A Simple Text Mining Approach for Ranking Pairwise Associations in Biomedical Applications

Finn Kuusisto, PhD[1], John Steill, MS[1], Zhaobin Kuang, MS[2], James Thomson, VMD, PhD[1,2], David Page, PhD[2], Ron Stewart, PhD[1]
[1]Morgridge Institute for Research, Madison, USA [2]University of Wisconsin, Madison, USA

Abstract

We present a simple text mining method that is easy to implement, requires minimal data collection and preparation, and is easy to use for proposing ranked associations between a list of target terms and a key phrase. We call this method KinderMiner, and apply it to two biomedical applications. The first application is to identify relevant transcription factors for cell reprogramming, and the second is to identify potential drugs for investigation in drug repositioning. We compare the results from our algorithm to existing data and state-of-the-art algorithms, demonstrating compelling results for both application areas. While we apply the algorithm here for biomedical applications, we argue that the method is generalizable to any available corpus of sufficient size.

Introduction

Many scientific discoveries are often subject to lengthy processes of trial and error before important and meaningful results are found. For example:

1. In biology, determining a set of defined transcription factors for differentiating or reprogramming cell types requires trying numerous combinations from lists of factors. The combinatorial growth of the search space quickly leads to intractability.
2. In medicine, discovering off-label uses of approved drugs can take years of collecting observational data and running post-approval trials. Once again, the search becomes time-consuming due to the enormous number of pairs of drugs and effects.
3. Similarly in medicine, detecting adverse drug events can require extensive observational data to detect potential correlations between drugs and events.

Because the search spaces are so large, proper prioritization of research directions in these cases is essential to reaching novel discoveries quickly, but this requires both extensive breadth and depth of knowledge within the domain. Furthermore, due to exponential growth in scientific literature,[1,2] it is becoming continually more challenging to keep up with current knowledge in any particular domain. We present a general text mining approach to address this prioritization problem by ranking a list of target terms (e.g. transcription factors or drugs) by their association with a key phrase (e.g. "embryonic stem cell" or "hypoglycemia"). This list provides researchers with a starting point for entering the literature domain and prioritizing potential research directions, thereby accelerating the discovery process. Our method is easy to implement, requires minimal data collection and preparation, and is easy to use.

To produce our ranked list of target terms associated with a key phrase, we leverage the vast collective knowledge available within the published scientific and medical literature. We use simple keyword matching and document counting to automatically identify significant correlations and rank them by their co-occurrence proportion. Owing to its simplicity, we call our method KinderMiner.

While we can imagine several applications of our approach, we focus our attention on the two former examples given above: determining important transcription factors for cell reprogramming and discovering off-label uses of approved drugs. To assess our approach, we compare rankings produced by our approach with three cell reprogramming tasks that have experimentally proven sets of defined factors from landmark publications. For fairness, we censor the literature in our experiments to publications from roughly two years prior to the relevant landmark publications. We also apply our approach to the task of discovering drugs that may be repurposed for reducing blood glucose. In both cases, we show that our method is able to reproduce sufficient sets of defined factors and many relevant drugs within the top hits, suggesting that our method will likely be useful in accelerating the discovery process.

The KinderMiner Algorithm

Algorithm 1 breaks KinderMiner down step-by-step. At a high level, KinderMiner ranks a list of target terms by their association with a specified key phrase. It does this via keyword matching and document counting within a specified, relevant, searchable text corpus.

Algorithm 1 The KinderMiner algorithm.

