Placing landMarks in the Knowledge Space: crowd-sourcing landmark publications for benchmarking text-mined predictions

Mark Thompson*, Herman H.H.B.M. van Haagen*,
Barend Mons, and Erik A. Schultes

Leiden University Medical Center

Abstract. It is generally perceived that text-mining systems have failed to deliver on the promise of predicting novel, meaningful relationships between biomedical concepts. Despite successes where novel relationships have been inferred and later confirmed by laboratory experiments, there are many more cases where text-mining did not predict the outcome of high-throughput experiments or population-based genetic studies. Here, we show that this apparent incongruity between text-mined predictions and experimental data results not from a failure of text-mining in principle, but rather, from the confounding of 4 distinct classes of data typically used in this research. Keeping this distinction in mind, and using a novel, crowd-sourced and crowd-curated test set of (among others) protein-protein interactions, we propose a more discriminating standard for the evaluation of text-mined predictions.

1 Introduction

Biological systems are composed of interactions between millions of components: genes, regulatory elements, RNAs, proteins, metabolites, nutrients and drug compounds which give rise to associated functions, healthy phenotypes and disease states that dynamically emerge over the life cycle of the organism. Although high-throughput methods routinely screen for associations between these components, the very large scale of these datasets precludes their analysis except by automated means [1]. In the last decade many text-mining systems have been developed to assist biologists in finding new associations in large and heterogeneous data. Typically, text-mining systems have two goals: (1) to annotate a set of genes with literature-based information, or (2) to infer new associations between concepts (e.g. a novel gene-disease relationships or a protein-protein interactions) that have never before been explicitly stated in literature or recorded in databases. For example, text-mining results were used to predict a relationship between fish oil and Raynauds syndrome [2], the physical interaction between the proteins CAPN3 and PARVB [3], and a novel gene involved in craniofacial development from a 2-Mb chromosomal region, deleted in some patients with

* Authors contributed equally to this work

DiGeorge-like birth defects [4]. Tune Pers et al. used five different sources to annotate results from a genome-wide association study and found the causative gene YWHAH for bipolar disorder [5].

Although these successes demonstrate the potential of exposing novel associations from existing biomedical texts, there are also many examples where text-mining was not able to predict experimental findings from microarrays, GWAS or other large-scale analyses. However, it is hard to evaluate the extent of such failures of text-mining as these cases, viewed as negative results, are not generally publishable. In any case, there is a growing consensus that text-mining is unreliable, and has not delivered on its promise of automated knowledge discovery [6, 7].

Here, we show that the perceived incongruities between text-mined predictions and laboratory studies often reflect confusion at a fundamental level about what text-mining is doing and what text-mined inferences actually represent. Essentially, text-mining exposes knowledge that is already there in the knowledge store (but has yet to be recognized by researchers), while experimental approaches can (and often do) establish novel associations that have no antecedents whatsoever in existing knowledge stores. As such text-mined predictions are, in general, not comparable to independent laboratory data and in such cases we should expect little, if any significant overlap between the two. This does not mean that text-mining systems can not be rigorously evaluated. To the contrary, the performance of text-mining systems can be very accurately assessed but only by directly testing the predictions in the laboratory.

2 Redefining the knowledge space

To help clarify these relations, we partition the knowledge space of potential associations by evidence derived from text-mining analyses and laboratory experiments (Figure 1). The evidence in both cases can be positive or negative, creating four types of conceptual associations.

Type I associations (top left) are cases where both the literature and experiments have provided confirmatory evidence for the association, and therefore represents well-established knowledge (Explicit Knowledge). Indeed, sometimes multiple independent lines of evidence confirm a particular finding making it more reliable. Typically literature is based on experimental evidence (e.g. a publication describing the experiment) so that text-mined Type I associations are often a re-discovery of what is already known, and in this way Type I associations provide confirmation that the text-mining method is working as intended. Although Type I associations enjoy consensus, they are not novel or surprising. An example of a Type I association would be a high association score between the gene huntingtin and Hungtintons disease.

