Mining Gene Ontology Annotations From Hyperlinks in the Gene Wiki

Benjamin M. Good, PhD, Andrew I. Su, PhD
The Genomics Institute of the Novartis Research Foundation, San Diego, CA

Abstract

The Gene Wiki is an informal collection of more than 10,000 Wikipedia articles about human genes. Through the continued contributions of many volunteers, it is continuously growing as a valuable repository of biomedical knowledge. While initial efforts were devoted to seeding Gene Wiki articles with data from public databases, we are now looking for ways to harvest the human-added knowledge accumulating in these articles. One of the sources of such potential knowledge are the links between Gene Wiki articles and articles describing biological concepts. Here, we assess the potential of such inter-wiki links to signal novel gene annotations. This analysis was performed by mapping the targets of Gene Wiki links to Gene Ontology terms and then comparing the resultant connections to known Gene Ontology annotations. We found a total of 12,828 potential annotations of which 5,005 (39%) correspond to existing annotations and 7,823 (61%) represent candidates for new gene annotations.

Introduction

Gene Ontology Annotations

Gene Ontology (GO) annotations facilitate the description of genes according to molecular function, biological process, and cellular component [1]. The primary public repository of gene ontology annotations is the Gene Ontology Annotation (GOA) database [2]. The breadth and depth of this gene index has enabled the creation of algorithms that are now the de facto standard starting point for the functional analysis of the gene lists characteristic of high-throughput molecular biology [3].

While clearly valuable in its current form, the gene ontology annotation database is far from complete. Currently, the Entrez Gene database lists records for 27,050 human genes (distinct genes of human origin with pseudogenes removed). At the same time, the GOA database documents only 17,792 human genes with any GO annotations; more than one third of the genes in Entrez Gene have no GO annotations at all. Furthermore, the annotations that do exist are spread disproportionately across the genome. As Figure 1 illustrates, a few genes are very well annotated while many genes are poorly annotated [4]. Even without considering the impact of new technologies that will expose new aspects of genetic variability with functional implications, it is clear that the work of describing the function of the components of the human genome is long from finished.

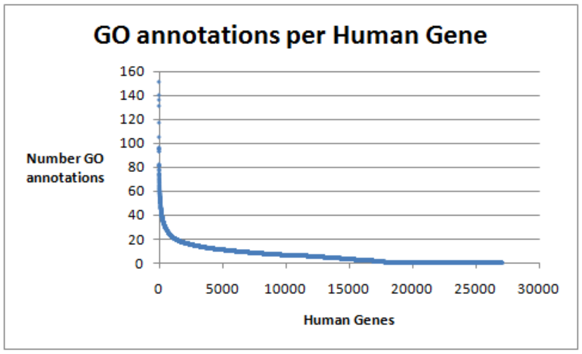

Figure 1: Number of gene ontology annotations per human gene. Points on the X axis correspond to the 27,050 distinct, non-pseudogene gene records found in Entrez Gene on Aug 5, 2010.

Generating Annotations

The Gene Ontology annotations provided by the GOA database are recorded by curators who utilize the scientific literature, computational tools and their own knowledge to identify new annotations [5]. While this process generates thousands of new annotations every year, its requirement for highly skilled, manual labor makes it very expensive to scale up as data and demands increase. As a result, increasing attention is being paid to methods that help to automate the process by predicting gene annotations based on sequence and by identifying gene names and ontology terms in the text of scientific articles [6].

In addition to professional curation pipelines and automation, approaches are emerging that engage researchers directly in the process of aggregating and synthesizing scientific knowledge. Such approaches attempt to harness the collective intelligence of the broad scientific community to address tasks, like gene annotation, where manual effort is needed. As suggested in a 2008 article in *Nature*,

> "*Sooner or later, the research community will need to be involved in the annotation effort to scale up to the rate of data generation.*" [7]

It is in this vein that the Gene Wiki was created [8]. The Gene Wiki is an informal collection of more than 10,000 Wikipedia articles about human genes. Like most Wikipedia entries, the contents of these pages can be edited by anyone. This openness combined with the impressive visibility of Wikipedia pages (which are often the top hit on Web searches for gene names) has helped to make the Gene Wiki one of the most successful collective intelligence applications in biology to date [9].