Input: $Corpus$, $TargetTerms$, $KeyPhrase$, $PThreshold$
$topTerms = \{\}$
$articleTotal = ArticleCount(Corpus)$
$kpTotal = ArticleCountWith(Corpus, KeyPhrase)$
for $term \in TargetTerms$ **do**
  $targKP = ArticleCountWithBoth(Corpus, term, KeyPhrase)$
  $targNoKP = ArticleCountWith(Corpus, term) - targKP$
  $noTargKP = kpTotal - targKP$
  $noTargNoKP = articleTotal - targKP - targNoKP - noTargKP$
  $p = OneSidedFisherExact(targKP, noTargKP, targNoKP, noTargNoKP)$
  if $p < PThreshold$ **then**
    $topTerms = topTerms \cup term$
  end if
end for
$sortedTerms = SortByKeyPhraseAndTermRatio(topTerms)$
return $sortedTerms$

First, KinderMiner requires a large corpus of documents for querying. While we focus on corpora of scientific literature, the corpus could also be a collection of plain text patient records taken from an electronic health record, a twitter feed, blog posts, or any other large indexed collection of plain text documents. The corpus must be queryable for document counts with exact matching of words and phrases. For evaluation purposes, it is also useful if the document queries can be date censored, reducing counts of documents to only those that have been published within a specified date range. This is not required, however.

Second, the user must specify a list of target terms to be ranked by their association with a specified key phrase. For example, for one of our cell reprogramming applications, we specify a list of transcription factors and rank them by their association with the key phrase "embryonic stem cell." The goal of this query is to identify the factors necessary for inducing an embryonic stem cell-like state. See Figure 1 for a more visual representation of this set of queries.

Next, for each target term, KinderMiner queries the corpus for documents that contain both, either, and neither the target term and the key phrase, producing a contingency table of document counts. KinderMiner then performs a one-sided Fisher's exact test on the resultant contingency table, and filters out target term, key phrase pairs that do not meet a prespecified significance level. KinderMiner uses the one-sided Fisher's exact test to assess significance only in the direction that there are more articles that contain both key phrase and target.

Finally, the selected target terms are ranked by the ratio of documents containing both the target term and the key phrase, over the total of those containing the key phrase; that is, they are ranked by the proportion of documents containing the target term that also contain the key phrase.

A great deal of work has been devoted to mining the biomedical literature. Our simple approach is related to prior work on co-occurrence statistics and relationship extraction[3,4] which often constrains search to particular types of relationships or relies on more sophisticated techniques such as part-of-speech tagging and named entity recognition. KinderMiner simply constrains the search space by relying on exact text matches to an input key phrase and target terms. Of course, KinderMiner could almost certainly benefit from NLP techniques such as text normalization and named entity recognition. Nevertheless, our goal with this work is to address whole literature information extraction using the simplest approach we can imagine to rank potential associations, using readily available tools and sources of data, and requiring little to no data annotation or processing. Despite its lack of sophistication, we find that our approach performs well when presented with a large corpus.

In the next two sections, we motivate two different applications, cell reprogramming and drug repositioning respec-

Target Terms

AATF	ABP1	…	ZXDC	ZYX

Key Phrase "Embryonic stem cell"

Censor Year 2004

Output Rank

1. Compute article count contingency table
2. Filter terms by one-sided Fisher Exact test
3. Sort terms by $\frac{Key\ Phrase\ \&\ Term}{Term\ Total}$

Example NANOG + "Embryonic stem cell" + 2004

	Term	⌐ Term	Total
Key Phrase	15	2,012	**2,027**
⌐ Key Phrase	44	17,010,295	**17,010,339**
Total	**59**	**17,012,307**	**17,012,366**

One-sided FET p: 5.219e-46 **Sort Ratio:** $\frac{15}{59} = 0.254$

Figure 1: Visual example of KinderMiner, with contingency table and associated Fisher's Exact Test (FET) analysis of the key phrase "embryonic stem cell" and the target term "NANOG." Target terms are filtered by significance of co-occurrence with the key phrase and then sorted by the co-occurrence ratio.

tively, and evaluate the KinderMiner algorithm in the context of these applications. We selected these particular applications not only for their significance to science and medicine, but also because of the availability of reasonable ground truth against which we can compare KinderMiner's findings.

