Chapter 6

Text Mining to Support Gene Ontology Curation and Vice Versa

Patrick Ruch

Abstract

In this chapter, we explain how text mining can support the curation of molecular biology databases dealing with protein functions. We also show how curated data can play a disruptive role in the developments of text mining methods. We review a decade of efforts to improve the automatic assignment of Gene Ontology (GO) descriptors, the reference ontology for the characterization of genes and gene products. To illustrate the high potential of this approach, we compare the performances of an automatic text categorizer and show a large improvement of +225 % in both precision and recall on benchmarked data. We argue that automatic text categorization functions can ultimately be embedded into a Question-Answering (QA) system to answer questions related to protein functions. Because GO descriptors can be relatively long and specific, traditional QA systems cannot answer such questions. A new type of QA system, so-called Deep QA which uses machine learning methods trained with curated contents, is thus emerging. Finally, future advances of text mining instruments are directly dependent on the availability of high-quality annotated contents at every curation step. Databases workflows must start recording explicitly all the data they curate and ideally also some of the data they do not curate.

Key words Automatic text categorization, Gene ontology, Data curation, Databases, Data stewardship, Information storage and retrieval

1 Introduction

This chapter attempts to concisely describes the role played by text mining in literature-based curation tasks concerned with the description of protein functions. More specifically, the chapter explores the relationships between the Gene Ontology (GO) and Text Mining.

Subheading 2 introduces the reader to basic concepts of text mining applied to biology. For a more general introduction, the reader may refer to a recent review paper by Zheng et al. [1].

Subheading 3 presents the text mining methods developed to support the assignment of GO descriptors to a gene or a gene product based on the content of some published articles. The section also introduces the methodological framework needed to assess the performances of these systems called automatic text categorizers.

Christophe Dessimoz and Nives Škunca (eds.), *The Gene Ontology Handbook*, Methods in Molecular Biology, vol. 1446,
DOI 10.1007/978-1-4939-3743-1_6,

Subheading 4 presents the evolution of results obtained today by GOCat, a GO categorizer, which participated in several BioCreative campaigns.

Finally, Subheading 5 discusses an inverted perspective and shows how GO categorization systems are foundational of a new type of text mining applications, so-called Deep Question-Answering (QA). Given a question, Deep QA engines are able to find answers, which are literally found in no corpus.

Subheading 6 concludes and emphasizes the responsibility of national and international research infrastructures, in establishing virtuous relationships between text mining services and curated databases.

2 State of the Art

This section presents the state of the art in text mining from the point of view of a biocurator, i.e., a person who is maintaining the knowledge stored in gene and protein databases.

2.1 Curation Tasks

In modern molecular biology databases, such as UniProt [2], the content is authored by biologists called biocurators. The work performed by these biologists when they curate a gene or a gene product encompasses a relatively complex set of individual and collaborative tasks [3]. We can separate these tasks into two subsets: sequence annotation—any information added to the sequence such as the existence of isoforms—and functional annotation—any information about the role of the gene or gene product in a given pathway or phenotype. Such a separation is partially artificial because a functional annotation can also establish a relationship between the role of a protein and some sequence positions but it is didactically convenient to adopt such a view.

The primary source of knowledge for genomics and proteomics is the research literature. In the context of biocuration, text mining can be defined as a process aimed at supporting biocurators when they search, read, identify entities, and store the resulting structured knowledge. The developments of benchmarks and metrics to evaluate how automatic text mining systems can help performing these tasks are thus crucial.

BioCreative is a community initiative to periodically evaluates the advances in text mining for biology and biocuration.[1] The forum explored a wide span of tasks with emphasis on named-entity recognition. Named-entity recognition covers a large set of methods that seek to locate and classify textual elements into predefined categories such as the names of persons, organizations, locations, genes, diseases, chemical compounds, etc. Thus, querying PubMed

[1] http://biocreative.sourceforge.net/

Table 1
Comparative curation steps supported by text mining

	[4]	[5]
1	Retrieval	Collection
2	Selection	Triage
3	Reading/Passage retrieval	
4	Entity extraction	Entity indexing
5	Entity normalization	
6		Relationship + evidence annotation
7		Extraction of evidences, e.g., images
8	Feed-back	
9		Check of records

Reference [4] describes the curation task as an iterative process (#8 Feed-back) whereas [6] describes it as a linear process (ending with #9 Check of records). Both descriptions are however consistent. Thus, it is possible to align steps #1, #2, and #4 in Table 1. Step #6 is optional in [4] as the process is regarded as an iterative process. This step is an "intelligent" follow up of the curation task, where already annotated functions/properties should receive less priority in the next Retrieval step. In contrast, steps #3 "Reading/passage retrieval" and #6 "Feed-back" is missed by [6], while the "Extraction of evidences" & "Check of record" is missed by [4] Step #5, i.e., the assignment of unique identifiers to descriptors, in [4] is implicit in step #4 of [6]

with the keywords "biocreative" and "information retrieval" returns 8 PMIDs, whereas 32 PMIDs are returned for the keywords "biocreative" and "named entity" [18th of November 2015].

The general workflow of a curation process supported by text mining instruments commonly comprises 6–9 steps as displayed in Table 1, which is a synthesis inspired by both [6] and [4].

Search is often the first step of a text mining pipeline, although information retrieval has received little attention from bioinformaticians active in Text Mining. Fortunately, information retrieval has been explored by other scientific communities and in particular by information scientists via the TREC (Text Retrieval Conferences) evaluation campaigns, *see* ref. 7 for a general introduction. From 2002 to 2015, molecular biology [8], clinical decision-support [9] and chemistry-related information retrieval [10] challenges have been explored by TREC. Interestingly, large-scale information retrieval studies have consistently shown that named-entity recognition has no or little impact on search effectiveness [11, 12].

