Functional Annotation of Protein Isoforms and Modified Forms

Harold J Drabkin¹, Cecilia N. Arighi², Cathy H Wu², Judith A Blake¹
¹Mouse Genome Informatics, The Jackson Laboratory, Bar Harbor, ME, USA
²Georgetown University Medical Center, Washington DC, US

Abstract Eukaryotic organisms can generate protein diversity through both post-transcriptional as well as post-translational processing events. The multiple proteins arising from the same gene may differ from each other in their temporal or tissue-specific expression, molecular function, cellular localization or participation in biological processes. Currently, many model organism databases associate functional and other annotations collectively to the gene encoding a protein object. Thus, functional, temporal and spatial distinctions between protein isoforms arising from a single gene are lost. In this paper, we discuss the strategies and challenges encountered by the Mouse Genome Informatics curations during the process of annotating protein isoforms.

Keywords Biological data mining and knowledge discovery, Bio-ontologies, Biological databases and information retrieval, Biological data visualization, Biological data integration

1 Introduction

One of the hallmarks of gene expression in eukaryotic organisms is the ability of a single gene to give rise to multiple gene products, including isoforms originated through mRNA processing, and a wide spectrum of protein forms as those derived from cleavage (such as signal peptide, or processing for activation) and/or post-translational modifications (such as phosphorylation, acetylation, glycosylation and ubiquination). This process can increase a gene's potential coding capacity several fold. Quite often, these isoforms share many functions. However, it is also the case that they can often differ in terms of tissue specificity of expression, subcellular localization, and functional properties.

Figure 1 Schema of the Protein Ontology Framework
The generation of protein diversity from a single gene can make the systematic functional annotation of gene products difficult. Many Model Organism Databases (MODs) utilize the Gene Ontology (GO) for functional annotation of gene products. Annotations provided in gene association files (GAF) for each organism are tied to a single object, either gene, transcript, or protein. However, in most cases, the annotation defaults to the level of the gene. Even when the annotation object is a protein, as in the case of GOA project, the object record may nonetheless represent a collection of various isoforms. Thus, annotations to “genes” are difficult to disambiguate to specific protein isoforms. The focus of this paper is to review the strategies and challenges to functional annotation of protein isoforms.

2 Background

The Protein Ontology (PRO)

The Protein Ontology (PRO) attempts to provide a framework that can provide formal descriptions of protein entities and their relationships and this framework can support the annotation of specific protein isoforms [1]. The PRO includes an ontology of proteins based on evolutionary relatedness (ProEvo), and an ontology of protein forms (ProForm) describing the multiple protein end-products produced from a given gene locus (Figure 1).

A working prototype of the PRO has been developed and submitted to the Open Biological Ontologies (OBO) foundry [2, http://obo.sourceforge.net]. Isoforms are represented as children of a protein class representing a collection of protein products of a given gene in any organism (in the case study only human and mouse protein forms are described). GO annotations for an isoform, and its modification state, if known, are indicated as xrefs for the isoform (Figure 2).

Mouse Genome Informatics (MGI)

The Mouse Genome Informatics resource (MGI, http://www.informatics.jax.org) provides high-integrated information about the genetics, genomics and phenotypes of the laboratory mouse for the scientific community. MGI leaders have made a commitment to capture not only isoform data, but also data on tissue, cell type and modification state of the protein. Although not all these data types are as yet fully represented in MGI, we have developed a simple framework to link current GO annotation to specific isoforms and modification states through the use of fields which are filled in with Dbxrefs referring to specific to sequence databases (UniProtKB and NCBI), as well as other ontologies available on the Obo Foundry site (http://obo.sourceforge.net). These include (psiMod (modification), MA (mouse adult anatomy), EMAP (embryonic mouse anatomy), CL (cell type), SO (sequence ontology), and others. By using the OBO Foundry
ontologies, a consistent manner of annotation can be adopted, which in theory, can also be reasoned on due to the structures of those ontologies.

The core of the functional annotation task is to extract information from the experimental literature and deposit it in a database resource. Figure 3 shows the detail of the MGI GO editorial interface (EI). Each line represents a single annotation. The curator can input the GO_ID, GO_ID qualifier, reference, evidence code, and evidence code qualifier (also known and the inferred from/with field1). Additionally, there is an annotation NOTES field, which was introduced in 2003, which when first opened presents the dialog shown in Panel B. There the curator can enter additional information using appropriate cross-references to other ontologies. Figure 4 shows several examples of entries for different annotations. An example of annotation for protein isoforms can be seen in the example of annotation of the protein products of the gene Zfpm2 (zinc finger protein, multitype 2). The gene encodes two variants, a long and a short form (Q8CCH7 and A5HE91, respectively). Although both bind to the transcription factor Gata2, the shorter form does not act as a transcriptional co-repressor and thus is not involved in the negative regulation of transcription as is the long form [3]. Because the annotation is tied to the gene, the resulting display appears to be conflicting (Figure 5). However, on the editorial interface, the notes fields for these data (Figure 6). At this time, approximately 595 annotations have an entry in the “Gene_product” field indicating annotation to a specific isoform2.