Type IV associations (Figure 1, right bottom) is the Negatome, those associations that have no evidence supporting them, whether they have been explicitly tested, or not. For example, a microarray experiment concludes that two genes are not differentially expressed, or that a SNP from GWAS is not significant. In

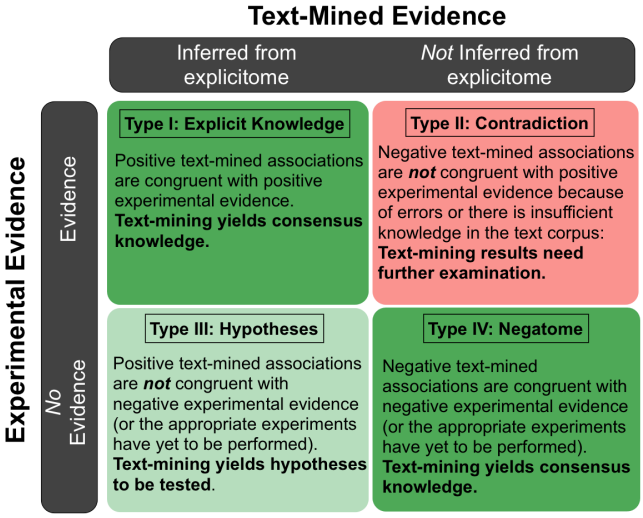

Fig. 1. The space of all possible assertions can be partitioned by whether evidence is derived from text-mining inference or experimental data, yielding 4 Types of assertions that play different roles in the evaluation of the text-mining system.

the text-mining case, a negative result may reflect a failure of the text-mining system, or simply that there is insufficient information in the literature to establish a significant association (a condition we call a Knowledge Vacuum). In any case, like Type I associations, there is a consensus between text-mining and experiments. Type IV associations are by far the largest class of associations and are often treated as a null set of randomly chosen concept pairs in statistical analysis [3].

The remaining associations, Types II and III, are characterized by conflicting results between experiments and text-mining. Type II and III associations are often confounded leading to confusion in the interpretation of text-mining results and erroneous conclusions about text-mining performance.

Type III associations (Figure 1, bottom left) is the case where text-mining can be most effectively used in knowledge discovery. Here, text-mining results predict novel associations that have yet to be tested experimentally, or have been tested but with negative results. In the former case, the predicted associations are treated as hypotheses to be tested, which is the ultimate goal of text-mining. In the latter case, as negative experimental results are always ambiguous, the positive text-mining results can be used as leads looking for associations under alliterative conditions. In either case, Type III associations can be viewed as a prioritized list of Hypotheses guiding the next-step decisions of experimental researchers.

In the case of Type II associations (Figure 1, top right), findings based on experimental evidence are not supported by text-mining, yielding a Contradic-

tion. Many Type II associations come from high-throughput screens or GWAS and are de novo discoveries such that no literature-based information is yet available. Although it is always a possibility that a text-mining method may simply be returning false negatives, failure to predict a positive experimental result could also reflect a text-mining Knowledge Vacuum. In any case, Type II associations necessitate further inquiry and possible trouble-shooting of the text-mining system.

Given the large number of biomedical concepts and their potential pair-wise associations, the Knowledge Vacuum is likely to be a large fraction of the Knowledge Space, that is, it is likely that the vast majority of associations have yet to be represented either explicitly or implicitly in the literature. For example, there are about 25,000 human genes yet only 12,000 of these entities have more than 5 PubMed abstracts, making them visible to text-mining systems. Hence, literature-based knowledge discovery is inherently limited to concepts that have been well-published upon, and can not be used to predict associations between concepts that have yet to be discussed in the literature. Although a large number of associations can, and should be mined, they should not be compared directly with experiments that test, de novo, a much wider class of associations. Hence, Type II associations that involve high-throughput experimental screens should not be viewed as a failure of text-mining, but rather text-mining and high-throughput experiments should be seen as complementary approaches to mapping the space of possible associations.