2.2 From Basic Search to More Advanced Textual Mining

Beyond information retrieval, more elaborated mining instruments can then be derived. Thus, search engines, which return documents or pointers to documents, are often powered with passage retrieval skills [7], i.e., the ability to highlight a particular sentence, a few phrases, or even a few keywords in a given context.

The enriched representation can help the end-user to decide upon the relevance of the document. If for MEDLINE records, such passage retrieval functionalities are not crucial because an abstract is short enough to be rapidly read by a human, passage retrieval tools become necessary when the search is performed on a collection of full-text articles like for instance in PubMed Central. Within a full-text article, the ability to identify the section where a given set of keywords can be very useful as matching the relevant keywords in a "background" section has a different value than matching them in a "results" section. The latter is likely to be a new statement while the former is likely to be regarded as a well-established knowledge.

2.3 Named-Entity Recognition

Unlike in other scientific or technical fields (finance, high energy physics, etc.), in the biomedical domain, named-entity recognition covers a very large set of entities. Such a richness is well expressed by the content of modern biological databases. Text Mining studies have been published for many of those curation needs, including sequence curation and identification of polymorphisms [13], posttranslational modifications [14], interactions with gene products or metabolites [15], etc. In this context, most studies attempted to develop instruments likely to address a particular set of annotation dimensions, serving the needs of a particular molecular biology database. The focus in such studies is often to design a Graphic User Interfaces and to simplify the curation work by highlighting specific concepts in a dedicated tool [16]. While most of these systems seem exploratory studies, some seem deeply integrated in the curation workflow, as shown by the OntoMate tool designed by the Rat Genome Database [17], the STRING DB for protein–protein interactions or the BioEditor of neXtProt [18].

From an evaluation perspective, the idea is to detect the beginning and the end of an entity and to assign a semantic type to this string. Thus in named-entity recognition, we assume that entity components are textually contiguous. Inherited from early corpus works on information extraction and computational linguistics [19], the goal is to assign a unique semantic category—e.g., Time, Location, and Person—to a string in a text [20].

Semantic categories are virtually infinite but some entities received more attention. Gene, gene products, proteins, species [21, 22], and more recently chemical compounds were significantly more studied than for instance organs, tissues, cell types, cell anatomy, molecular functions, symptoms, or phenotypes [23].

The initial works dealing with the recognition of GO entities were disappointing (Subheading 3.2), which may explain part of the reluctance to address these challenges. We see here one important limitation of named entities: it is easy to detect a one or two words terms into a document, while the recognition of a protein function does require a "deeper" understanding or combination of

biological concepts. Indeed a complex GO concept is likely to combine subconcepts belonging to various semantic types, including small molecules, atoms, protein families, as well as biological processes, molecular functions, and cell locations.

2.4 Normalization and Relationship Extraction

In order to compensate for the limitations of named-entity recognition frameworks, two more complementary approaches have been proposed: entity normalization and information (or relationship) extraction.

Normalization can be defined as the process by which a unique semantic identifier is assigned to the recognized entities [24]. The identifiers are available in different resources such as several onto-terminologies or knowledge bases. The assignment of unique identifiers can be relatively difficult in practice due to a linguistic phenomenon called lexical ambiguity. Many strings are lexically ambiguous and therefore can receive more than one identifier depending on the context (e.g., *HIV* could be a disease or a virus). The difficulty is amplified in cascaded lexical ambiguities. Many entities require the extraction of other entities to receive an unambiguous identifier. For instance, the assignment of an accession number to a protein may depend on the recognition of an organism or a cell line somewhere else in the text.

Further, the extraction of relationships requires the recognition of the specific entities, which can be as various as a location, an interaction (binding, coexpression, etc.) [25], an etiology or a temporal marker (cause, trigger, simultaneity, etc.) [26]. For some information extraction tasks such as protein–protein interactions, the normalization and relationship extraction may require first the proper identification of other entities such as the experimental methods (e.g., yeast 2-hybrid) used to generate the prediction. Furthermore, additional information items may be provided such as the scale of the interaction or the confidence in the interaction [27].

To identify GO terms, named-entity recognition and information extraction is insufficient due to two main difficulties: first, the difficulty of defining all (or most) strings describing a given concept; second, the difficulty of defining the string boundaries of a given concept. The parsing of texts to identify GO functions and how they are linked with a given protein demands the development of specific methods.

2.5 Automatic Text Categorization

Automatic text categorization (ATC) can be defined as the assignment of any class or category to any text content. The interested reader can refer to [28], where the author provides a comprehensive introduction to ATC, with a focus on machine learning methods.

In both ATC and in Information Retrieval, documents are regarded as "bag-of-words." Such a representation is an approximation but it is a powerful and productive simplification. From this bag, where all entities and relationships are treated as flat and

independent data, ATC attempts to assign a set of unambiguous descriptors. The set of descriptors can be binary as in triage tasks, where documents can be either classified as relevant for curation or irrelevant, or it can be multiclass. The scale of the problem is one parameter of the model. In some situations, ATC systems do not need to provide a clear split between relevant and irrelevant categories. In particular, when a human is in the loop to control the final descriptor assignment step, ATC systems can provide a ranked list of descriptors, where each rank expresses the confidence score of the ATC system. ATC systems and search engines share here a second common point: compared to named-entity recognition, which is normally not interactive, ATC and Information Retrieval are well suited for human–computer interactions.

3 Methods

With over 40,000 terms—and many more if we account for synonyms—assigning a GO descriptor to a protein based on some published document is formally known as a large multiclass classification problem.