Cell Reprogramming Applications

An increasingly common task in modern biology is the process of taking cells of one type and reprogramming them to exhibit the characteristics of another cell type. Reprogramming in this case often involves introducing a set of transcription factors that put the source cells on track to behave like a different target cell type. A particularly important example of reprogramming is that of somatic cells to an induced pluripotent stem (iPS) cell as iPS cells behave like embryonic stem cells, wherein they have the potential to differentiate into nearly all fetal or adult cell types.[5,6,7,8] Reprogramming can also be accomplished through transdifferentiation, which is when one somatic cell type is directly converted into another somatic cell type.[9] Reprogramming is important because researchers often need particular cell types to create models, study the effects of disease, develop therapies, or perform basic science, but primary cells of certain types are not always available in abundant quantities, if at all.

Altering the expression of transcription factors is also useful in the maturation of cells. For instance, methods exist for differentiating and culturing immature hepatocytes, the main cells of the liver responsible for metabolism of drugs and toxins, but these cells are difficult to mature. Immature hepatocytes cannot serve as reasonable surrogates for hepatocyte function, drug toxicity, or metabolism. Recent publications[10,11] describe methods for partial maturation of hepatocytes using transcription factors. For similar reasons, having methods for differentiating cardiomyocytes, muscle cells of the heart, is useful, and transcription factor sets for differentiating cells into cardiomyocytes have recently been described.[12,13]

Determining a set of important transcription factors for converting one cell type into another is, however, a challenging task that involves a great deal of domain expertise as well as trial and error. There are roughly 2,000 transcription factors to choose from,[14] and researchers must rely on their reading of the literature and intuition to decide which combinations to try and in what order. This search is time consuming, and we propose that our algorithm can assist researchers by accelerating the trial and error process. Instead of trying combinations from the entire list of transcription factors based on intuition, researchers can prioritize their experiments by exploring a much smaller number of possible combinations from only the top ranked factors provided by our algorithm.

To demonstrate our algorithm in this domain, we refer to three well-established sets of factors for reprogramming. The first is for creating induced pluripotent stem cells (iPS cells), the second is for creating cardiomyocytes, and

the third is for the maturation of hepatocytes. We use our algorithm to mine scientific and medical literature and rank a list of transcription factors by correlation with the key phrases "embryonic stem cell," "cardiomyocyte," and "hepatocyte." We then compare the top hits in each list with the experimentally determined factors known to produce cells representative of these cell states. For fairness, we censor the literature available to our algorithm by roughly two years in advance of the earliest publications that demonstrate these conversions.

Drug Repurposing Application

Despite increases in R&D spending, the biopharmaceutical industry has struggled to improve cost and throughput of de novo drug discovery.[15] Due to advances in key technologies and the increasing availability of data, drug repositioning, the detection of new uses for existing drugs, has become more feasible.[16] Furthermore, repositioned drugs do not require a costly development process and can reach clinical trials much faster than traditionally developed drugs. These advantages have led repositioned drugs to constitute approximately 30% of drugs and vaccines newly approved by the US Food and Drug Administration[17].

There have been several computational drug repositioning (CDR) approaches proposed. Computational methods often rely on heterogeneous data sources containing genetic and phenotypic information, drug molecular structure, electronic health records, or plain-text literature as we do here.[16,18,19] We propose that our algorithm is a useful addition to the CDR toolbox, despite being far simpler than other methods.

To demonstrate our algorithm in this domain, we focus on the task of identifying drugs that may reduce blood glucose. We use our algorithm to mine the literature and rank a list of drugs and devices by correlation with the key phrase "hypoglycemia" (i.e. low blood sugar). We manually assess how well our method is able to identify drugs and devices that are specifically used to treat diabetes in the top hits, and then assess the potential of those top hits that are not specifically for treatment of diabetes. We do not censor the date for this task.

Materials and Methods

For our experiments, we used the Europe PMC (EPMC) corpus.[20] We implemented our queries with EPMC's RESTful API, using the *profile* search module with counts coming taken from the *ALL* publication type. We form our queries using quoted, exact matches for both the target terms and key phrases, and we use the FIRST_PDATE parameter to censor publication year from 1900 through the specified year. For example, a query for co-occurrence of the term NANOG and key phrase "embryonic stem cell," censored to the end of 2004, would appear as follows:

```
``NANOG'' AND ``embryonic stem cell'' AND (FIRST_PDATE:[1900-01-01 TO
                              2004-12-31])
```

At time of writing, the EPMC corpus contains a total of approximately 27.5 million publications. Approximately 20 million of the articles were published during or before 2008 and 17 million were published during or before 2004.

