Research Article

Text mining of CHO bioprocess bibliome: Topic modeling and document classification

Qinghua Wang1,2*, Jonathan Olshin1,2*, K. Vijay-Shanker1, Cathy H. Wu1,2,3

1 Department of Computer and Information Sciences, University of Delaware, Newark, Delaware, United States of America, 2 Center for Bioinformatics and Computational Biology, University of Delaware, Newark, Delaware, United States of America, 3 Department of Biochemistry and Molecular & Cellular Biology, Protein Information Resource, Georgetown University Medical Center, Washington, The District of Columbia, United States of America

* These authors contributed equally to this work.

* wangq@udel.edu

Abstract

Chinese hamster ovary (CHO) cells are widely used for mass production of therapeutic proteins in the pharmaceutical industry. With the growing need in optimizing the performance of producer CHO cell lines, research on CHO cell line development and bioprocess continues to increase in recent decades. Bibliographic mapping and classification of relevant research studies will be essential for identifying research gaps and trends in literature. To qualitatively and quantitatively understand the CHO literature, we have conducted topic modeling using a CHO bioprocess bibliome manually compiled in 2016, and compared the topics uncovered by the Latent Dirichlet Allocation (LDA) models with the human labels of the CHO bibliome. The results show a significant overlap between the manually selected categories and computationally generated topics, and reveal the machine-generated topic-specific characteristics. To identify relevant CHO bioprocessing papers from new scientific literature, we have developed supervised models using Logistic Regression to identify specific article topics and evaluated the results using three CHO bibliome datasets, Bioprocessing set, Glycosylation set, and Phenotype set. The use of top terms as features supports the explainability of document classification results to yield insights on new CHO bioprocessing papers.

1. Introduction

1.1 CHO bibliome

Chinese hamster ovary (CHO) cells are widely used for biological and medical research [1, 2]. They are the predominant host for mass production of many therapeutic proteins such as recombinant monoclonal antibodies in the pharmaceutical industry [3]. With the increasing market demand and growing need in optimizing the performance of producer CHO cell lines, research on CHO cell line development and bioprocess engineering continuously increases in recent decades [4, 5]. In 2016, Golabgir et al. [6] reported a manual bibliographic compilation of the published CHO cell studies from January 1995 to June 2015, which were retrieved with keywords “CHO cells” and/or “Chinese hamster ovary” in the
The initial article set (10,279 abstracts) was manually filtered to identify a bioprocess (BP) set (1157 abstracts) that focus on CHO cell bioprocesses and biotechnologies, including host cell line engineering, strain selection/screening, and cell culture media design, etc. The non-BP set covers the remaining abstracts describing studies irrelevant to CHO bioprocess. For each BP abstract in the CHO bibliome, one or more category labels from a total of 16 research categories were manually assigned based on the types of phenotypic and bioprocess data contained therein [6].

The CHO bibliome continues to grow since its last compilation in 2015, with over 500 PubMed citations annually. To automate text analysis of the CHO bibliome and gain insight to key topics and trends in CHO bioprocessing and biotechnologies, we have applied topic modeling to explore and classify CHO literature and compared results with those manually assigned category labels in the CHO bibliome. When coupled with our classifiers trained with supervised machine learning methods, the resulting models can automatically classify the newly published CHO cell studies after 2015 into bioprocess categories and help researchers select CHO cell research articles of their interest.

1.2 Topic modeling and document classification

Natural language processing (NLP) allows machines to interpret human language with either unsupervised or supervised approaches [7, 8]. For text analysis to uncover the main topics in an unlabeled set of documents, probabilistic topic models are considered an effective framework for unsupervised topic discovery [9, 10]. Latent Dirichlet Allocation (LDA) is a widely used topic modeling method [11] with many applications [12]. It is a generative probabilistic model of a corpus. The basic principle is that documents are represented as random mixtures over latent (hidden) topics, where each topic is characterized by a distribution over words in the corpus. In this study, LDA is adopted for an automatic exploration of latent topics in the CHO bioprocess bibliome, which are then compared and contrasted with those previously manually assigned research categories. One of the motivations for conducting LDA is to find out how well machine-generated topics agree with human labels in general. It is possible that when human labels categories, they rely on their domain knowledge and objectives for document classification, whereas topic labeling with LDA would generate topics that are not biased by existing knowledge. For example, in the CHO bibliome human labeled Glycosylation set is quite homogenous, whereas the Phenotype set has heterogeneous sub-topics. In general such study can provide insight into practical performance of LDA topic models in comparison with human manual category labels, and potential benefits from applying topic modeling to identify significant topics.