3.1 Automatic Text Categorization

The two basic approaches to solve the GO assignment problem are the following: (1) exploit the lexical similarity between a text and a GO term and its synonyms [29]; (2) use some existing database to train a classifier likely to infer associations beyond string matching. The second approach uses any scalable machine learning techniques to generate a model trained on the Gene Ontology Annotation (GOA) database. Several machine learning strategies have been used but the trade-off between effectiveness, efficiency, and scalability often converges toward an approach called k-Nearest Neighbors (k-NN); *see* also ref. 30.

3.2 Lexical Approaches

Lexical approaches for ATC exploit the similarities between the content of a text and the content of a GO term and its related synonyms [31]. Additional information can be taken into account to augment the categorization power such as the definitions of the GO terms. The ranking functions take into account the frequency of words, their specificity (measured by the "inverse document frequency," the inverse of how many documents contain the word), as well as various positional information (e.g., word order); *see* ref. 32 for a detailed description.

The task is extremely challenging if we consider that some GO terms contain a dozen words, which makes those terms virtually unmatchable in any textual repository. The results of the first BioCreative competition, which was addressing this challenge, were therefore disappointing. The best "high-precision" system achieved an 80 % precision but this system covered less than 20 % of

the test sample. In contrast, with a recall close to 80%, the best "high-recall" systems were able to obtain an average precision of 20–30% [33]. At that time, over 10 years ago, such a complex task was consequently regarded are practically out of reach for machines.

3.3 k-Nearest Neighbors

The principle of a k-NN is the following: for an instance X to be classified, the system computes a similarity measure between X and some annotated instances. In a GO categorizer, an instance is typically a PMID annotated with some GO descriptors. Instances on the top of the list are assumed "similar" to X. Experimentally, the value of k must be determined, where k is the number of similar instances (or neighbors), which should be taken into account to assign one or several categories to X.

When considering a full-text article, a particular section in this article, or even a MEDLINE record, it is possible to compute a distance between this section and similar articles in the GOA database because in the curated section of GOA, many GO descriptors are associated with a PMID—those marked up with an EXP evidence code [34]. The computation of the distance between two arbitrary texts can be more or less complex—starting with counting how many words they share—and the determination of the *k* parameters can also be dependent on different empirical features (number of documents in the collection, average size a document, etc.) but the approach is both effective and computationally simple [7]. Moreover, the ability to index a priori all the curated instances makes possible to compute distances efficiently.

The effectiveness of such machine learning algorithms is directly dependent on the volume of curated data. Surprisingly GO categorizers seem not affected by any concept drift, which affects database and data-driven approaches in general. Even old data, i.e., protein annotated with an early version of the GO, seem useful for k-NN approaches [35]. To give a concrete example, consider proteins curated in 2005 with a version of the Gene Ontology and a MEDLINE reports available at that time: it is difficult to understand why a model containing mainly annotations from 2010 to 2014 would outperform a model containing data from 2003 to 2007 using data exactly centered on 2005. While the GO itself has been expanded by at least a factor 4 in the past decade, the consistency of the curation model has remained remarkably stable.

3.4 Properties of Lexical and k-NN Categorizers

In Fig. 1, we show an example output of GOCat [35], which is maintained by my group at the SIB Swiss Institute of Bioinformatics. The same abstract is processed by GOCat using two different types of classification methods: a lexical approach and a k-NN.

In this example, the title of an article ([36]; "Modulation by copper of p53 conformation and sequence-specific DNA binding: role for Cu(II)/Cu(I) redox mechanism") is used as input to contrast the behavior of the two approaches: This reference is used in

#	Score	GO ID	Name
1	1.00	GO:0003677	DNA binding +/-
2	0.77	GO:0043565	sequence-specific DNA binding (synonym sequence specific dna binding) +/-
3	0.31	GO:0070712	RNA cytidine-uridine insertion (synonym rna cu insertion) +/-
4	0.22	GO:0071103	DNA conformation change (synonym dna conformation modification) +/-
5	0.22	GO:0005488	binding +/-
6	0.21	GO:0051982	copper-nicotianamine transmembrane transporter activity (synonym cu-na chelate transporter activity)+/-
7	0.21	GO:0004008	copper-exporting ATPase activity (synonym cu(2+)-exporting atpase activity) +/-
8	0.19	GO:0009455	redox taxis +/-
9	0.19	GO:0016491	oxidoreductase activity (synonym redox activity) +/-
10	0.19	GO:0051776	detection of redox state (synonym redox sensing) +/-
11	0.17	GO:0002039	p53 binding +/-
12	0.16	GO:0005507	copper ion binding (synonym copper binding) +/-
13	0.15	GO:0000393	spliceosomal conformational changes to generate catalytic conformation +/-

#	Score	GO ID	Name
1	1.00	GO:0005507	copper ion binding +/-
2	0.42	GO:0046688	response to copper ion +/-
3	0.22	GO:0008270	zinc ion binding +/-
4	0.21	GO:0003677	DNA binding +/-
5	0.19	GO:0004784	superoxide dismutase activity
6	0.16	GO:0006878	cellular copper ion homeostasis +/-
7	0.13	GO:0035434	copper ion transmembrane transport +/-
8	0.13	GO:0015677	copper ion import +/-
9	0.13	GO:0071280	cellular response to copper ion +/-
10	0.12	GO:0005375	copper ion transmembrane transporter activity +/-
11	0.12	GO:0005886	Plasma membrane
12	0.11	GO:0016531	copper chaperone activity +/-
13	0.10	GO:0055114	oxidation-reduction process
14	0.10	GO:0019430	removal of superoxide radicals
15	0.10	GO:0046914	transition metal ion binding +/-
16	0.10	GO:0006825	copper ion transport +/-
17	0.09	GO:0006801	superoxide metabolic process
18	0.09	GO:0010273	detoxification of copper ion +/-

Fig. 1 Comparative outputs of lexical vs. k-NN versions of GOCat

UniProt to support the assignment of the "copper ion binding" descriptor to *p53*. We see that the lexical system (left panel) is able to assign the descriptor at rank #12, while the k-NN system (right panel) provides the descriptor in position #1.