The Gene Wiki was created by seeding Wikipedia article stubs with content from existing annotation databases. Through the continued work of the global community, these initial seeds have grown into a useful, though loosely structured resource for biologists and laypersons to learn about gene function. Now, two years after its creation, we are starting to consider the reverse flow of information. Specifically, we seek to mine structured data from the Gene Wiki to improve the databases that were used to initially seed Gene Wiki content. Here, we explore an implicit mechanism through which contributions to the Gene Wiki can be used to identify new potential Gene Ontology annotations.

WikiLinks as potential sources of Gene Annotations

One of the fundamental aspects of the Gene Wiki is its interconnected nature. Each page links to many other pages forming a useful, directed network of information sources. We hypothesized that if the pages linked from a Gene Wiki article can be mapped to specific terms in the Gene Ontology, then each link provides a candidate annotation for that gene. Figure 2 displays an example of this situation. It shows how the article for the gene '5HT1A', which maps to NCBI Gene:3350, contains a link to the page for 'vasodilation' which in turn can be mapped to the GO term with the same name (GO:0042311).

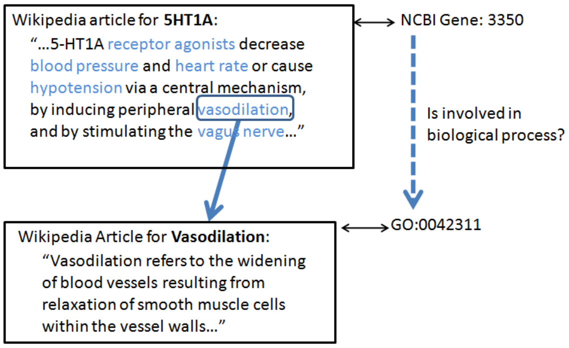
Wikipedia article for **5HT1A**:
"...5-HT1A receptor agonists decrease blood pressure and heart rate or cause hypotension via a central mechanism, by inducing peripheral vasodilation, and by stimulating the vagus nerve..."
NCBI Gene: 3350
Is involved in biological process?
Wikipedia Article for **Vasodilation**:
"Vasodilation refers to the widening of blood vessels resulting from relaxation of smooth muscle cells within the vessel walls..."
GO:0042311

Figure 2: Link from the 5HT1A Gene Wiki article to the article for 'vasodilation' suggests a new GO annotation for 5HT1A.

The annotation of 'vasodilation' for 5HT1A (also known as HTR1A) suggested by the link shown in Figure 2 is not presently recorded in the GOA database and hence represents a novel candidate annotation.

To evaluate such candidate annotations, bio-curators would subsequently consult the literature and secondary sources. In this particular case, we manually identified supporting evidence for the potential annotation in the Drugbank database [10]. We found that HTR1A is one of two targets for the drug Ergoloid mesylate (DB01049) and that Ergoloid mesylate is a vasodilator.

In sum, these data strongly suggest that 5HT1A (HTR1A) has a bona fide role in vasodilation, and that links in the Gene Wiki can be a fruitful source of candidate annotations.

Methods

To comprehensively mine the Gene Wiki for candidate gene annotations, we first extracted a set of links from Gene Wiki articles to Wikipedia articles that could be mapped to GO terms and then compared the implied annotations to the annotations in the GOA database.