To identify new CHO bioprocessing papers from PubMed (especially for publications after 2015), a classifier is needed to separate BP from non-BP studies and identify their bioprocessing topics by learning how the existing CHO bibliome classifies them. For this task, a supervised approach, Logistic Regression, is utilized to classify the bibliome using three data-sets, one for the overall “Bioprocess” category (BP set), and two on the specific bioprocessing categories of “Phenotype and Production Characteristics” (Phenotype set) and “Glycosylation” (Glycosylation set), respectively. Logistic regression allows for different term representations to be used in classification efficiently, ranging from a term’s binary presence/absence method, term frequency (tf), and term frequency-inverse document frequency (tf-idf). Our objective is to determine if each category of interest includes unique terms that could be used for document classification. If the model is able to predict the category of a document in a dataset with high accuracy, it suggests that the documents in that category share an
adequate amount of unique terms for classification, which may yield insights on new CHO bioprocessing papers.

2. Methods

The CHO bibliome processing and analysis workflow consists of document processing, unsupervised topic modeling, and supervised document classification, as summarized in Fig 1.

2.1 Document processing

2.1.1 Literature corpus. For both topic modeling and document classification tasks in this study, we used the abstracts that were compiled in the CHO bibliome paper [6]. To retrieve the abstract texts for the citations in the bibliome, PubMed was used for obtaining PMIDs based on matching of title, doi, and/or journal information. PubTator API was used to retrieve the documents containing both title and abstract text with PMID as query [13]. The resulting dataset consisted of 9689 documents, including 1049 documents in the Bioprocessing (BP) set and 8640 documents in the non-BP set.

2.1.2 Pre-processing. The text processing included typical NLP steps: removal of special characters and numbers, removal of stop words (NLTK package [14]), tokenization, and term lemmatization (with Part of Speech (POS) Tagging allowed for 'NOUN', 'VERB', 'ADJ', 'ADV','PROPN', 'NUM'; with spaCy library [15]).

Further text processing for document classification was conducted using Python. The dataset of 9689 documents (BP + non-BP) were mapped to their designated topic classification as marked in [6]. The articles were labeled with a 0 or 1 signifying each document’s allocation to the negative and positive set, respectively, for each bioprocess category of interest. These articles with known bioprocess classification allow a training set and test set to be created. 5-fold cross-validation was used for checking the accuracy of the model, where 80% of the dataset was used as the training set and 20% as the test set each time.

2.1.3 Document datasets. Three document datasets were compiled to study the efficiency and accuracy of the classifiers for predicting previously unseen documents. The BP set consisting of 1049 BP and 8640 non-BP documents was used to discern bioprocessing-related papers from all CHO cell publications. Two human-labeled categories of the BP documents, Glycosylation and Phenotype, were used to study how well the system can automatically identify articles containing these two bioprocess categories of interest. The Glycosylation set consisted of 70 Glycosylation and 979 non-Glycosylation BP documents, while the Phenotype set consisted of 547 Phenotype and 502 non-Phenotype BP documents.
2.2 Topic modeling using LDA

LDA is among the most widely applied probabilistic topic modeling approach [12]. Python’s GENSIM package [16] was used for LDA applications in this study. Bigrams and trigrams were created with GENSIM phrase detection and added to the dictionary. Words that appear in less than 5 documents were filtered out, resulting in 2534 words in the final dictionary. Lastly, BP documents were included in training the LDA model with Python’s GENSIM package. Method of grid search is employed to select the best set of hyperparameters (i.e., the number of topics, alpha, and eta) for the final LDA model. The resulting document-to-topic probabilities based on the chosen model were analyzed and compared with the previously reported manual category assignments. Python library pyLDAvis [17] is used for interactive topic model visualization.

2.3 Document classification

2.3.1 Logistic regression. It was implemented for classification of three document datasets, BP, Glycosylation, and Phenotype. Multiple trials were conducted. The first trial was run using a binary term representation feature with the entirety of the vocabulary. The binary feature specifying whether a term was present or not in the article was considered. The results thus represented a baseline for the performance of the logistic regression on classification tasks.

The second trial involved the use of tf-idf, which is often used to capture the importance of the terms in the document. Only terms with minimum document frequency (df_min) of 0.05 and a maximum document frequency (df_max) of 0.95 were used. This allowed words to be removed that only occurred a very limited amount of times in the dataset. The use of the entire feature list allows the feature list length to be altered in subsequent trials. Tf-idf was applied to the dataset, first with the entire feature list, followed by altered feature list with decreasing size to see if a more optimal set occurred within the entire list of features. The code used for the logistic regression trials can be found here: https://github.com/udel-biotm-lab/Chinese-Hamster-Ovary-Cell-Logistic-Regression.git.

