Aliases and Ambiguity: A case study of gene aliases, and implications for information curation and AI

Chandler Armstrong
University of Illinois, Urbana-Champaign
Department of Sociology
57 Computing Applications Building
605 E Springfield Ave
Champaign, IL 61820
carmstr3@illinois.edu

ABSTRACT

This research seeks to understand how names and aliases of concepts are used in scientific literature. Natural language processing tools, and data curation in general, depend upon unique concept identifiers for information, and aliases only provide more oppurtunity for ambiguiyt; despite this, aliases seem to persist in literature and daily life. As a case study, gene names are analyzed. This article presents a discussion on patterns of alias usage, and implications this has for bioinformatics librarianship. Observation suggests that research scientists in the bio-medical fields think about information organization from their contextual perspective, and organize information to be most applicable to their daily research tasks. Information scientist might think about information from a more generalized perspective, and prefer categorizations that minimize ambiguity. Aliases used by scientists probably emerge for functional reasons, each providing distinct semantic roles, despite that they create ambiguity from an information curation perspective. In light of this, information science must be careful to consider contexual needs of information users, and word sense disambiguation models will become increasingly important to deal with the increasingly complex grammar for talking about concepts in scientific research.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Gene Name Disambiguation

Keywords

Gene Name Disambiguation, Name Ambiguity, Synonyms

1. INTRODUCTION

This research seeks insight into name ambiguity, to better understand why single concepts possesses multiple names, and asks how to resolve these situations. It is motivated by the degree of ambiguity found in the names and aliases of concepts used by everyone on a daily basis, in both professional and casual capacities. In bioinformatics, this ambiguity has garnered a special interest. Gene names and aliases, in particular, present thorny difficulties and are the focus of specialized areas of model building and research [5].

Gene-name ambiguity arises where gene aliases have multiple senses; it can be envisioned as a special case of word sense disambiguation (WSD). Genes always have an official name, and commonly have an alias or set of aliases; coincidentally, aliases often serve as abbreviations for full gene names. Chen et al [3] investigate the names and aliases for mouse genes, and reports that while gene name ambiguity is low at about fourteen percent, the aliases are far more ambiguous, at eighty-five percent. Schumie et al [4] analyzed both abstracts and articles for gene symbol versus gene alias usage, and found that full gene names were used only thirty percent of the time in abstracts, and eighteen percent of the time in full text; the remainder the time only a gene alias or abbreviation is used. Current attempts at disambiguation vary from 77 to 100 percent accuracy, depending on the details of the model and the species to which the gene belongs (with genes found in humans being the most ambiguous) [1][6]. In summary, aliases are used more often than full names, aliases are very ambiguous, and disambiguation attempts, while promising, are lacking in key areas.

The hypothesis was that, over time, a gene would come to posses fewer aliases; with ideally all the usage converging into a single name. This expectation arises from the intuition that having fewer names for a thing make it easier to find all the information about it. In the limited amount of manual observation made, we do not see the hypothesized patterns. Rather, usage of all aliases increases across time. This suggests that each alias possesses some distinct semantic function.

The observed patterns makes sense, although why may not be immediately apparent. We reason that most genes appear in multiple species, or be involved in multiple bio-medical concepts (such as diseases and drugs). Thus, genes have a lot of range, and are interdisciplinary. From the user's perspective, having multiple aliases eases the taks of finding information on a given gene as it relates to a specific area of research. Following analysis, it was clear that the original hypothesis was formed based upon a certain set of assumptions, and these assumptions simply do not hold for the bio-medical research scientist using and creating gene aliases.

The findings suggest that information science needs to be aware that information users may have different needs which will compete with the best practices of information organization. Multiple names, even at the price of ambiguity and diffusion of information, possibly provide some benefit to information users.

If our observations represent the state of affairs, automated disambiguation models will become increasingly important, since eliminating ambiguity through tightly controlled concept naming may not be possible. Disambiguation models are increasingly dependent on metadata provided in curated databases. Where data-mining and AI once attempted to *learn* a word sense, contemporary models are dependent on *infering* a word sense from other information available in the database. This means data curation may not need to focus on providing unique concept identifiers for everything, but should provide a diversity of information about items.