This analysis was initiated by identifying 10,102 Gene Wiki pages with direct matches to entries in the Entrez Gene database using the Gene Wiki plugin for BioGPS [11]. Next, we used the Wikipedia API [12] to collect the formatted wikitext from each of the articles and applied a regular expression to extract the WikiLinks from the article text. (Links that were automatically added via 'transclusion' of templates were not considered.) For each of the pages that were linked from a Gene Wiki page we attempted to find a match in the GO. The matching process works as follows. For each linked page:

1. Find all the redirects for that page. If the page itself is a redirect, find all other redirects that refer to the same page.
2. Compare the title of the page as well as all of its redirects to the terms in the GO to identify any matches. (All String comparisons ignore case, spaces, commas, and quotation marks.) We compared using:
    a. the GO_wikipedia_xref mapping (is the Wikipedia page linked as an xref for a GO term?)
    b. the title of the GO terms
    c. all the synonyms provided for the GO terms

With the exception of matches to 'protein', which were discarded, all matches were recorded along with the evidence for the match. Evidence documents whether the match came from the linked page title or a redirect and whether the preferred label for the GO term was used or a synonym. Because the GO applies multiple kinds of synonyms, the specific synonym type ('exact', 'broader', 'narrower', or 'related') was also recorded. For matches to obsolete GO terms, the terms that replaced the matched term were added as matches using the same evidence type as the match to the obsolete term. When no replacement term was identified, its obsolescence was recorded and the match was kept for the analysis. For the results presented here, all forms of matching evidence are used unless otherwise noted.

Once the mappings from the WikiLink targets to the GO were established, the gene2go table provided by the NCBI was used to collect GO annotations of human genes. These annotations were then compared to the WikiLinks associated with GO terms. A candidate annotation was classified as 'known' if it matched an existing annotation directly or if it matched a less specific GO term from an existing annotation. The less specific terms (ancestors) included the full, transitive closure of the 'is a', 'part of', and 'regulated by' relationships. For example, the page for Fibronectin suggests a candidate annotation of 'cell adhesion' but, since Fibronectin is already annotated with 'substrate adhesion-dependent cell spreading' (a descendent of 'cell adhesion') the candidate annotation is considered to be known.

Results

The Gene Wiki articles contained a total of 96,898 links to 21,422 distinct pages within Wikipedia. Of these, 1,655 were mapped to Gene Ontology terms. Figure 3 summarizes the process and key results of mining and classifying known and unknown annotations in the links from Gene Wiki articles. As it shows, we identified 5,005 known GO annotations in these links as well as 7,823 unknown, candidate GO annotations.

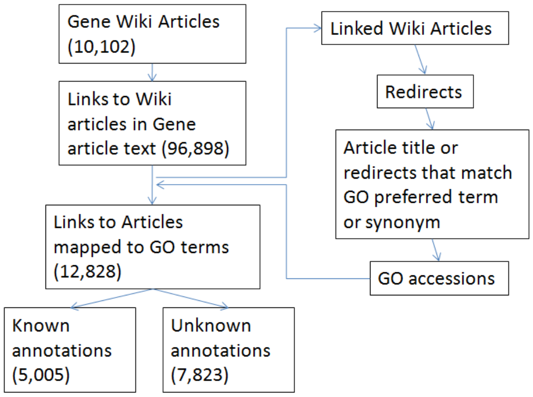
Gene Wiki Articles (10,102)
Links to Wiki articles in Gene article text (96,898)
Links to Articles mapped to GO terms (12,828)
Known annotations (5,005)
Unknown annotations (7,823)
Linked Wiki Articles
Redirects
Article title or redirects that match GO preferred term or synonym
GO accessions

Figure 3. GO annotation mining process.

Table 1 illustrates the influence of reasoning (of inferring that a gene bears the annotations of the terms that are broader than the terms that it is directly annotated with). Without the application of reasoning, the number of known annotations decreases dramatically - from 5,005 (39%) to 1139 (9%). This suggests that wikilinks tend to match less specific terms than are typically used in gene annotations. This makes sense when viewed from the perspective that Wikipedia is meant to be a general-purpose encyclopedia while the Gene Ontology is meant to enable extremely precise scientific description.

	Total candidates that match GOA annotations	Total candidates that do not match GOA annotations
without reasoning	1,139	11,689
with reasoning	5,005	7,823

Table 1: Comparison of candidate GO annotations to known GO annotations with and without reasoning.