2.3.2 Under-sampling. It is common practice to use under-sampling when the distribution of the sets is quite skewed, as in the cases of BP/non-BP (1049/8640 documents) and Glycosylation/non-Glycosylation BP (70/979 documents). We applied under-sampling of the majority class by varying the under-sampling rate with each iteration. The statistics of the overall efficacy of the model were taken with each trial. The fraction of the majority set was varied between 0.1 and 0.9 by intervals of 0.1. This means that the majority set, negative set, was cut down to 10% and all the way up to 90%. The test set and training set remained as specified above. As the under-sampling rate was adjusted, the feature set length was altered by 10% each time as well, thus, as the fraction of the negative set increased towards 90% by 10% intervals the feature set length increased at the same rate. Once under-sampling was conducted, logistic regression was run using the tf-idf.

3. Results and discussion

3.1 LDA topic modeling

3.1.1 Comparative analysis of LDA topics and manual categories. We have compared the topics ("Topics") uncovered by the LDA models with bioprocess categories ("Categories") manually compiled in the CHO bibliome [6]. The 15 human labeled categories and their document dataset sizes (S1 Table) are: Phenotype and Production Characteristics ("Phenotype", 547 documents), Enzyme Analysis ("Enzyme", 152), Glycosylation ("Glycosylation",
The LDA model discovered 9 topics from the bioprocessing documents. The top four topics, covering 20.5% to 12.6% of tokens (i.e., terms or words including bigrams and trigrams) of the corpus, account for a total of 65.5% of tokens in the whole corpus.

LDA allows multiple topics for each document, by showing the probability of each topic. For example, document for PMID 9043639 has probability of 0.49 for Topic-4 and 0.37 for Topic-1 according to the LDA model predictions. To simplify further analysis, each BP document was assigned a representative Topic ID corresponding to the highest probability score (e.g., PMID 9043639 is assigned Topic-4 as its representative topic). To compare how LDA topics align with human category labels, heatmaps were generated where the columns show the human Category label and rows correspond to documents (PMIDs shown on the left) broken into different LDA topic groups.

Fig 3 shows the comparative analysis of automatically generated LDA topics and the manually annotated categories. The overall distribution readily reveals that human category labels are differentially captured by LDA topics (Fig 3A). Among the four largest topics (containing
153 to 218 documents), Topic-1 is mapped to Categories "Phenotype", "Transcriptomics", "Proteomics", and several other categories; Topic-2 to Categories "Phenotype", "Expression", "Cell Line", "Secretion"; Topic-3 to "Phenotype", "Glycosylation", "Purification", "Enzyme"; and Topic-4 to "Enzyme" and "Phenotype" (Fig 3B). While Topic-1, -2 and -3 spread over several topics, Topic-4 has only two major categories. The document datasets for the remaining topics are much smaller and predominant with Category "Phenotype" (Fig 3A). Among the top 4 categories, "Phenotype", "Enzyme", "Glycosylation" and "Purification", the largest is "Phenotype" which accounts for over 50% of all bioprocess publications in the CHO bibliome. Not surprising, it has a diverse distribution over many Topics (Fig 3C). In contrast, the other three categories all have only one dominant Topic each, "Enzyme" is dominant with Topic-4, and "Glycosylation" and "Purification" are dominant with Topic-3.

3.1.2 Interpretable terms in topic models. A basic question to ask about a topic model is whether the topics are interpretable to human. LDA represents documents as a mixture of topics, and a topic as a mixture of words, with different weights as the probability of those words appearing in the topic. Fig 4 shows the top terms for each topic and a pyLDAvis display from interactive topic model visualization (S2 File). In Fig 4A, each of the LDA topics is featured with top 30 most frequent terms with term weights (i.e., probabilities). Here, word “cell” is among the top 15 most frequent words for all topics. It is ranked first in Topic-4, -5, -6, -7, with the highest weight for Topic-4. On the contrary, word “mutant” is exclusive to Topic-4 among the top 30 words for all topics, therefore it is a discriminative key term in capturing a document into Topic-4. Fig 4B shows the pyLDAvis display of top 30 most frequent words for
Fig 4. LDA topics with term probability. (A) The top 30 most frequent terms from nine LDA topics with weights. (B) Visualization of topic modeling results using pyLDAvis. Left shows semantic topic space, where each circle is a single topic and its size represents its importance in the model. The proximity between two circles reflects the semantic similarity of their concepts. Right shows Top-30 most salient terms for Topic-4. The terms (red bars) are in descending order of probability, and the blue bars show the terms’ frequency over whole corpus (i.e., a pair of overlaid bars represent both the corpus-wide frequency of a given term as well as the topic-specific frequency of the term). For a given term, when the red bar is almost the same length as the blue bar, it means it is a salient term almost exclusive to the topic.

