Topic Modeling as an Approach to Enzyme Fingerprinting

Alexander Bock
Final Project Report
CS150: Computational Systems Biology
Professor Soha Hassoun
Tufts University
Spring 2019

I. Introduction

In the era of such massive volumes of proteomic data, enabled by advanced in technologies like high-throughput sequencing, efficient computational methods are needed to make sense of trends and patterns in proteomic datasets that could have significant implications for biological and medical research. One such technique that can be implemented with computational solutions is a standardized representation of amino acid sequences that can enable efficient comparison between and classification of studied amino acid sequences.

In the field of chemoinformatics, "fingerprinting" techniques have been developed to represent simple to complex organic molecules with computationally understandable representations. Such representations include SMILES, a linear string-based representation system encoding atoms, bonds, and some chemical substructures [1], and more recent systems that have developed bit-vector encodings of more complex chemical composition and structure (e.g. MACCS, PubChem, and BCI). However, these methods are restricted to atomic-level molecular representations, and there is a general lack in comparable methods for representing amino acid sequences other than the sequences themselves. Peptide mass fingerprinting is a technique that uses mass spectrometry data from the analysis of subdivided amino acid sequences, using generated mass spectra to construct fingerprint representations of full sequences. However, this method relies on empirical MS data and comparisons against existing spectral datasets, and the analysis of novel sequences would involve MS experimentation on them to establish a reliable MS spectrum profile.

This project proposes the use of topic modeling to generate fixed-size and standardized fingerprint representations of amino acid sequences and presents the results of preliminary analysis of such a method by applying topic modeling to a dataset of enzyme amino acid sequences and comparing resulting fingerprints in selected Enzyme Commission (EC) taxa. Topic modeling refers to a set of approaches common in the text mining field that seek to classify lexical components of a corpus of documents into semantically significant groups (or "topics"). Blei notes the extra-linguistic applications of such methods, citing uses in genetic, image, and social network datasets [2]. Schneider et al. demonstrate how topic modeling can be applied to small (non-peptide) organic molecules by analyzing patterns in chemical substructures [3]. The linguistic intent of such approaches is to discover connections between words that compose documents in a corpus. In addition, many topic model implementations are able to, after training, consider some novel document and provide a profile of the document by estimating the prevalence of each generated topic in that document. Using this capability provided by topic modeling, this project assesses how the relative prevalence values of a document provided by a trained topic model can act as a fingerprint for an amino acid sequence.

II. Methods

The methodological profile of this project will be discussed in four stages: Developing a text mining-proteomics analogy and processing amino acid sequences, constructing and training a topic model, evaluating the topic model, and using related enzyme amino acid sequences as test sets to generate example results.

A. Parsing amino acid sequences

The first step in applying topic modeling to a non-text dataset is mapping relevant components of text corpora to components of the dataset in question. In practice, this involves defining a "document" and a "word" in non-text contexts in which such elements are not inherently apparent. Because the topic modeling implementation this project uses abstracts "words" with numerical identifiers and trains a topic model using those identifiers, the existence of actual linguistic words is irrelevant. In this implementation, a sequence of amino acids is treated as a "document" and subdivided into (overlapping) n-grams (in practice, trigrams), representing "words." Ultimately, each sequence is represented as a list of its constituent trigrams; a sequence of length l will necessarily have a trigram representation of length $l-2$ (and an n-gram representation of length $l-(n-1)$).

The amino acid sequence dataset [1] used by this project contains sequences of 254,016 unique enzymes across a wide spectrum of the range of EC classifications [5]. Each amino

[1] This dataset was provided by Professor Soha Hassoun's research group at Tufts University.

acid sequence used for topic model training and testing underwent n-gram parsing using Python to generate the representations described above.

B. Constructing and training a topic model

This project uses a latent Dirichlet allocation (LDA) topic model to sort amino acid trigrams into topics. LDA is a probabilistic topic modeling method that uses probability distributions to identify patterns in the co-occurence of words in documents [2]. This project uses Gensim[2], an open-source Python library providing an implementation of LDA topic modeling.