Aside from the application of reasoning to GO annotations, the other key variable in this analysis was the choice of mapping strategy. In the results presented above, the loosest mapping was employed. To achieve higher precision in the mappings, we tested two additional strategies - 'medium' and 'tight'. The medium level mapping included all matches with the exception of 'broader' and 'related' GO synonyms. The tight mappings again removed 'broader' and 'related' synonyms but also removed all redirects - only exact matches to the article title were kept. Table 2 illustrates the reduction of matches with increasing stringency from 12,828 for the loosest strategy, to 11,125 for the medium level, and 3,747 for the tightest level. The relative proportions of the number of matches to existing GO annotations did not vary substantially at 39%, 42% and 34% respectively.

Mapping stringency	Total candidates	Total matching GOA	Total novel candidates
Loose	12,828	5,005	7,823
Medium	11,125	4,640	6,485
Tight	3,747	1,279	2,468

Table 2: Comparison of the results of different article-to-GO mapping strategies.

Rating potential new GO annotations

If the predicted annotations generated by this approach are to enter into public databases, a high level of confidence in them must be established - ideally through manual verification. In support of such downstream analysis, we evaluated one method for prioritizing candidate annotations that, like the Gene Wiki itself, takes advantage of community intelligence.

We hypothesized that the more frequently a gene name and a GO term appear together versus separately the more likely it is that the GO term might represent a good annotation for the gene. This 'semantic relatedness' of the gene and the GO concept can be estimated simply by counting the number of hits returned by a search engine for the terms individually and together. Using hit-count data acquired with the Yahoo BOSS search engine API, we calculated such a semantic relatedness score (known as the "Normalized Google Distance" [13]) for each candidate annotation and then used this score to rank all of the candidates.

To evaluate this ranking we treated the 1,139 exact matches to known annotations (see Table 1) as 'true positives' and built a Receiver Operator Curve to illustrate the ability of the ranking system to identify them. (Note that this evaluation treats the unknown, potential annotations as false positives though many of them may be true positives.) As Figure 4 shows, the ranking is substantially better than what we would expect by chance.

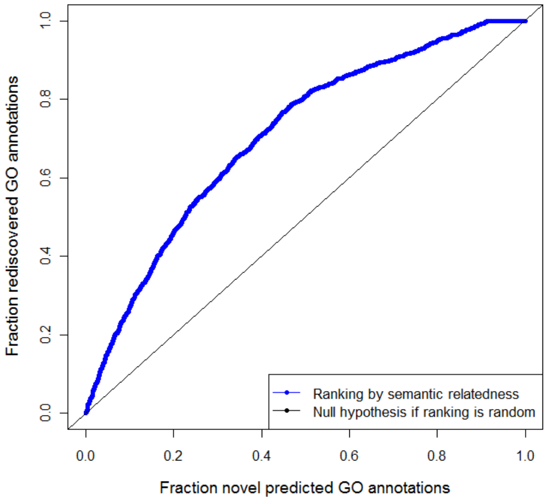

Figure 4: Performance of search engine-based semantic relatedness estimation for ranking candidate annotations.

Discussion

The continued activity of the thousands of volunteers that edit Gene Wiki articles is a testament to its success at harnessing collective intelligence for the purpose of generating *unstructured* biological information. This unstructured information, encapsulated in an expanding collection of free text, images and hyperlinks, is providing an increasingly useful resource for interested readers to learn about gene function. However, these unstructured data are not amenable to computational analyses.

In this paper, we describe our efforts to make the collective knowledge in the Gene Wiki as accessible for computation as it is for human comprehension. An essential step in this process is the translation of the Gene Wiki into a semantic network of concepts whose nodes and edges bear unambiguous meaning.

Here, we have demonstrated one technique with which it is possible to infer structured data from the articles in Wikipedia related to human genes. While we have identified interesting candidate gene annotations, we have barely scratched the surface of the knowledge that resides within the Gene Wiki. Even if we just consider the links - ignoring for the moment the majority of the knowledge that remains locked in unlinked text - we were only able to find matches in the GO for 8% of the linked titles.