3 An alternative evaluation method

A more relevant evaluation of text-mining systems can be based directly on the text-mined predictions themselves. We propose the use of retrospective analyses that use benchmark sets of known associations that takes into consideration the taxonomy of potential associations as shown in Fig. 1. In particular, the benchmark datasets makes a distinction between associations that are (or can be) inferred from the explicitome (Type I and III), and those that are not (Type II and IV). We then perform a retrospective analysis using only the predicted associations (Type III) until a certain date and evaluate the prediction by comparing the result against the consensus knowledge after that date (Type I). In this way the evaluation method will not discredit a text-mining result that fails to predict relationships that are inherently unknowable due to a lack of information available in literature.

The key problem is to identify the set of benchmark associations that can be inferred from the explicitome. Benchmark data sets have to define very precisely the concepts that make up the association and the date of first publication of the association. We note some particular problems with trying to automatically generate such a benchmark, for example, using automatically retrieved first co-occurrences. Although first co-occurrences of terms can be determined automatically, mapping those terms to concepts can still not be done with complete accuracy. Moreover, a finding may have been reported first in a publication

that was not in the co-occurrence data set or it may have been reported as a hypothetical relationship, thus co-occurring before it is presented together with any kind of evidence. For such reasons we propose building a curated benchmark by means of crowd-sourcing.

We have developed landMark a landmark publication crowd sourcing tool. A landmark publication refers to the first occurrence in literature of an association between concepts for which experimental evidence is given. It has been developed to allow easy and accurate registration of curated landmarks in a form that we refer to as the "landmark claim": Article X is the first to show a relationship between concept A and concept B The target audience for this tool are publication authors (who register their own landmark findings) as well as, for example, curators who may register landmark findings on behalf of the authors. From a sufficiently large set of such landmark claims we will be able to derive high quality curated benchmark test sets.

These benchmark sets will be made publicly available as a valuable resource for text-mining and knowledge discovery researchers worldwide. For the purpose of simplicity we initially limit ourselves to protein-protein interactions, gene-disease relationships and drug-disease relationships. We think this represents an important subset of landmark findings while making for interesting targets in the current state-of-the-art of knowledge discovery.

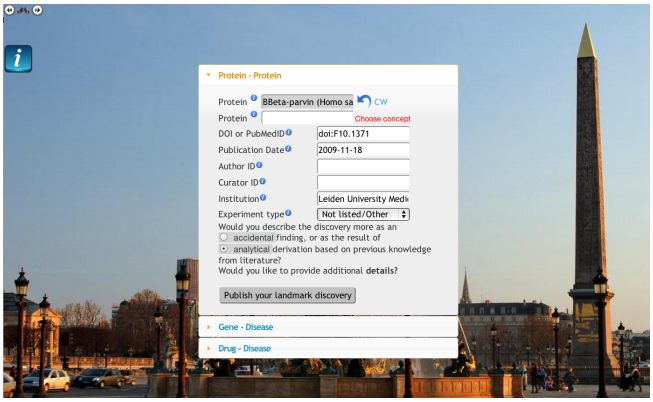

Fig. 2. Screenshot of landMark

The landMark interface presents the user with a very concise web-form that helps the user to quickly and unambiguously enter a landmark claim. Disambiguation is achieved in two steps: first the user selects one of the three relationship categories from the accordion widget, and types a term describing each of the concepts. As the user types, an auto-complete feature queries the ConceptWiki [8] for concepts (within the selected category) that match the (partially) typed term. In case multiple concepts match the term, the user can review their ConceptWiki summaries to manually disambiguate them. The interface also has

fields asking the user to provide a DOI or PubMedID of the landmark paper, its publication date and the first author and his institutional affiliation. The optional curator field identifies the curator making the claim on behalf of the author. Two final questions are asked to identify the type of discovery, from which we can infer the Type of the landmark and thereby the suitability of this landmark towards the evaluation of prediction mechanisms as discussed in the previous section.

A final processing step is required to transform the entire collection of landmark claims into the required benchmark set. For example, consider the situation where an author challenges a previous claim by submitting a new claim that refers to an earlier article. Due to the careful and unambiguous selection of concepts, we can later identify whether two claims refer to the same concepts and include only the information from the claim that refers to the earliest paper in the final benchmark test set.