First, Gensim's `Dictionary` module converts a set of word-list representations of documents to a numerical abstraction, building a one-to-one unique word-to-numerical identifier mapping and representing documents as a list of tuples (n, f), where n is a unique word's numerical identifier and f is the number of times word n occurs in the document. This representation of the dataset is expected by Gensim's `LdaModel` module for training. (Remaining consistent with the analogy described, amino acid n-grams are words and full amino acid sequences are documents, in practice.)

Next, Gensim's `LdaModel` module uses this dataset abstraction and one-to-one word-identifier mapping to train a topic model that sorts observed words into a given number of topics. The number of topics provided can (and will) be used as a graduated independent variable in evaluating trained topic models. Gensim provides an interface to prepare and initiate topic model training; similar to the use of open-source machine learning packages (e.g. Scikit-Learn), the actual implementation of the entire training process is performed by Gensim.

In this project, topic models were built on 10,000 amino acid sequences of selected enzymes from the dataset.

C. Evaluating the topic model

Using Gensim's interface, evaluation statistics can be generated for a trained topic model. Coherence statistics are widely used in topic modeling to assess the quality of probabilistically generated topics [4]. The statistics aim to provide a (somewhat more deterministic) estimate of how often words grouped within the same topic actually occur in the same document. One such statistic, UMass coherence, is implemented by Gensim. UMass coherence is a relatively straightforward measure of coherence, producing a statistic that, given a pair of words (x, y) in the same topic, represents the ratio between the number of documents containing both x and y and the number of documents containing x or y alone, but not both, thus providing a measure of how often these words truly co-occur. Gensim provides a method of calculating UMass coherence.

Another metric used to evaluate topic models is log perplexity, a common measure of how well probability distributions are able to predict given samples. Because LDA topic models are built on probability distributions, this can also serve as a useful measure of how well a topic model has sorted words into topics. Gensim also provides a simple method of obtaining the log perplexity of a trained model.

In this project, 10 topic models with varying topic numbers (10, 20, 30, ... 100) were trained on the same set of 10,000 parsed amino acid sequences, and each individual model was evaluated by UMass coherence and log perplexity. Trends in each of these evaluation statistics were analyzed to determine an optimal number of topics with which to train a model on the 10,000-sequence dataset.

D. Testing the topic model

In order to assess the real-world performance of topic models (beyond the numerical assessments the metrics discussed above provide), certain generated topic models were provided with parsed amino acid sequences from EC classifications not a part of the 10,000 sequences the models were initially trained on, analogous to "unseen" testing sets used in machine learning classification problems. These EC classification groups were 109 oxidoreductase sequences (EC 1.14.12.-) and 1,569 RNA helicase sequences (EC 3.6.4.13). Using Gensim's interface, a trained topic model can be provided with Gensim-abstracted representations of a document and produces a vector of values $\{p_1, p_2, p_3, ...p_n\}$, where n is the number of topics the model was trained on and p_i is the probability that the document can be described as discussing topic i, itself a function of the share of words of topic i among words in the document. For any amino acid sequence that undergoes the n-gram parsing and Gensim abstraction processes described in this section, its representation can be provided to a trained topic model which produces the sequence's corresponding topic vector. This topic vector, of fixed length equal to the number of topics on which the model is trained, is the end result of the pipeline that this project describes, and can potentially act as a fingerprint for amino acid sequences analogous to bit vector representations of small organic molecules described in the previous section. Distributions of topics among enzyme amino acid sequences with strong sequence similarity or shared EC classification can then be analyzed to assess the level of regularity (and diversity) of topic distributions within these sequence groups.