To gauge what other forms of structured data may be present in these links we executed an exploratory analysis using the Unified Medical Language System (UMLS) knowledge server [14]. Using the 'exact match' option of their term identification service we found 12,074 concept matches for pages linked to from Gene Wiki articles. As Figure 3 illustrates, there are many links to concepts outside the purview of the Gene Ontology such as other genes, drugs, diseases, and body parts.

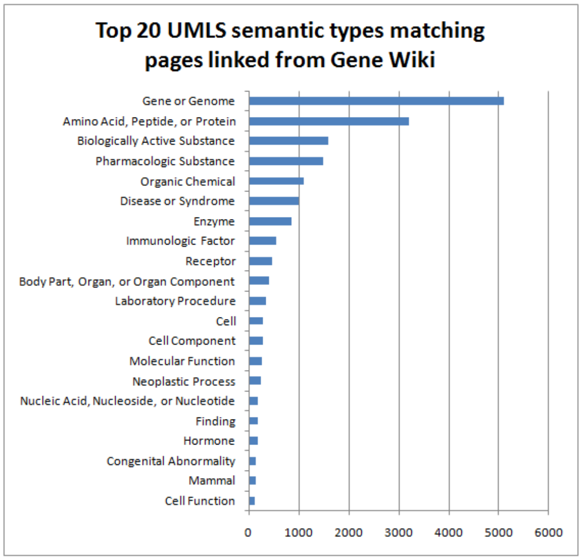

Figure 5: UMLS semantic types for Concepts matching pages linked from Gene Wiki articles.

Looking forward, there is clearly a large and diverse body of data trapped in the links on the Gene Wiki. There is also an even larger body of data residing in the unlinked text of these articles. The key to unlocking this potential lies in defining robust, high-throughput processes to take the next step: to verify the predictions.

References

1. Ashburner M, Ball C, Blake J, et al.: **Gene Ontology: tool for the unification of biology**. *Nat Genet* 2000, **25**:25-29.

2. Camon E, Barrell D, Brooksbank C, Magrane M, Apweiler R: **The Gene Ontology Annotation (GOA) Project-Application of GO in SWISS-PROT, TrEMBL and InterPro.** *Comparative and Functional Genomics* 2003, **4**:71-74.

3. Huang DW, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.** *Nature protocols* 2009, **4**:44-57.

4. Su AI, Hogenesch JB: **Power-law-like distributions in biomedical publications and research funding.** *Genome biology* 2007, **8**:404.

5. Hirschman J, Berardini TZ, Drabkin HJ, Howe D: **A MOD(ern) perspective on literature curation.** *Molecular genetics and genomics : MGG* 2010, **283**:415-25.

6. Camon EB, Barrell DG, Dimmer EC, et al.: **An evaluation of GO annotation retrieval for BioCreAtIvE and GOA.** *BMC bioinformatics [electronic resource].* 2005, **6 Suppl 1**:S17.

7. Howe D, Costanzo M, Fey P, et al.: **Big data: The future of biocuration.** *Nature* 2008, **455**:47-50.

8. Huss JW, Orozco C, Goodale J, et al.: **A gene wiki for community annotation of gene function.** *PLoS biology* 2008, **6**:e175.

9. Huss JW, Lindenbaum P, Martone M, et al.: **The Gene Wiki: community intelligence applied to human gene annotation.** *Nucleic acids research* 2010, **38**:D633-9.

10. Wishart DS, Knox C, Guo AC, et al.: **DrugBank: a knowledgebase for drugs, drug actions and drug targets.** *Nucleic acids research* 2008, **36**:D901-6.

11. Wu C, Orozco C, Boyer J, et al.: **BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources.** *Genome biology* 2009, **10**:R130.

12. **mediawiki api**. /
http://en.wikipedia.org/w/api.php

13. Cilibrasi R, Vitanyi P: **The Google Similarity Distance**. *IEEE Transactions on Knowledge and Data Engineering* 2007, **19**:370-383.

14. Bodenreider O: **The Unified Medical Language System (UMLS): integrating biomedical terminology.** *Nucleic acids research* 2004, **32**:267-270.