https://doi.org/10.1371/journal.pone.0274042.g004
Table 1. Logistic regression utilizing binary representation of terms.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>F1-Score</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phenotype/non-Phenotype 547/502</td>
<td>0.76</td>
<td>0.78</td>
<td>0.75</td>
<td>0.76</td>
</tr>
<tr>
<td>BP/non-BP 1049/8640</td>
<td>0.61</td>
<td>0.66</td>
<td>0.57</td>
<td>0.92</td>
</tr>
<tr>
<td>Glycosylation/non-Glycosylation 70/979</td>
<td>0.22</td>
<td>0.5</td>
<td>0.14</td>
<td>0.93</td>
</tr>
</tbody>
</table>

3.2 Logistic regression for classification

3.2.1 Binary representation. The results for binary representation of a term (presence or absence of a term in a document) serve as a baseline to see if each subsequent configuration improved the overall effectiveness of the classifier (Table 1). The Phenotype set obtained the best classification results. The BP set had lower F1-score, precision, and recall. The Glycosylation set had large fluctuations in the results and this could be the first sign that the small positive set and the large size difference between the positive and negative set is affecting the results of the classification.

3.2.2 Term frequency-inverse document frequency (tf-idf). Utilizing tf-idf, the length of the feature set can be seen. This allows feature engineering to find optimal configurations of the classifier. The df_min and df_max values were set to remove terms with very low occurrence rates. The results show that by using tf-idf the performance (especially precision) improved slightly for the BP vs. non-BP (Table 2). The statistics for logistic regression with a shortened feature list of top terms were quite similar to the trial run with all terms (Table 2). However, this configuration may be more desirable as the same results can be achieved with a much smaller set of terms.

For the Glycosylation set, the performance dropped to all zeros with a high accuracy level. This trend indicates that the model was predicting all of the documents as negatives and because the true number of positives was so limited the accuracy was 93%. This confirms the need for under-sampling. The accuracy is also much higher than the other statistics for the BP vs. non-BP which shows that under-sampling may be needed here as well.
3.2.3 Under-sampling. Previous results showed that the results were better for the phenotype/non-phenotype sets than with the other two classifications. We believe this is because the phenotype/non-phenotype document distribution is reasonably balanced, unlike the other two cases. We conducted under-sampling to address the unbalanced size of the positive and negative data for the Glycosylation/non-Glycosylation (70/972) and the BP/non-BP (1049/8640) sets. When the ratio of the positive to the negative data was set to be roughly 1:1, the performance was greatly improved (Table 3).

Table 2 shows that the best results were obtained in the case of phenotype/non-phenotype dataset. These results are based on the simple choice of representation of terms—whether or not the term appeared in the document. Table 3 shows the results of the same three classifications with use of tf-idf for the terms. We observed that this commonly-used representation of terms in information retrieval doesn’t have any significant impact on phenotype/non-phenotype classification but offers a slight improvement for BP/non-BP classification, with a significant gain in precision. This experiment also used a cut-off for the use of terms by applying a threshold for minimum and maximum document frequency (i.e., how many documents does a given term appear in). Further restrictions to terms by using top terms only did not show much change in the composite F1 scores.

We also noted that the BP/non-BP and Glycosylation/non-Glycosylation datasets are imbalanced with distribution being heavily skewed towards the negative set. This clearly impacted the results of these two classification tasks in contrast to the better balanced phenotype/non-phenotype. To address this situation, we applied under-sampling of the majority class and show in Table 3 the F1 scores improves significantly, especially in the gain of recall as is to be expected with undersampling the majority, negative class.

4. Conclusions

In this work, we have described approaches to analyze CHO cell literature that would be of general interest to researchers of CHO bioprocessing research and broader bioengineering and biotechnology community. It makes use of the existing CHO bibliome dataset previously manually labeled with research categories in 2016. Our unsupervised topic modeling enabled a detailed comparison between human labels and machine-generated topics, which empowers qualitative and quantitative understanding of the CHO literature set. Even though the size of the corpus for LDA is relative small in our current study from NLP perspective, select topics notably mirror some human manually assigned topic categories. With the insight gained from

---

Table 2. Logistic regression utilizing tf-idf with all terms and chosen top terms.