For our cell reprogramming applications, we query our lab's list of 2,243 transcription factors against the key phrases "embryonic stem cell," "cardiomyocyte," and "hepatocyte." We use a one-sided FET p-value threshold of 1×10^{-5}. We collect the top 20 transcription factors from each of these queries and use two standards for comparison. First, we search our top factors for factors from landmark publications that have previously been shown experimentally to reprogram somatic cells to iPS cells, cardiomyocytes, and to partially mature hepatocytes. Specifically, the relevant factors we consider for iPS cells are MYC, KLF4, LIN28, NANOG, POU5F1, and SOX2.[6,8,7] The relevant factors we consider for cardiomyocytes are GATA4, HAND2, MEF2C, NKX2-5, and TBX5.[12,13] The relevant factors for hepatocyte maturation are GATA4, HNF1A, FOXA3, FOXA2, HNF4A, CEBPB, and MYC.[10,11]

Second, we identify our top selected transcription factors that are also indicated as being relevant by the Mogrify algorithm, a state-of-the-art algorithm to predict transcription factors for reprogramming between several cell types.[21] Mogrify starts from gene expression data to score differentially expressed genes between cell types of interest and background expression levels. It then combines these differential expression scores with regulatory network information to rank transcription factors in each cell type by regulatory influence. Finally, Mogrify selects optimal sets of transcription factors with the greatest regulatory influence over differentially expressed genes in a given target cell type in comparison to a given starting cell type. Importantly, Mogrify requires a large amount of processed data that may not be readily available and would be costly and time prohibitive to collect.

For the Mogrify comparison, we collect the complete lists of predicted transcription factors from `http://www.`

mogrify.net. For the iPS cell comparison, we use the conversion between *dermal fibroblast* and *H9 embryonic stem cells*. For the cardiomyocyte comparison, we use the conversion between *dermal fibroblast* and *heart - adult*. For the hepatocyte comparison, we use the conversion between *dermal fibroblast* and *liver - adult*.

For the iPS cell queries, we censor the publication date range through the end of the year 2004. This time frame roughly corresponds to two years prior to the first publications on direct reprogramming in mouse cells.[6] For the cardiomyocyte queries, we censor the publication date range through the end of the year 2008, which also corresponds to two years prior to the first major publications on cardiomyoctye reprogramming in mice.[12] We censor to the year 2009 for the hepatocyte applications as it corresponds to roughly two years prior to the first major publication on induction of functional hepatocytes from mouse fibroblasts.[10]

To evaluate our algorithm on the drug repositioning application, we query the same list of 2,609 drugs and devices used by Kuang et al.[18] against the key phrase "hypoglycemia" (low blood glucose). Again, we use a p-value threshold of 1×10^{-5}. Note that we use an exact match of drug names in this case (e.g. *Glucotrol* and *Glucotrol XL* are treated as different) even though there may be multiple names for the same drug. To evaluate our method, we first manually annotate the top 50 hits as either advertised specifically to treat diabetes or not. We then compare those that were not identified as diabetes drugs to a curated list of drugs[22] that are known to cause hypoglycemia, hyperglycemia, or both, reporting those correctly and incorrectly identified as reducing blood glucose. Finally, we mark the top hits that also match hits in the full list of drugs and devices predicted to reduce blood glucose by the state-of-the-art approach proposed by Kuang et al. using electronic health records.[18] Kuang et al. extend the self-controlled case series model[23] to handle continuous numeric responses. The self-controlled case series, which has been widely used for detecting adverse drug events, divides patient time-course data into control and risk periods corresponding to periods before and after exposure to a drug. Patients thus serve as their own control cases and relative incidence of adverse events can be measured in the control and risk periods. Importantly, this approach requires a large amount of time-course electronic health record data, which is difficult to acquire.