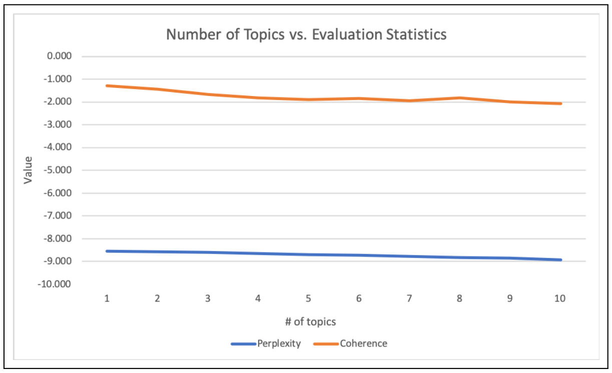
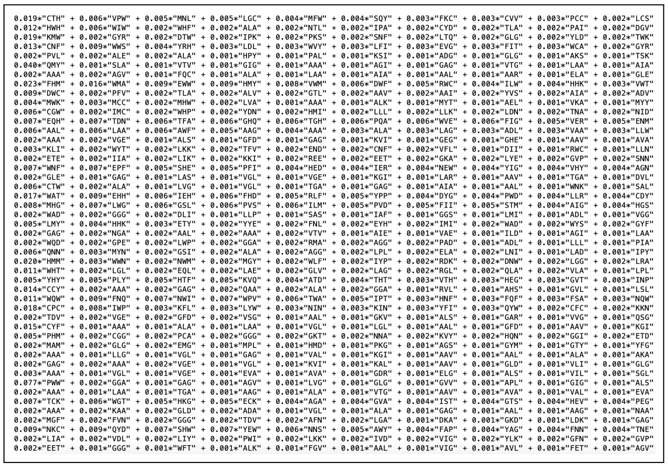


Fig. 1. Relationship between number of topics generated and coherence and perplexity scores of trained topic models.