<table>
<thead>
<tr>
<th>Document Dataset (Size)</th>
<th>Feature Size</th>
<th>F1-Score</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phenotype/non-Phenotype (547/502)</td>
<td>All terms (474)</td>
<td>0.76</td>
<td>0.79</td>
<td>0.73</td>
<td>0.76</td>
</tr>
<tr>
<td>BP/non-BP (1049/8640)</td>
<td>All terms (706)</td>
<td>0.62</td>
<td>0.75</td>
<td>0.53</td>
<td>0.93</td>
</tr>
<tr>
<td>Phenotype/non-Phenotype (547/502)</td>
<td>Top terms (250)</td>
<td>0.75</td>
<td>0.76</td>
<td>0.75</td>
<td>0.75</td>
</tr>
<tr>
<td>BP/non-BP (1049/8640)</td>
<td>Top terms (500)</td>
<td>0.63</td>
<td>0.75</td>
<td>0.54</td>
<td>0.93</td>
</tr>
</tbody>
</table>

Table 3. Logistic regression utilizing under-sampling and tf-idf with chosen top terms (fraction of majority set = 0.1).

<table>
<thead>
<tr>
<th>Document Dataset</th>
<th>F1-Score</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glycosylation/non-Glycosylation</td>
<td>0.64</td>
<td>0.7</td>
<td>0.58</td>
<td>0.76</td>
</tr>
<tr>
<td>BP/non-BP</td>
<td>0.71</td>
<td>0.65</td>
<td>0.78</td>
<td>0.66</td>
</tr>
</tbody>
</table>

https://doi.org/10.1371/journal.pone.0274042.t002

https://doi.org/10.1371/journal.pone.0274042.t003
our LDA model, we further applied supervised learning for document classification to address the pressing need of classifying new unseen publications automatically instead of time-consuming manual labeling. Making use of terms as features for given topics, the effect of different feature representations on classifier performance is studied. Our study showcases important applications of text analytics on a biological scientific corpus: it discovers structural relations between topics and documents, summarizes corpus via visualization, and discusses challenges and future studies for consideration.

We have explored supervised deep learning method BioBERT (Bidirectional Encoder Representations from Transformers) due to its strength in classifying biomedical literature [18]. We used the Google Cloud platform and conducted preliminary studies where the learning rate, epoch amount, and token limits, as well as other variables can be controlled [18, 19]. This method is a possible path for supervised text classification of datasets such as the CHO bioprocess bibliome, but more research must be performed to test its applicability. BERT models are powerful models and tend to overfit when training data is not sufficiently large. This was a factor for not including their use in this work.

Supporting information

S1 Fig. Overview of CHO bibliome bioprocessing set with manual categories.
(PDF)

S2 Fig. Snip shot of Topic4Term.html in S3 File to show PMID 9043639 document with salient terms in color.
(PDF)

S3 Fig. Heatmap of documents by LDA topic-1 to -9 and manual categories.
(PDF)

S1 Table. Human labels of categories for CHO bioprocess set.
(PDF)

S1 File. Cross comparisons between LDA topics and manually assigned categories for a subset with single dominant LDA topic assigned.
(PDF)

S2 File. LDA topics summarized with pyLDAvis in interactive html.
(HTML)

S3 File. Compressed folder of the html documents by LDA topics with salient terms in color.
(ZIP)

Acknowledgments

We thank CHO Genome to Phenome CHOg2p project community for their helpful discussions and suggestions, and appreciate Dr. Sarah Harcum (Clemson University) and Dr. Kelvin Lee (University of Delaware) groups for sharing their ideas at the early stage. This work is completed with the kind aid of Dr. Debarati Roy Chowdhury, Dr. Sachin Gavali, Dr. Cecilia Arighi, and Dr. Peng Su from University of Delaware.

Author Contributions

Conceptualization: Qinghua Wang, K. Vijay-Shanker, Cathy H. Wu.
Funding acquisition: Cathy H. Wu.
Investigation: Qinghua Wang, Jonathan Olshin.
Supervision: K. Vijay-Shanker, Cathy H. Wu.
Visualization: Qinghua Wang, Jonathan Olshin.
Writing – original draft: Qinghua Wang, Jonathan Olshin.
Writing – review & editing: K. Vijay-Shanker, Cathy H. Wu.

References