Results

Table 1(a) shows the top 20 ranked transcription factors from our query using a list of 2,243 transcription factors and the key phrase "embryonic stem cell," censored to a publication date range through 2004. Factors that match the landmark papers for producing iPS cells are highlighted gray, and factors that match Mogrify's list of predicted factors are marked with *. Note that our naive approach is able to reproduce a sufficient list of factors (NANOG, POU5F1, and SOX2) for direct reprogramming[24] in the top 12 hits. Additionally, five of the top 20 match the list of 70 factors produced by Mogrify.

Table 1(b) shows the top 20 ranked hits from our query using a list of transcription factors and the key phrase "cardiomyocyte," censored to a publication date range through 2008. Again, factors that match the landmark papers for direct reprogramming to cardiomyocytes are highlighted in gray, and factors that match Mogrify's list of predicted factors are marked with *. Similar to the iPS cell query, our approach reproduces the complete list of early published transcription factors in the first nine hits, and nine of the top 20 hits match the list of 57 factors predicted by Mogrify.

Table 1(c) shows the top 20 ranked hits from our query using a list of transcription factors and the key phrase "hepatocyte," censored to a publication date range through 2009. Again, factors that match the landmark papers for direct reprogramming to hepatocytes are highlighted in gray, and factors that match Mogrify's list of predicted factors are marked with *. KinderMiner successfully reproduces four of the six factors for maturation from the landmark literature, and nine of the top 20 hits match the 27 predicted by Mogrify.

Table 2 shows the top 50 drugs and devices ranked by our method as being relevant to hypoglycemia (low blood sugar). Drugs that are advertised specifically to treat diabetes are not highlighted. The highlighted drugs are not specifically advertised to treat diabetes. Drugs highlighted green are labeled as drugs that may reduce blood sugar, drugs highlighted red may increase blood sugar, and drugs highlighted gray are not present in our labeled list.[22]

Perhaps unsurprisingly, 43 of our top 50 hits are specifically for treatment of diabetes, due in part to the abundance of diabetes drugs and various brand names thereof. We note that the hit *premeal* is likely a result of correlation to premeal insulin. These 43 hits are a positive result as they suggest that our method successfully finds relevant correlations, but the more interesting hits are those that are not diabetes drugs as our goal is to reposition drugs. Of the seven hits that are not specifically diabetes drugs, Zestoretic, Avalide, and Demadex have been shown to potentially increase blood glucose, whereas Zebeta, Tiazac, and Calan SR have been shown to potentially decrease blood glucose. Tequin is not

Table 1: Top 20 hits for each of our cell reprogramming queries. Hits that match the landmark papers are highlighted in gray, and hits that match transcription factors predicted by Mogrify are marked with *.

(a) Transcription factors - “embryonic stem cell” - 2004

Term	Term + KP Count	Term Count	Co-occur Ratio
*NANOG	15	59	0.254
*UTF1	5	25	0.200
CBX4	2	21	0.095
*POU5F1	24	272	0.088
EZH1	2	25	0.080
SOX1	8	103	0.078
IRX4	2	26	0.077
*FOXD3	4	54	0.074
MYF6	5	79	0.063
HOXB4	8	158	0.051
LMO2	12	240	0.050
*SOX2	11	230	0.048
EOMES	3	65	0.046
LMX1B	5	112	0.045
LHX2	3	76	0.040
HOXD9	3	78	0.039
HOXD11	3	80	0.038
OTX1	5	140	0.036
HAND1	4	117	0.034
HOXB3	3	88	0.034