Finally, we see how both categorizers are also flexible instruments as they basically learn to rank a set of a priori categories. Such systems can easily be used as fully automatic systems—thus taking into account only the top N returned descriptors by setting up an empirical threshold score—or as interactive systems able to display dozens of descriptors including many irrelevant ones, which then can be discarded by the curator.

Today, GO k-NN categorizers do outperform lexical categorizers; however, the behavior of the two systems is complementary. While the latter is potentially able to assign a GO descriptor, which has rarely or never been used to generate an annotation, the former is directly dependent on the quantity of [GO; PMID] pairs available in GOA.

3.5 Inter-annotator Agreement

An important parameter when assessing text mining tools is the development of a ground truth or gold standard. Thus, typically for GO annotation, we assume that the content of curated

databases is the absolute reference. This assumption is acceptable from a methodological perspective, as text mining systems need such benchmarks. However, it is worth observing that two curators would not absolutely agree when they assign descriptors, which means that a 100% precision is purely theoretical. Thus, Camon et al. [37] reports that two GO annotators would have an agreement score of about 39–43%. The upper score is achieved when we consider that the assignment of a generic concept instead of a more specific one (children) is counted as an agreement.

4 Today's Performances

Today, GOCat is able to assign a correct descriptor to a given MEDLINE record two times out of three using the BioCreative I benchmark [35], which makes it useful to support functional annotation. Another type of systems, can be used to support complementary tasks of literature exploration (GoPubMed: [38]) or named-entity recognition [39]. While GOCat attempts to assign GO descriptors to any input with the objective to help curating the content of the input, GoPubMed provides a set of facets (Gene Ontology or Medical Subject Headings) to navigate the result of a query submitted to PubMed.

It is worth observing that GO categorizers work best when they assume that the curator is involved in selecting the input papers (performing a triage or selection task as described in Table 1). Such a setting, inherited from the BioCreative competitions, [33, 40] is questionable for at least two reasons: (1) Curators read full-text articles and not only the abstracts—captions and legends seem especially important; (2) The triage task, i.e., the ability to select an article as relevant for curation, could mostly be performed by a machine, provided that fair training data are available. In 2013, the campaign of BioCreative, under the responsibility of the NCBI, revisited the task [41]. The competitors were provided with full-text articles and they were asked not only to return GO descriptors but also to select a subset of sentences. The evaluation was thus more transparent. A small but high-quality annotated sample of full-text papers was provided [42].

The main results from these experiments are the following; *see* ref. 41 for a complete report describing the competition metrics as well as the different systems participating in the challenge. First, the precision of categorization systems improved by about +225% compared to BioCreative 1. Second, the ability to detect all relevant sentences seems less important than being able to select a few high content-bearing sentences. Thus GOCat achieved very competitive results for both recall and precision in GO assignment task, but interestingly the system performed relatively poorly when focusing on the recall of the sentence selection task, *see* Figs. 2 and 3 for

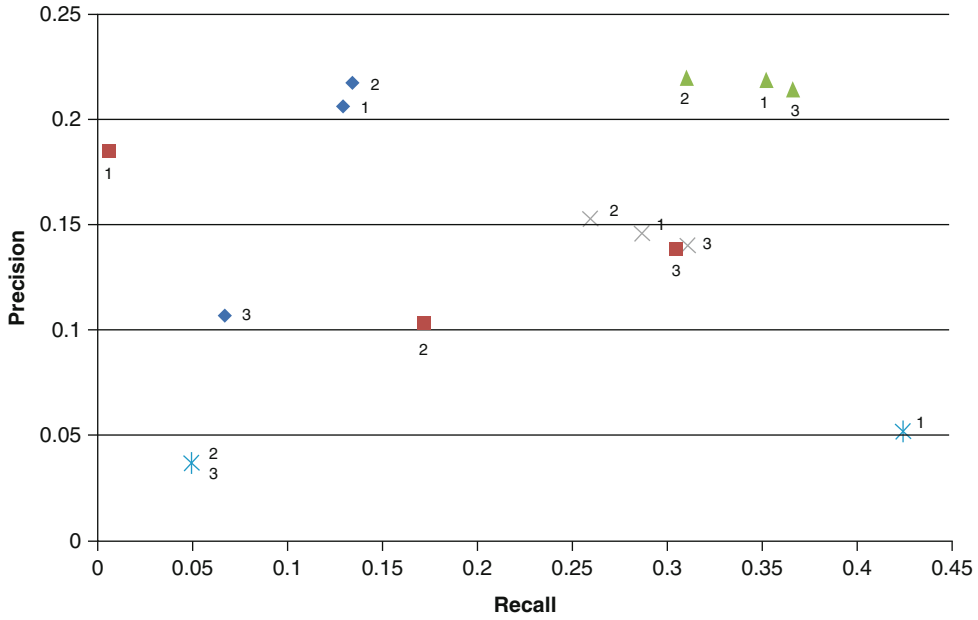

Fig. 2 Relative performance of the sentence triage module of GOCat4FT (GOCat for full-text, *blue diamond*) at the official BioCreative IV competition. Courtesy of Zhiyong Lu, National Institute of Health, National Library of Medicine