[2] https://radimrehurek.com/gensim/



```
0.019*"CTH" + 0.006*"VPW" + 0.005*"MNL" + 0.005*"LGC" + 0.004*"MFW" + 0.004*"SQY" + 0.003*"FKC" + 0.003*"CVV" + 0.003*"PCC" + 0.002*"LCS"
0.012*"HWH" + 0.006*"WIW" + 0.002*"WHF" + 0.002*"ALA" + 0.002*"NTL" + 0.002*"IPA" + 0.002*"CYD" + 0.002*"TLA" + 0.002*"PAI" + 0.002*"DGV"
0.019*"KMW" + 0.002*"GYR" + 0.002*"DTW" + 0.002*"IPK" + 0.002*"PKS" + 0.002*"SNF" + 0.002*"LTQ" + 0.002*"GLG" + 0.002*"YLD" + 0.002*"TWK"
0.013*"CNF" + 0.009*"WWS" + 0.004*"YRH" + 0.003*"LDL" + 0.003*"WVY" + 0.003*"LFI" + 0.003*"EVG" + 0.003*"FIT" + 0.003*"WCA" + 0.002*"GYR"
0.002*"PVL" + 0.002*"ALE" + 0.002*"ALA" + 0.001*"HPY" + 0.001*"PAL" + 0.001*"KSI" + 0.001*"ADG" + 0.001*"GLG" + 0.001*"AKS" + 0.001*"TSK"
0.040*"QMY" + 0.001*"SLA" + 0.001*"VTV" + 0.001*"GIG" + 0.001*"AAA" + 0.001*"AGI" + 0.001*"GAG" + 0.001*"VTG" + 0.001*"LAA" + 0.001*"AIA"
0.002*"AAA" + 0.002*"AGV" + 0.001*"FQC" + 0.001*"ALA" + 0.001*"LAA" + 0.001*"AIA" + 0.001*"AAL" + 0.001*"AAR" + 0.001*"ELA" + 0.001*"GLE"
0.023*"FHM" + 0.016*"WMA" + 0.009*"EWW" + 0.009*"HMY" + 0.008*"VWM" + 0.006*"DWF" + 0.005*"RWC" + 0.004*"ILW" + 0.004*"HHK" + 0.003*"VWT"
0.009*"DWC" + 0.002*"PFV" + 0.002*"TLA" + 0.002*"ALV" + 0.002*"GTL" + 0.002*"AAV" + 0.002*"AAI" + 0.002*"YVS" + 0.002*"AIA" + 0.002*"ADV"
0.004*"MWK" + 0.003*"MCC" + 0.002*"MHW" + 0.002*"LVA" + 0.001*"AAA" + 0.001*"ALK" + 0.001*"MYT" + 0.001*"AEL" + 0.001*"VKA" + 0.001*"MYY"
0.006*"CGW" + 0.002*"IMC" + 0.002*"WHP" + 0.002*"YDN" + 0.002*"HMI" + 0.002*"LLL" + 0.002*"LLK" + 0.002*"LDN" + 0.002*"TNA" + 0.002*"NID"
0.007*"EQH" + 0.007*"TDN" + 0.006*"TFA" + 0.006*"GHQ" + 0.006*"TGH" + 0.006*"PQA" + 0.006*"WVE" + 0.006*"FIG" + 0.005*"VER" + 0.005*"ENM"
0.006*"AAL" + 0.006*"LAA" + 0.006*"AWF" + 0.005*"AAG" + 0.004*"AAA" + 0.003*"ALA" + 0.003*"LAG" + 0.003*"ADL" + 0.003*"VAA" + 0.003*"LLW"
0.002*"AAA" + 0.002*"VGE" + 0.001*"ALS" + 0.001*"GFD" + 0.001*"GAG" + 0.001*"KVI" + 0.001*"GEG" + 0.001*"GHE" + 0.001*"AAV" + 0.001*"AVA"
0.003*"KLI" + 0.002*"WYT" + 0.002*"LKK" + 0.002*"TFV" + 0.002*"END" + 0.002*"CNF" + 0.002*"VFL" + 0.001*"DII" + 0.001*"RWC" + 0.001*"LLN"
0.002*"ETE" + 0.002*"IIA" + 0.002*"LIK" + 0.002*"KKI" + 0.002*"REE" + 0.002*"EET" + 0.002*"GKA" + 0.002*"LYE" + 0.002*"GVP" + 0.002*"SNN"
0.007*"WNF" + 0.007*"EPF" + 0.005*"SHE" + 0.005*"PFI" + 0.004*"HED" + 0.004*"IER" + 0.004*"NEW" + 0.004*"YIG" + 0.004*"VHY" + 0.004*"AGN"
0.002*"GLE" + 0.001*"GAG" + 0.001*"LAS" + 0.001*"VGL" + 0.001*"VGE" + 0.001*"KGI" + 0.001*"LAR" + 0.001*"AAV" + 0.001*"TGA" + 0.001*"DVL"