2. METHOD

Gene names and aliases were taken from the Entrez Gene database. Each gene in this database possesses a unique identification number: its GeneID. The database also includes the set of known aliases for each GeneID. We search the entire collection of abstracts available from Pubmed for all gene names and aliases. The search was a simple deterministic search with some processing to maximize matching. First, the abstracts were tokenized into single, bi, tri, and four grams. This is because gene names are often multiple tokens, so it is important to attempt to match sequences of tokens up to some length. Most gene names and aliases will be found with up to a four gram search, however not all will (eg. "A. Thaliana Receptor Kinase 1" requires a sequence of five tokens to possibly match). Tokenized abstracts were searched without modifying case, and with case of all tokens uppered. This is because occasionally gene names will be written in upper case. Finally, if a token contained a dash, it was tried as a single token with the dash, and as two tokens without it. This is because authors occasionally add dashes to multipart names.

The concept of a 'gene' is not well defined by this research. The Entrez Gene data has been used naively; its list of genes may not all be, indeed, genes. In reality, many components go into making a gene, and Entrez Gene data may consist of any of these components. Therefore, the data we are investigating may be components in a process that includes whole genes, but also RNA, protiens, and enzymes. Therefore, we do not simply assume that each of our datapoints are genes. However, we do define our data as information on named entities directly relevant to the scientific field of gene research.

The results of the search are text files listing each GeneID, the alias, and all PubmedIDs containing that alias. The set of PubmedIDs for each alias can then be used to quickly look up information such as dates and authors. Using this information, aliases are grouped by the GeneID to which they refer, and then each mapped to the dates of the articles containing them. Then GeneID's family of aliases is plotted into bar graphs for frequency of usage by date. These plots allow quick visual inspection for interesting patterns in usage.

The results of the plotting are, literally, hundreds of thousands of plots. To find interesting ones, many are randomly selected and inspected. We are interested in plots that have two or more aliases with at least a total of thirty instantiations of an alias (and preferably many more). Patterns which might provide evidence supporting the original hypothesis were also specifically sought.

3. RESULTS

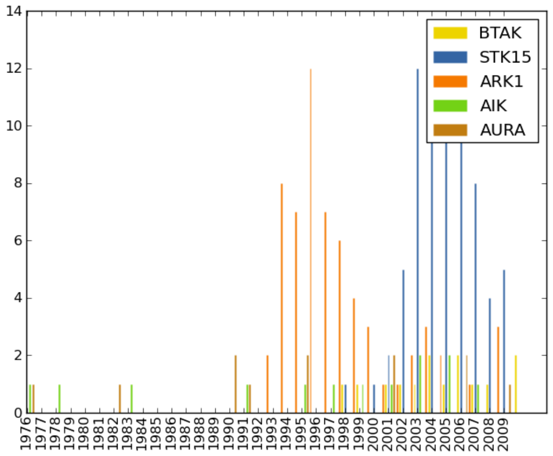
Initially, some plots did seem to support the original hypothesis. However an in-depth analysis revealed that what appeared to be 'advisarial' competition between two aliases was simply an artifact created by ambiguous usage of one of the aliases. One particular case is geneID 6790, the AURKA or Aurora Kinase gene. This gene is present in Humans, and the mold S. Pombe. In S. Pombe the gene is a variant: the Aurora Kinase B. In the literature, when referencing this gene in S. Pombe, the string 'ark1' is most commonly used. 'ark1' is an alias for both Aurora Kinase, and Aurora Kinase B. Additionally this string is a component of 'beta-ark1'; which is not the Aurora Kinase gene (not even the 'Aurora Kinase B' gene, despite the 'beta'). Therefore, a simple match for 'ark1' will often match 'beta-ark1'.

The orange bars in figure 1 below are apparent instantions of GeneID 6790 through the alias 'ark1'. Notice that these instantiations do not align, in terms of variance patterns, with the other aliases; it appears that as the oranga bars decrease, the blue ones are rising. However, the usage seen in the orange bars is almost entirely ambigous: they result from references to 'beta-ark1'. The usage pattern seen in the orange bars is not reflecting how the concept of GeneID 6790 is actually being used. As an aside, when GeneID 6790 (Aurora Kinase) is referenced in Humans, it is usually referred to using 'STK15' (serine/threonine kinase 15), notice that this is the most used alias.

An important element of the ambiguous usage of 'ark1' is that it seemed to always come from other genes. 'ark1' serves as a referent for the gene A. Thaliana Receptor Kinase. More disturbing, 'ark1' is often used in literature to refer to genes who are not listed as possessing that alias. For example, both ARBORKNOX1 and the armadillo repeat-containing protein (an RNA name listed in Entrez Gene) use 'ark1', although this is not listed as a known alias for these concepts. The presence of mispellings and usage of unrecognized names remains a primary source of error in gene name disambiguation models. Current models to resolve this sort of error attempt predict how scientists produce aliases, and generate the likely mispellings and misuses of aliases for a gene [2].