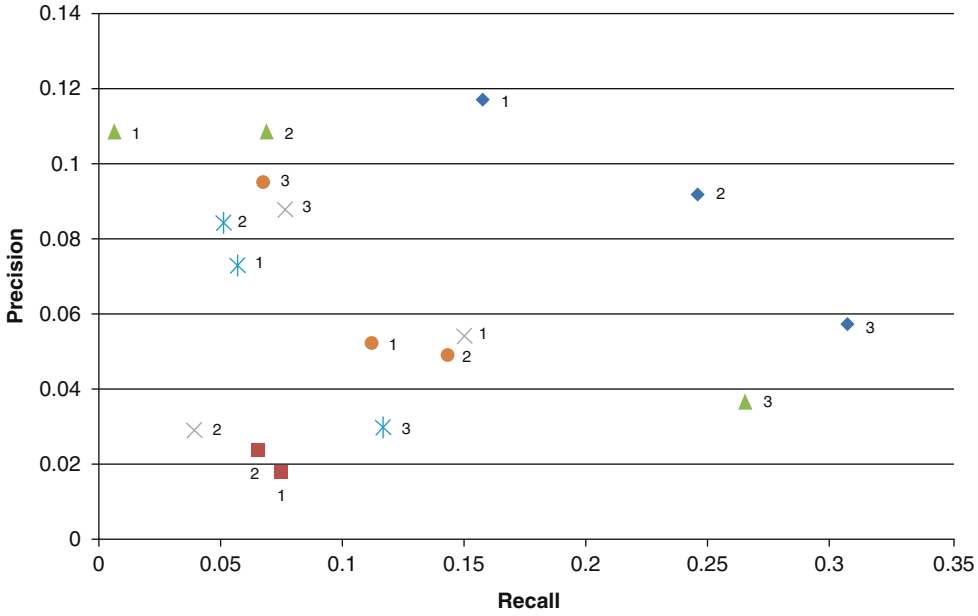

Fig. 3 Relative performance of GOCat4FT (*blue diamond*) when fed with the sentences selected by the three sentence triage systems evaluated in Fig. 2

comparison. We see that two of the sentence ranking systems developed for the BioCreative IV competition (orange dots) outperform other systems in precision but not in recall. References [40, 43] conclude from these experiments that the content in a full-text article is so (highly) redundant that a weak recall is acceptable provided that the few selected sentences have good precision. The few high relevance sentences selected by GOCat4FT (Gene Ontology Categorizer for Full Text) are sufficient to obtain highly competitive results when GO descriptors are assigned by GOCat (orange dots) regarding both recall and precision as the three official runs submitted by SIB Text Mining significantly outperforms other systems. Such a redundancy phenomenon is probably found not only in full-text contents but more generally in the whole literature.

Together with GO and GOA, which was used by most participants in the competition, some online databases seem particularly valuable to help assigning GO descriptors. Thus, Luu et al. [44] uses the cross-product databases [45] with some effectiveness.

5 Discussion

Although a fraction of it is likely to be sufficient to obtain the top-ranked GO descriptors, the results reported in the previous section are obtained by using only 10–20% of the content of an article. This suggests that 80–90% of what is published is unnecessary from an information-theoretic perspective.

5.1 Information Redundancy and Curation-Driven Data Stewardship

New and informative statements are rare in general. They are moreover buried in a mass of relatively redundant and poorly content-bearing claims. It has been shown that the density and precision of information in abstracts is higher [5, 46] than in full-text reports while the level of redundancy across papers and abstracts is probably relatively high as well.

We understand that the separation of valuable scientific statements is labor intensive for curators. This filtering effort is complicated within an article but also between articles at retrieval time. We argue that such task could be performed by machines provided that high-quality training data are available. The training data needed by text mining systems are unfortunately lost during the curation process. Indeed, the separation between useful and useless materials (e.g., PMIDs and sentences) is performed—but not recorded—by the curator during the annotation process but they are unfortunately not stored in databases.

In some cases, the separation is explicit, in other cases, it is implicit but the key point is that a mass of information is definitely lost with no possible recovery. The capture of the output of the selection process—at least for the positive content but ideally also for a fraction of the negative content—is a minimal requirement to

improve text mining methods. The expected impact of the implementation of such simple data stewardship recommendation is likely a game changer for text mining far beyond any hypothetical technological advances.

5.2 Assigning Unmatchable GO Descriptors: Toward Deep QA

Some GO concepts describe entities which are so specific that they can hardly be found anywhere. This has several consequences. Traditional QA systems were recently made popular to answer Jeopardy-like questions with entities as various as politicians, town, plants, countries, songs, etc., *see* ref. 47. In the biomedical field, Bauer and Berleant [48] compare four systems, looking at their ergonomics. With a precision in the range of 70–80% [49], these systems perform relatively well. However, none of these systems is able to answer questions about functional proteomics. Indeed, how can a text mining system find an answer if such an answer is not likely to be found on Earth in any corpus of book, article, or patent? The ability to accurately process questions, such as *what molecular functions are associated with tp53* requires to supply answers, such as "RNA polymerase II transcription regulatory region sequence-specific DNA binding transcription factor activity involved in positive regulation of transcription" and only GO categorizers are likely to automatically generate such an answer.

We may think that such complex concepts could be made simpler by splitting the concept into subconcepts, using clinical terminological resources such as SNOMED CT [50, 51] or ICD-10 [52], *see* also Chap. 20 [53]. That might be correct in some rare cases but in general, complex systems tend to be more accurately described using complex concepts. The post-coordination methods explored elsewhere remain effective to perform analytical tasks but they make generative tasks very challenging [52]. Post-coordination is useful to search a database or a digital library because search tasks assume that documents are "bag of words" and they ignore the relationships between these words. However, other tasks such as QA or curation do require to be able to meaningfully combine concepts. In this context, the availability of a pre-computed list of concepts or controlled vocabulary is extremely useful to avoid generating ill-formed entities.