3. Annotation Efforts: Challenges

There are many challenges to be dealt with for isoform annotation. These are at the information management level as well as at the annotation level.

3.1. Information management challenges

3.1.1 Identification of genes having isoforms.

Nearly 25,000 genes in MGI have one or more curated protein sequences linked to them. One may get an idea of the number of potential isoforms from the number of alternative transcripts that may be either predicted (e.g., NM RefSeqs) or described (cDNAs, ESTs). However, additional experimental evidence at the protein level may be required.
For example, as shown in figure 7, although two RefSeqs entries are associated with the gene Atp2b2 (NM_009723, and NM_001036684, Figure 7a), the protein that these predict although labeled isoform 1 and 2 (NP_033853 and NP_001031761, respectively, Figure 7b), are in fact identical (see comparison, Figure 7c).

UniProt has introduced the protein existence line (PE line) in UniProtKB records to indicate what the evidences are of the existence of a protein. Approximately 3657 markers in MGI have a UniProtKB ID whose record contains information on multiple isoforms (6.3%).

3.1.2. Lack of unique identifier in data resources

Not all protein objects receive unique identifiers. For example, Notch 1 activation occurs through binding of an extracellular ligand which results in a cleavage in the intracellular domain of Notch, releasing the notch intracellular domain (NICD). This fragment in turn travels to the nucleus, where it binds RBJ-1 and acts as a transcription cofactor, stimulating the transcription of downstream genes (4). Notch1 gene, or even the Notch1 protein, would have cellular localization annotation to integral to membrane. However, the fragment, now as a separate functional unit, has an annotation to nucleus. The UniProtKB record for Notch1 (Q01705) indicates two isoforms due to alternative splicing. However, neither of these represent the NICD fragment alone. Thus, the annotations to nucleus and transcription cofactor/activator can not be properly represented.

Another case is seen in a study of Mybl2 (myeloblastosis oncogene-like 2,) where Horstmann et al. [5] described the properties of a “short form” of murine Mybl2. This form is exon 9A. Unlike the “long form”, the short form does not activate transcription. However, a unique sequence identifier is not found in either Genbank or UniProt resources.

In some cases, PRO may be able to address some of these issues. For example, unique identifier may be given to the NICD fragment with a “derives_from”relationship to the Notch protein. This would then become annotated object.

3.1.3. Published literature fails to identify the isoform

Often, the specific isoform being used in an experiment is not always readily discernible or even mentioned in the article. The gene Trp63 has six isoforms. Cheng et al. [6] have implicated a role for an isoform \( \Delta N6 \) in ventral bladder development. There are three forms of \( \Delta N6 \) (\( \Delta N6\alpha \), \( \Delta N6\beta \), and \( \Delta N6\gamma \)), but it is not clear that the three \( \Delta N6 \) isoforms...
are involved. Thus, the annotation would be applied to the entire gene.

### 3.2. Functional Annotation Challenges:

#### 3.2.1 Updating older annotations with isoform data

As mentioned earlier, the notes field has been available in its current form since 2004. However, MGI curators have been using GO for functional annotation since 2000. Thus, we need to reassess the annotations made prior to the time when isoform specific data could be captured. Several strategies are being used to aid in the identification of publications used for GO annotations that may actually be specific to isoforms. The simplest of these is to see if any GO annotations have been made using papers that mention “isoform” or “alternative splicing” in their title or abstract.

#### 3.2.2 Functional annotation of protein isoforms outside of GO.

The private notes field may also play a pivotal role in linking functional annotation outside of GO to specific isoforms. For example, one can describe specific isoform expression during embryonic development, or identify an isoform specific allele. Since the notes field have slots for anatomy (adult or embryonic), and the GO inferred from field tracks specific alleles, there may be a way to connect the isoform slot to other areas of the MGI database.

### 4. Future Goals

With the notes field in place, GO curators can now enter isoform data during new curation. Our current estimate is that between June 2007 and February 2008, we have added over 200 new isoform specific annotations for a total of over 600 annotations. The next goal is to enable public access to these notes. Ideally, one would like to annotate directly to a specific isoform object, and have the objects and their annotations collected under the gene object. This is actually a goal of the MGI sequences as objects (SAO) project, but will require massive schema changes to the database. It will be easier, in theory, to at least be able to display the note field of a specific annotation as a pop-up, or to provide an additional column in standard GO summary page to contain the isoform id.

Figure 5 GO annotation summary for Zfpm2 as seen at [http://www.informatics.jax.org](http://www.informatics.jax.org), showing two apparently conflicting annotations (participates in negative regulation of transcription from RNA polymerase II promoter, and does NOT negative regulation of transcription from RNA polymerase II promoter. Both make reference to the same publication (J:121803)
Figure 7 Atp2b2 has two RefSeqs coding for the same protein. A. Mouse G-browse view of the RefSeqs for Atp2b2. B. The protein RefSeqs and their sequence. C. Blast2 alignment of the two protein RefSeqs shows 100% identity.
References