The original hypothesis assumed it is intrinsically beneficial, from an information retrieval perspective, to have fewer names for a given concept. The rejection of the hypothesis

Figure 1: GeneID 6790, frequency of use over time



demonstrates that this is not the case. Rather, it may be the case that it is at least occasionally beneficial to have multiple names for a single concept, in order to represent that concept in a different context. This observation, if true, has important implications for librarianship and data curation. What is ambiguity from the perspective of concept sense, is disambiguity from the perspective of context sense. This also means data-mining and AI models for disambiguating concept senses will become increasingly important.

Many contemporary gene name disambiguation models are based upon data-curation, and using the information provided in the databases (such as author and MeSH terms) to disambiguate gene word sense [1][6]. Because feature selection is swinging away from characterizing the data directly, and instead infering classifications using other data in the database, errors in the models are likely arising where the integrity of the database breaksdown: with mispellings, unknown aliases, and ambiguity in the curated data (such as author name ambiguity). Improving accuracy of these models must solve these problems, and data curation and ontology research will also become increasingly critical.

Importantly, the problems of ambiguity of concept senses in these databases are interelated. Any name in a database, such as author or journal name, are candidates for ambiguity. Given that gene name disambiguation models may use author or journal of an article as a feature, improving performance of these models depends on disambiguating other named concepts in the database.

4. CONCLUSION

The original hypothesis is a 'natural selection' one in which pressures on the way names are used favors fewer names, and that we might expect to see competitions between names for 'usage resources'. Analysis did not support this hypothesis. Multiple aliases appear to have a function at least some of the time. This has implications for science and research in general, where single concepts are utilized across a wide range of fields, so that it becomes beneficial for the concept to possess a different name for each of these contexts. If this is the case, word sense disambiguation, especially in scientific research, will become increasingly important, as ambiguity will always be rising. The methods and features used by data-mining and AI models are changing to utilize the curation of this mass of data, and utilizing inference from data curation and ontologies to make discoveries and disambiguate word sense.

5. FUTURE

Grouping aliases of a gene and using this as a vehicle for making comparisons and finding relationships has shown to be a productive mechanic. It is likely that other methods of grouping genes may be similarly fruitful. An immediate possibility is grouping the genes of an alias, and using this is a challenge for disambiguating their usage. We found that the nature of gene names appears to produce congruent aliases between many different genes (eg aurora receptor kinase, a. thaliana receptor kinase, beta-adrenegeric receptor kinase, all using the string 'ark1'). Other possibilities include grouping authors of a gene, aliases of genes of a species, journals of a gene, or any other variable which may exhibit a clustering of gene usage, such that certain genes tend to cluster around certain objects (ie around certain authors, journals, disciplines, species, etc).

Given that current source of error in gene sense disambiguation are mispellings and usage of unknown words, models for identifying these anomolies are required. Such models will need to depend even more heavily on data curation and ontologies than ever before, as a word sense may have to be predicted in a set of tokens without utilizing the token itself. Needless to say, this is a difficult problem.

6. REFERENCES

[1] Richard Farkas. The strength of co-authorship in gene name disambiguation. *BMC Bioinformatics*, 9:69, 2008.

[2] Jorg Hakenburg. What's in a gene name? automated refinement of gene name dictionaries. *Biological, translational, and clinical language processing*, pages 153–160, 2007.

[3] C Friedman L Chen, H Lui. Gene name ambiguity of eukaryotic nomenclatures. *Bioinformatics*, 21:2248–2256, 2005.

[4] MJ Schumie, M Weeber, BJ Schijvenaars, EM van Mulligan, CC van der Eijk, R Jelier, B Mons, and JA Kors. Distribution of information in biomedical abstracts and full text publications. *Bioinformatics*, 20:2597–2604, 2004.

[5] Lorraine Tanabe and W John Wilbur. Tagging gene and protein names in biomedical text. *Bioinformatics*, 18:1124–1132, 2002.

[6] Hua Xu, Jung-Wei Fan, George Hripcsak, Eneida A. Mendonca, Marianthi Markatou, and Carol Friedman. Gene symbol disambiguation using knowledge-based profiles. *Bioinformatics*, 23:1015–1022, 2007.