Answering functional omics questions is truly original: it requires the elaboration of a new type of QA engines such as the DeepQA4GO engine [54]. For GO-type of answers, DeepQA4GO is able to answer the expected GO descriptors about two times out of three, compared to one time out of three for traditional systems. We propose to call these new emerging systems: Deep QA engines. Deep QA, like traditional QA engines are able to screen through millions of documents, but since no corpus contain the expected answers, Deep QA is needed to exploit curated biological databases in order to generate useful candidate answers for curators.

6 Conclusion

While the chapter started with introducing the reader to how text mining can support database annotation, the conclusion is that next generation text mining systems will be supported by curated databases. The key challenges have moved from the design of text mining systems to the design of text mining systems able to capitalize on the availability of curated databases. Future advances in text mining to support biocuration and biomedical knowledge discovery are largely in the hands of database providers. Databases workflows must start recording explicitly all the data they curate and ideally also some of the data they do not curate.

In parallel, the accuracy of text mining system to support GO annotation has improved massively from 20 to 65 % (+225 %) from 2005 to 2015. With almost 10,000 queries a month, a tool like GOCat is useful in order to provide a basic functional annotation of protein with unknown and/or uncurated functions [55] as exemplified by the large-scale usage of GOCat by the COMBREX database [56, 57]. However, the integration of text mining support systems into curation workflows remains challenging. As often stated, curation is accurate but does not scale while text mining is not accurate but scales. National and international Research Infrastructures should play a central role to promote optimal data stewardship practices across the databases they support. Similarly, innovative curation models should emerge by combining the quality and richness of curation workflows, more cost-effective crowd-based triage, and the scalability of text mining instruments [58].

Funding Open Access charges were funded by the University College London Library, the Swiss Institute of Bioinformatics, the Agassiz Foundation, and the Foundation for the University of Lausanne.

References

1. Zeng Z, Shi H, Wu Y, Hong Z (2015) Survey of natural language processing techniques in bioinformatics. Comput Math Methods Med 2015:674296. doi:10.1155/2015/674296, Epub 2015 Oct 7
2. Dimmer EC, Huntley RP, Alam-Faruque Y, Sawford T, O'Donovan C, Martin MJ, Bely B, Browne P, Mun Chan W, Eberhardt R, Gardner M, Laiho K, Legge D, Magrane M, Pichler K, Poggioli D, Sehra H, Auchincloss A, Axelsen K, Blatter MC, Boutet E, Braconi-Quintaje S, Breuza L, Bridge A, Coudert E, Estreicher A, Famiglietti L, Ferro-Rojas S, Feuermann M, Gos A, Gruaz-Gumowski N, Hinz U, Hulo C, James J, Jimenez S, Jungo F, Keller G, Lemercier P, Lieberherr D, Masson P, Moinat M, Pedruzzi I, Poux S, Rivoire C, Roechert B, Schneider M, Stutz A, Sundaram S, Tognolli M, Bougueleret L, Argoud-Puy G, Cusin I, Duek-Roggli P, Xenarios I, Apweiler R (2012) The UniProt-GO Annotation database in 2011. Nucleic Acids Res 40(Database issue):D565–D570. doi:10.1093/nar/gkr1048, Epub 2011 Nov 28
3. Poux S, Magrane M, Arighi CN, Bridge A, O'Donovan C, Laiho K; UniProt Consortium (2014) Expert curation in UniProtKB: a case study on dealing with conflicting and erroneous data. Database (Oxford):bau016. doi:10.1093/database/bau016
4. Vishnyakova D, Emilie Pasche E, Patrick Ruch P (2012) Using binary classification to prioritize and curate articles for the Comparative Toxicogenomics Database. Database 2012
5. Lin J (2009) Is searching full text more effective than searching abstracts? BMC Bioinformatics 10:46. doi:10.1186/1471-2105-10-46
6. Lu Z, Hirschman L. Biocuration workflows and text mining: overview of the BioCreative 2012 Workshop Track II. Database 2012
7. Singhal A (2001) Modern information retrieval: a brief overview. IEEE Data Eng Bull 24:35–43
8. Hersh W, Bhupatiraju RT, Corley S (2004) Enhancing access to the Bibliome: the TREC Genomics Track. Stud Health Technol Inform 107(Pt 2):773–777
9. Simpson MS, Voorhees ES, Hersh W (2014) Overview of the TREC 2014. Clinical Decision Support Track. TREC 2014
10. Lupu M, Huang J, Zhu J, Tait J (2009) TREC-CHEM: large scale chemical information retrieval evaluation at TREC. SIGIR Forum 43(2):63–70
11. Abdou S, Savoy J (2008) Searching in Medline: query expansion and manual indexing evaluation. Inf Process Manag 44(2):781–789
12. Pasche E, Gobeill J, Kreim O, Oezdemir-Zaech F, Vachon T, Lovis C, Ruch P (2014) Development and tuning of an original search engine for patent libraries in medicinal chemistry. BMC Bioinformatics 15(Suppl 1):S15
13. Yip YL, Lachenal N, Pillet V, Veuthey AL (2007) Retrieving mutation-specific information for human proteins in UniProt/Swiss-Prot Knowledgebase. J Bioinform Comput Biol 5(6):1215–1231
14. Veuthey AL, Bridge A, Gobeill J, Ruch P, McEntyre JR, Bougueleret L, Xenarios I (2013) Application of text-mining for updating protein post-translational modification annotation in UniProtKB. BMC Bioinformatics 14:104. doi:10.1186/1471-2105-14-104
15. Xu S, An X, Zhu L, Zhang Y, Zhang H (2015) A CRF-based system for recognizing chemical entity mentions (CEMs) in biomedical literature. J Cheminform 7(Suppl 1 Text mining for chemistry and the CHEMDNER track):S11. doi:10.1186/1758-2946-7-S1-S11. eCollection 2015
16. Dowell KG, McAndrews-Hill MS, Hill DP, Drabkin HJ, Blake JA (2009) Integrating text mining into the MGI biocuration workflow. Database (Oxford):bap019. Epub 2009 Nov 21
17. Liu W, Laulederkind SJ, Hayman GT, Wang SJ, Nigam R, Smith JR, De Pons J, Dwinell MR, Shimoyama M (2015) OntoMate: a text-mining tool aiding curation at the Rat Genome Database. Database (Oxford):bau129
18. SIB Swiss Institute of Bioinformatics Members (2015) The SIB Swiss Institute of Bioinformatics' resources: focus on curated databases. Nucleic Acids Res 44(D1):D27–D37
19. Black WJ, Gilardoni L, Dressel R, Rinaldi F (1997) Integrated text categorisation and information extraction using pattern matching and linguistic processing. RIAO
20. Chinchor N (1997) Overview of MUC-7. Message Understanding Conferences (MUC).
21. Hirschman L, Yeh A, Blaschke C, Valencia A (2005) Overview of BioCreAtIvE: critical assessment of information extraction for biology. BMC Bioinformatics 6(Suppl 1):S1
22. Smith L, Tanabe LK, Ando RJ, Kuo CJ, Chung IF, Hsu CN, Lin YS, Klinger R, Friedrich CM, Ganchev K, Torii M, Liu H, Haddow B, Struble CA, Povinelli RJ, Vlachos A, Baumgartner WA Jr, Hunter L, Carpenter B, Tsai RT, Dai HJ, Liu F, Chen Y, Sun C, Katrenko S, Adriaans P, Blaschke C, Torres R, Neves M, Nakov P, Divoli A, Maña-López M, Mata J, Wilbur WJ (2008) Overview of