(b) Transcription factors - “cardiomyocyte” - 2008

Term	Term + KP Count	Term Count	Co-occur Ratio
MESP1	26	89	0.292
THRAP1	4	15	0.267
*TBX20	30	114	0.263
*GATA4	302	1294	0.233
*NKX2-5	122	528	0.231
*TBX5	104	481	0.216
GATA5	40	194	0.206
*MEF2C	151	825	0.183
*HAND2	52	297	0.175
CSRP3	8	46	0.174
IRX4	10	64	0.156
HDAC9	26	179	0.145
NFATC4	23	173	0.133
*IRX5	8	68	0.118
MKL2	5	43	0.116
ISL1	51	470	0.109
*GATA6	55	526	0.105
*HAND1	30	292	0.103
HES2	6	60	0.100
TBX18	7	73	0.096

(c) Transcription factors - “hepatocyte” - 2009

Term	Term + KP Count	Term Count	Co-occur Ratio
HNF1A	781	849	0.920
HNF1B	639	699	0.914
*HNF4A	466	596	0.782
*ONECUT1	105	140	0.750
HNF4G	23	36	0.639
*FOXA3	137	217	0.631
ONECUT3	6	10	0.600
*FOXA1	313	571	0.548
*FOXA2	523	1055	0.496
TCF2	136	276	0.493
MLX	325	687	0.473
*NR0B2	54	138	0.391
*NR1I3	66	171	0.386
*NR1H4	66	171	0.386
HMBOX1	5	13	0.385
NR1I2	74	200	0.370
ONECUT2	14	40	0.350
TCF1	137	460	0.298
*CREB3L3	7	25	0.280
CUTL2	13	47	0.277

in our labeled list. It is an antibiotic that has been shown to increase patient risk of dysglycemia (either hypoglycemia or hyperglycemia).[25]

Overall, in all of our evaluation tasks, our method finds numerous relevant hits and demonstrates overlap with the results of far more sophisticated methods designed specifically for the separate tasks presented.

Conclusions and Future Work

In this work, we present a simple and general text mining method for predicting pairwise associations between a key phrase and target terms. We demonstrate the use of this method for identifying transcription factors that are important for three cell reprogramming tasks and for discovering candidate drugs for alternative uses. In both of our application domains, we find that KinderMiner identifies numerous relevant hits and overlaps with state-of-the-art methods designed specifically for each domain. In historically censored searches of factors for reprogramming cell states, KinderMiner highly ranks transcription factors that would, years later, be shown to be important for reprogramming to

Table 2: Top 50 drug and device hits for our drug repositioning query against the key phrase "hypoglycemia." Hits that are diabetes drugs are not highlighted. Hits that are not diabetes drugs, but which are known to decrease blood sugar are highlighted in green, and hits that increase blood sugar are highlighted in red. Hits that are not diabetes drugs, but were also not in our labeled list, are highlighted in gray. Hits that are exact matches to those in Kuang et al.[18] are marked with *.

Drug	Drug + KP Count	Drug Count	Co-occur Ratio
GLYBURIDE MICRONIZED	3	4	0.750
GLYNASE	16	27	0.593
MICRONASE	24	41	0.585
NOVOLIN N	28	48	0.583
STARLIX	26	46	0.565
TOLINASE	14	26	0.538
GLIPIZIDE XL	7	13	0.538
*GLUCOTROL XL	15	28	0.536
*INSULIN DETEMIR	547	1107	0.494
PREMEAL	477	975	0.489
SUBCUTANEOUS INSULIN INFUSION PUMP	37	76	0.487
*INSULIN ASPART	723	1509	0.479
*INSULIN LISPRO	717	1515	0.473
NPH INSULIN	787	1665	0.473
PRECOSE	31	66	0.470
PRANDIN	27	59	0.458
LANTUS	290	640	0.453
*GLUCOTROL	41	91	0.451
NOVOLOG	90	203	0.443
ZESTORETIC	3	7	0.429
*HUMALOG	210	495	0.424
*AMARYL	47	113	0.416
INSULIN NPH	281	691	0.407
GLYBURIDE-METFORMIN	46	117	0.393
REGULAR INSULIN	1182	3048	0.388
*GLIMEPIRIDE	935	2487	0.376
BYETTA	136	370	0.368
ZEBETA	4	11	0.364
HUMULIN N	50	140	0.357
*GLUCOPHAGE XR	16	45	0.356
PRAMLINTIDE ACETATE	21	60	0.35
JANUVIA	87	252	0.345
LIRAGLUTIDE	885	2589	0.342
INSULIN REGULAR HUMAN	41	121	0.339
AVALIDE	3	9	0.333
DEMADEX	3	9	0.333
NATEGLINIDE	351	1098	0.320
REPAGLINIDE	458	1486	0.308
AVANDAMET	17	56	0.304
*EXENATIDE	1136	3843	0.296
*GLIPIZIDE	669	2278	0.294
GLUCAGEN	29	99	0.293
*BLOOD-GLUCOSE METER	539	1926	0.280
WELCHOL	22	79	0.278
TIAZAC	4	15	0.267
GLUCOVANCE	18	68	0.265
TEQUIN	11	43	0.256
NOVOLIN R	68	270	0.252
NOVOLIN	143	570	0.251
CALAN SR	4	16	0.25