As with all curation efforts, the quality of the benchmark test set will depend on the quality and the amount of contributions. We hope to incentivize authors to make their landmark claims by offering to turn their landmark claim in nanopublications [9–12]. A nanopublication is a permanent, immutable, semantic-web representation of the smallest unit of publishable information that consists of an assertion and provenance. The landMark web tool will store each landmark claim as a nanopublication assertion with the author, publication date and additional information as nanopublication provenance. As the nanopublication becomes part of the web of linked data, a landmark nanopublication offers a simple way for authors to gain attribution for key parts of their published research and for curators to receive credit for the important (but often underappreciated) effort of data curation.

Currently the landmark nanopublication web application is in an extensive user testing phase at Leiden University Medical Center. We believe usability is an important factor in the adoption of this tool. By reducing the effort required to submit a claim we make it easy for authors, curators and others to submit claims and thus help the creation of a high-quality, curated benchmark test sets.

4 Conclusions

Text-mining results can be partitioned by experimental evidence and text-mined evidence. We clarified that text-mining prediction always has literature as a starting point and is therefore not particularly suitable for predicting associations between concepts for which literature has no (or very little) information. This is often the case for serendipitous findings of high-throughput experiments, such as for example, microarray experiments. We propose an alternative method of evaluation based on a high-quality, curated benchmark data set of landmark associations in literature. We demonstrated an implementation of a web tool that will be made available to the community to crowd-source the creation of such a benchmark set. We hope it will serve as a new and open standard for text-mining and prediction research.

References

1. Attwood, T.K., Kell, D.B., McDermott, P., Marsh, J., Pettifer, S.R., Thorne, D.: Calling international rescue: knowledge lost in literature and data landslide! Biochemical Journal **424**(3) (2009) 317–333
2. Swanson, D.R.: Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspectives in biology and medicine **30**(1) (1986) 7–18
3. van Haagen, H.H.H.B.M., 't Hoen, P.A.C., Botelho Bovo, A., de Morrée, A., van Mulligen, E.M., Chichester, C., Kors, J.A., den Dunnen, J.T., van Ommen, G.J.B., van der Maarel, S.M., Kern, V.M., Mons, B., Schuemie, M.J.: Novel protein-protein interactions inferred from literature context. PLoS ONE **4**(11) (11 2009) e7894
4. Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., Tranchevent, L.C., De Moor, B., Marynen, P., Hassan, B., Carmeliet, P., Moreau, Y.: Gene prioritization through genomic data fusion. Nat Biotechnol **24**(5) (May 2006) 537–544
5. Pers, T.H., Hansen, N.T., Lage, K., Koefoed, P., Dworzynski, P., Miller, M.L., Flint, T.J., Mellerup, E., Dam, H., Andreassen, O.A., Djurovic, S., Melle, I., Børglum, A.D., Werge, T., Purcell, S., Ferreira, M.A., Kouskoumvekaki, I., Workman, C.T., Hansen, T., Mors, O., Brunak, S.: Meta-analysis of heterogeneous data sources for genome-scale identification of risk genes in complex phenotypes. Genetic Epidemiology **35**(5) (2011) 318–332
6. Veuthey, A.L., Pillet, V., Yip, Y.L., Ruch, P.: Text mining for swiss-prot curation: A story of success and failure. In: Nature Precedings. (2009)
7. Dai, H.J., Chang, Y.C., Tzong-Han Tsai, R., Hsu, W.L.: New challenges for biological text-mining in the next decade. Journal of Computer Science and Technology **25** (2010) 169–179
8. `http://www.conceptwiki.org`
9. Groth, P., Gibson, A., Velterop, J.: The anatomy of a nanopublication. Information Services and Use **30**(1) (01 2010) 51–56
10. Mons, B., van Haagen, H., Chichester, C., Hoen, P.B.t., den Dunnen, J.T., van Ommen, G., van Mulligen, E., Singh, B., Hooft, R., Roos, M., Hammond, J., Kiesel, B., Giardine, B., Velterop, J., Groth, P., Schultes, E.: The value of data. Nat Genet **43**(4) (Apr 2011) 281–283
11. Thompson, M., Mons, B., Roos, M., Schultes, E.: Data publishing using nanopublications. In: Tiny Transaction on Computer Science (TinyToCS). (2012)
12. `http://www.nanopub.org`