BioCreative II gene mention recognition. Genome Biol 9(Suppl 2):S2

23. Tran LT, Divita G, Carter ME, Judd J, Samore MH, Gundlapalli AV (2015) Exploiting the UMLS Metathesaurus for extracting and categorizing concepts representing signs and symptoms to anatomically related organ systems. J Biomed Inform. pii: S1532-0464(15)00192-6. doi:10.1016/j.jbi.2015.08.024
24. Morgan AA, Lu Z, Wang X, Cohen AM, Fluck J, Ruch P, Divoli A, Fundel K, Leaman R, Hakenberg J, Sun C, Liu HH, Torres R, Krauthammer M, Lau WW, Liu H, Hsu CN, Schuemie M, Cohen KB, Hirschman L (2008) Overview of BioCreative II gene normalization. Genome Biol 9(Suppl 2):S3. doi:10.1186/gb-2008-9-s2-s3, Epub 2008 Sep 1
25. Bell L, Chowdhary R, Liu JS, Niu X, Zhang J (2011) Integrated bio-entity network: a system for biological knowledge discovery. PLoS One 6(6):e21474
26. Perfetto L, Briganti L, Calderone A, Perpetuini AC, Iannuccelli M, Langone F, Licata L, Marinkovic M, Mattioni A, Pavlidou T, Peluso D, Petrilli LL, Pirrò S, Posca D, Santonico E, Silvestri A, Spada F, Castagnoli L, Cesareni G (2015) SIGNOR: a database of causal relationships between biological entities. Nucleic Acids Res 44:D548–D554
27. Bastian FB, Chibucos MC, Gaudet P, Giglio M, Holliday GL, Huang H, Lewis SE, Niknejad A, Orchard S, Poux S, Skunca N, Robinson-Rechavi M (2015) The Confidence Information Ontology: a step towards a standard for asserting confidence in annotations. Database:bav043 doi:10.1093/database/bav043
28. Sebastiani F (2002) Machine learning in automated text categorization. ACM Comput Surv 34(1):1–47
29. Ruch P (2006) Automatic assignment of biomedical categories: toward a generic approach. Bioinformatics 22(6):658–664, Epub 2005 Nov 15
30. Lena PD, Domeniconi G, Margara L, Moro G (2015) GOTA: GO term annotation of biomedical literature. BMC Bioinformatics 16:346
31. Couto F, Silva M, Coutinho P (2005) FiGO: finding GO terms in unstructured text. BioCreative Workshop Proceedings
32. Ehrler F, Geissbühler A, Jimeno A, Ruch P (2005) Data-poor categorization and passage retrieval for gene ontology annotation in Swiss-Prot. BMC Bioinformatics 6(Suppl 1):S23, Epub 2005 May 24
33. Blaschke C, Leon E, Krallinger M, Valencia A (2005) Evaluation of BioCreAtIvE assessment of task 2. BMC Bioinformatics 6(Suppl 1):S16
34. Gaudet et al. Primer on gene ontology. GO handbook
35. Gobeill J, Pasche E, Vishnyakova D, Ruch P. Managing the data deluge: data-driven GO category assignment improves while complexity of functional annotation increases. Database 2013
36. Hainaut P, Rolley N, Davies M, Milner J (1995) Modulation by copper of p53 conformation and sequence-specific DNA binding: role for Cu(II)/Cu(I) redox mechanism. Oncogene 10(1):27–32
37. Camon EB, Barrell DG, Dimmer EC, Lee V, Magrane M, Maslen J, Binns D, Apweiler R (2005) An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. BMC Bioinformatics 6(Suppl 1):S17, Epub 2005 May 24
38. Doms A, Schroeder M (2005) GoPubMed: exploring PubMed with the Gene Ontology. Nucleic Acids Res 33(Web Server issue):W783–W786
39. Rebholz-Schuhmann D, Arregui M, Gaudan S, Kirsch H, Jimeno A (2008) Text processing through Web services: calling Whatizit. Bioinformatics 24(2):296–298
40. Yeh A, Morgan A, Colosimo M, Hirschman L (2005) BioCreAtIvE task 1A: gene mention finding evaluation. BMC Bioinformatics 6(Suppl 1):S2, Epub 2005 May 24
41. Mao Y, Van Auken K, Li D, Arighi CN, McQuilton P, G Hayman T, Tweedie S, Schaeffer ML, Laulederkind SJF, Wang S-J, Gobeill J, Ruch P, Luu AT, Kim J-J, Chiang J-H, De Chen Y, Yang C-J, Liu H, Zhu D, Li Y, Yu H, Emadzadeh E, Gonzalez G, Chen J-M, Dai H-J, Lu Z (2014). Overview of the gene ontology task at BioCreative IV. Database (Oxford) 2014
42. Van Auken K, Schaeffer ML, McQuilton P, Laulederkind SJ, Li D, Wang SJ, Hayman GT, Tweedie S, Arighi CN, Done J, Müller HM, Sternberg PW, Mao Y, Wei CH, Lu Z (2014) BC4GO: a full-text corpus for the BioCreative IV GO task. Database (Oxford). pii: bau074. doi:10.1093/database/bau074
43. Gobeill J, Pasche E, Dina V, Ruch P. (2014) Closing the loop: from paper to protein annotation using supervised Gene Ontology classification. Database:bau088
44. Luu AT, Kim JJ, Ng SK (2013) Gene ontology concept recognition using cross-products and statistical methods. In: The Fourth BioCreative Challenge Evaluation Workshop, vol. 1, Bethesda, MD, USA, pp 174–181
45. Mungall CJ, Bada M, Berardini TZ et al (2011) Cross-product extensions of the gene ontology. J Biomed Inform 44:80–86