cell states of interest, thus providing a short, ordered list of candidates for biologists that would have greatly simplified the challenging combinatorial task they faced. Importantly, the domain-specific approaches require domain-specific data, whereas KinderMiner only requires an indexed text corpus. We argue that our method is a valuable new tool that can be used to help prioritize research directions despite its naiveté. Furthermore, we anticipate that our method may prove valuable in domains other than biomedicine by mining other large plain text corpora.

We view the simplicity of KinderMiner as a strength, but this simplicity also leads to limitations. For example, KinderMiner does not explicitly implement any actual natural language processing. Thus, terms like the transcription factor T (Brachyury) are likely to match many articles that do not reference the T gene, but may in fact be matches to middle initials or similar. We do not observe this particular phenomenon in our lists of top 20 hits presented here, but we anticipate that this may be a problem for other queries. While we believe there is value in the simplicity of our method, we expect that the addition of techniques such as text normalization and named entity recognition may help alleviate this issue and, therefore, propose it as future work.

Furthermore, we observe that some of our queries have low total counts of articles for sorting by ratio. For example, THRAP1 in Table 1(b) counts a total of 15 articles that contain the term, four of which contain both the term and key phrase. This may pose a greater challenge when using smaller corpora, or when querying terms or key phrases that are relatively new within the literature. A query that counts a total of four articles, three of which have both term and key phrase may be ranked well by ratio, but is unlikely to actually represent compelling evidence of association. In general, there will always be a horizon of discovery defined by the quantity of published literature for particular key phrases and target terms, but we will explore the use of thresholding, pseudocounts, and other Bayesian approaches to modulate the rank of such cases in future work.

Finally, we note that constructing a search engine around large corpora is non-trivial. We were fortunate with our applications in that Europe PMC offers a web API on which we built KinderMiner, but not all corpora will afford such convenience. We do not propose any specific suggestions for how to address this issue, but instead expect that time will assist with the continued democratization of search tools (e.g. Apache Lucene and SOLR). We anticipate that the availability of easy-to-use software packages will continue to grow, and we propose evaluating applications of KinderMiner using such software on open data as future work.

Acknowledgements

The authors acknowledge support from the National Institutes of Health (NIH) grant number UH3TR000506-05 and the National Institute of General Medical Sciences (NIGMS) grant number R01GM097618-05. The authors also thank Marv and Mildred Conney for a grant to R.S. and J.A.T., and Erin Syth for editorial assistance.

References

[1] Pautasso M. Publication growth in biological sub-fields: patterns, predictability and sustainability. Sustainability. 2012;4(12):3234–3247.

[2] Bornmann L, Mutz R. Growth rates of modern science: a bibliometric analysis based on the number of publications and cited references. Journal of the Association for Information Science and Technology. 2015;66(11):2215–2222.

[3] Cohen AM, Hersh WR. A survey of current work in biomedical text mining. Briefings in Bioinformatics. 2005;6(1):57–71.

