At5g06410

<table>
<thead>
<tr>
<th>Protein</th>
<th>Add</th>
<th>About</th>
</tr>
</thead>
<tbody>
<tr>
<td>At5g06410 (AT5G06410.1)</td>
<td>A. thaliana</td>
<td></td>
</tr>
</tbody>
</table>

No Synonyms

Sequence, Domains, Structure, Locus, Literature

No structure information available

No annotation available
New evaluation metrics

• Time spent in system (please include PAUSE button!)

Utility
• Increased TPs and TNs, decreased FPs and FNs
• Novelty
• Number of URLs visited outside application
• Indicator of resources provided within application and how often curator has to leave to accomplish task

• Subjective metric
  – Did curator enjoy the user experience?
Other questions

System specification questions
• Is any identifier better than none?
  – Entrez, UniProt, IMAGE, TAIR, etc.
• Is one application with many functions better than multiple applications (e.g. browser plus spreadsheet)
• Is finding the identifier the hardest part of GN, or determining whether an identifier is appropriate?

More research needed?
• Is there a point at which too many FPs are worse than FNs?
• Do we need to know how curators normalize gene mentions?
• Do we need to know the frequency of GN difficult tasks?
• Will someone build a first mention corpus?
  – Here we describe LINGO-1, a nervous system-specific transmembrane protein that binds NgR1 and p75 and that is an additional functional component of the NgR1/p75 signaling complex. (pmid:14966521)
  – We describe a yeast enzyme, Doa4, that is integral to the degradation of ubiquitinated proteins and is required in diverse physiological processes. PMID: 8247125
Who is curation tool for?

- Domain expert
- Domain novice

- Curation expert
- Database curator
- Joe/Jane Scientist

- Author
- Curation novice