46. Jimeno-Yepes AJ, Plaza L, Mork JG, Aronson AR, Díaz A (2013) MeSH indexing based on automatically generated summaries. BMC Bioinformatics 14:208
47. Ferrucci D (2012) Introduction to « This is Watson ». IBM J Res Dev 56(3.4):1–15
48. Bauer MA, Berleant D (2012) Usability survey of biomedical question answering systems. Hum Genomics 6:17
49. Gobeill J, Patsche E, Teodoro D, Veuthey AL, Lovis C, Ruch P. Question answering for biology and medicine. Information Technology and Applications in Biomedicine, 2009. ITAB 2009
50. Campbell WS, Campbell JR, West WW, McClay JC, Hinrichs SH (2014) Semantic analysis of SNOMED CT for a post-coordinated database of histopathology findings. J Am Med Inform Assoc 21(5): 885–892
51. Dolin RH, Spackman KA, Markwell D (2002) Selective retrieval of pre- and post-coordinated SNOMED concepts. Proc AMIA Symp:210–214
52. Baud RH, Rassinoux AM, Ruch P, Lovis C, Scherrer JR (1999) The power and limits of a rule-based morpho-semantic parser. Proc AMIA Symp:22–26
53. Denaxas SC (2016) Integrating bio-ontologies and controlled clinical terminologies: from base pairs to bedside phenotypes. In: Dessimoz C, Škunca N (eds) The gene ontology handbook. Methods in molecular biology, vol 1446. Humana Press. Chapter 20
54. Gobeill J, Gaudinat A, Pasche E, Vishnyakova D, Gaudet P, Bairoch A, Ruch P (2015) Deep question answering for protein annotation. Database (Oxford):bav081
55. Mills CL, Beuning PJ, Ondrechen MJ (2015) Biochemical functional predictions for protein structures of unknown or uncertain function. Comput Struct Biotechnol J 13:182–191
56. Anton BP, Chang YC, Brown P, Choi HP, Faller LL, Guleria J, Hu Z, Klitgord N, Levy-Moonshine A, Maksad A, Mazumdar V, McGettrick M, Osmani L, Pokrzywa R, Rachlin J, Swaminathan R, Allen B, Housman G, Monahan C, Rochussen K, Tao K, Bhagwat AS, Brenner SE, Columbus L, de Crécy-Lagard V, Ferguson D, Fomenkov A, Gadda G, Morgan RD, Osterman AL, Rodionov DA, Rodionova IA, Rudd KE, Söll D, Spain J, Xu SY, Bateman A, Blumenthal RM, Bollinger JM, Chang WS, Ferrer M, Friedberg I, Galperin MY, Gobeill J, Haft D, Hunt J, Karp P, Klimke W, Krebs C, Macelis D, Madupu R, Martin MJ, Miller JH, O'Donovan C, Palsson B, Ruch P, Setterdahl A, Sutton G, Tate J, Yakunin A, Tchigvintsev D, Plata G, Hu J, Greiner R, Horn D, Sjölander K, Salzberg SL, Vitkup D, Letovsky S, Segrè D, DeLisi C, Roberts RJ, Steffen M, Kasif S (2013) The COMBREX Project: design, methodology, and initial results. PLoS Biol 11(8):e1001638
57. Škunca N, Roberts RJ, Steffen M (2016) Evaluating computational gene ontology annotations. In: Dessimoz C, Škunca N (eds) The gene ontology handbook. Methods in molecular biology, vol 1446. Humana Press. Chapter 8
58. Burger J, Doughty E, Khare R, Wei CH, Mishra R, Aberdeen J, Tresner-Kirsch D, Wellner B, Kann M, Lu Z, Hirschman L (2014) Hybrid curation of gene-mutation relations combining automated extraction and crowdsourcing. Database (Oxford) 22:2014