[4] Zweigenbaum P, Demner-Fushman D, Yu H, Cohen KB. Frontiers of biomedical text mining: current progress. Briefings in Bioinformatics. 2007;8(5):358–375.

[5] Mitalipov S, Wolf D. Totipotency, pluripotency and nuclear reprogramming. In: Engineering of Stem Cells. Springer; 2009. p. 185–199.

[6] Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell. 2006;126(4):663–676.

[7] Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, Tomoda K, et al. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell. 2007;131(5):861–872.

[8] Yu J, Vodyanik MA, Smuga-Otto K, Antosiewicz-Bourget J, Frane JL, Tian S, et al. Induced pluripotent stem cell lines derived from human somatic cells. Science. 2007;318(5858):1917–1920.

[9] Graf T, Enver T. Forcing cells to change lineages. Nature. 2009;462(7273):587–594.

[10] Huang P, He Z, Ji S, Sun H, Xiang D, Liu C, et al. Induction of functional hepatocyte-like cells from mouse fibroblasts by defined factors. Nature. 2011;475(7356):386–389.

[11] Kogiso T, Nagahara H, Otsuka M, Shiratori K, Dowdy SF. Transdifferentiation of human fibroblasts into hepatocyte-like cells by defined transcriptional factors. Hepatology International. 2013;7(3):937–944.

[12] Ieda M, Fu JD, Delgado-Olguin P, Vedantham V, Hayashi Y, Bruneau BG, et al. Direct reprogramming of fibroblasts into functional cardiomyocytes by defined factors. Cell. 2010;142(3):375–386.

[13] Addis RC, Ifkovits JL, Pinto F, Kellam LD, Esteso P, Rentschler S, et al. Optimization of direct fibroblast reprogramming to cardiomyocytes using calcium activity as a functional measure of success. Journal of Molecular and Cellular Cardiology. 2013;60:97–106.

[14] Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. A census of human transcription factors: function, expression and evolution. Nature Reviews Genetics. 2009;10(4):252–263.

[15] Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. Nature Reviews Drug Discovery. 2004;3(8):673–683.

[16] Li J, Zheng S, Chen B, Butte AJ, Swamidass SJ, Lu Z. A survey of current trends in computational drug repositioning. Briefings in Bioinformatics. 2016;17(1):2–12.

[17] Jin G, Wong ST. Toward better drug repositioning: prioritizing and integrating existing methods into efficient pipelines. Drug Discovery Today. 2014;19(5):637–644.

[18] Kuang Z, Thomson J, Caldwell M, Peissig P, Stewart R, Page D. Computational Drug Repositioning Using Continuous Self-controlled Case Series. In: Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining. ACM; 2016. .

[19] Andronis C, Sharma A, Virvilis V, Deftereos S, Persidis A. Literature mining, ontologies and information visualization for drug repurposing. Briefings in Bioinformatics. 2011;12(4):357–368.

[20] Consortium EP, et al. Europe PMC: a full-text literature database for the life sciences and platform for innovation. Nucleic Acids Research. 2014;.

[21] Rackham OJ, Firas J, Fang H, Oates ME, Holmes ML, Knaupp AS, et al. A predictive computational framework for direct reprogramming between human cell types. Nature Genetics. 2016;.

[22] 390 Drugs That Can Affect Blood Glucose Levels;. Accessed: 2016-09-08. http://www.diabetesincontrol.com/drugs-that-can-affect-blood-glucose-levels/.

[23] Farrington C. Relative incidence estimation from case series for vaccine safety evaluation. Biometrics. 1995;p. 228–235.

[24] Huangfu D, Osafune K, Maehr R, Guo W, Eijkelenboom A, Chen S, et al. Induction of pluripotent stem cells from primary human fibroblasts with only Oct4 and Sox2. Nature Biotechnology. 2008;26(11):1269–1275.

[25] Park-Wyllie LY, Juurlink DN, Kopp A, Shah BR, Stukel TA, Stumpo C, et al. Outpatient gatifloxacin therapy and dysglycemia in older adults. New England Journal of Medicine. 2006;354(13):1352–1361.