0.006*"CTW" + 0.002*"ALA" + 0.001*"LVG" + 0.001*"VGL" + 0.001*"TGA" + 0.001*"GAG" + 0.001*"AIA" + 0.001*"AAL" + 0.001*"WNK" + 0.001*"SAL"
0.017*"WAT" + 0.009*"EHH" + 0.006*"IEH" + 0.006*"FHD" + 0.005*"RLF" + 0.005*"YPP" + 0.004*"DYG" + 0.004*"PWD" + 0.004*"LLR" + 0.004*"CDY"
0.008*"MHG" + 0.007*"LWG" + 0.006*"GSL" + 0.006*"PVS" + 0.006*"ILM" + 0.005*"PVD" + 0.005*"FII" + 0.005*"STM" + 0.004*"AIG" + 0.004*"HGS"
0.002*"WAD" + 0.002*"GGG" + 0.002*"DLI" + 0.001*"LLP" + 0.001*"SAS" + 0.001*"IAF" + 0.001*"GGS" + 0.001*"LMI" + 0.001*"ADL" + 0.001*"VGG"
0.005*"LMY" + 0.004*"HHK" + 0.003*"ETY" + 0.002*"YYE" + 0.002*"FNL" + 0.002*"EYH" + 0.002*"IMI" + 0.002*"WAD" + 0.002*"WYS" + 0.002*"GYF"
0.002*"GAG" + 0.002*"NGA" + 0.002*"AAL" + 0.002*"AAA" + 0.002*"VTV" + 0.001*"AIE" + 0.001*"VAE" + 0.001*"ILD" + 0.001*"AGI" + 0.001*"LAA"
0.002*"WQD" + 0.002*"GPE" + 0.002*"LWP" + 0.002*"GGA" + 0.002*"RMA" + 0.002*"AGG" + 0.002*"PAD" + 0.001*"ADL" + 0.001*"LLL" + 0.001*"PIA"
0.006*"QNN" + 0.003*"MYN" + 0.002*"GSI" + 0.002*"ALA" + 0.002*"AGG" + 0.002*"LPL" + 0.002*"ELA" + 0.002*"LNI" + 0.001*"LAD" + 0.001*"IPY"
0.020*"HMM" + 0.003*"WWN" + 0.002*"NWM" + 0.002*"MGY" + 0.002*"WLF" + 0.002*"IYP" + 0.002*"RDK" + 0.002*"DNW" + 0.002*"LGG" + 0.002*"LRA"
0.011*"WHT" + 0.002*"LGL" + 0.002*"EQL" + 0.002*"LAE" + 0.002*"GLV" + 0.002*"LAG" + 0.002*"RGL" + 0.002*"QLA" + 0.002*"VLA" + 0.001*"LPL"
0.005*"YHY" + 0.005*"PLY" + 0.005*"HTF" + 0.005*"KVQ" + 0.004*"ATD" + 0.004*"THT" + 0.003*"VTH" + 0.003*"HEG" + 0.003*"GVT" + 0.003*"INP"
0.014*"CCY" + 0.002*"AAA" + 0.002*"GAG" + 0.002*"QAA" + 0.002*"ALA" + 0.002*"GGA" + 0.001*"RVL" + 0.001*"AHS" + 0.001*"GVL" + 0.001*"LSL"
0.011*"WQW" + 0.009*"FNQ" + 0.007*"NWI" + 0.007*"WPV" + 0.006*"TWA" + 0.005*"IPT" + 0.003*"HNF" + 0.003*"FQF" + 0.003*"FSA" + 0.003*"NQW"
0.018*"CPC" + 0.008*"IWP" + 0.003*"KFL" + 0.003*"LYW" + 0.003*"NIN" + 0.003*"KIN" + 0.003*"YFI" + 0.003*"QYW" + 0.002*"CFC" + 0.002*"KKN"
0.002*"TDV" + 0.002*"VGE" + 0.002*"GFD" + 0.002*"VSG" + 0.001*"AAL" + 0.001*"GKV" + 0.001*"ALS" + 0.001*"GAR" + 0.001*"VVG" + 0.001*"QSG"
0.015*"CYF" + 0.001*"AAA" + 0.001*"ALA" + 0.001*"LAA" + 0.001*"VGL" + 0.001*"LGL" + 0.001*"AAL" + 0.001*"GFD" + 0.001*"AAV" + 0.001*"KGI"
0.005*"PHM" + 0.002*"CGG" + 0.002*"PCA" + 0.002*"GGG" + 0.002*"GKT" + 0.002*"NNA" + 0.002*"KVY" + 0.002*"HQN" + 0.002*"GGI" + 0.002*"ETD"
0.002*"MAM" + 0.002*"GLG" + 0.002*"EMG" + 0.001*"MPL" + 0.001*"HMD" + 0.001*"PKG" + 0.001*"AGS" + 0.001*"GYM" + 0.001*"GTY" + 0.001*"YFG"
0.002*"AAA" + 0.001*"LLG" + 0.001*"VGL" + 0.001*"GAG" + 0.001*"VAL" + 0.001*"KGI" + 0.001*"AAV" + 0.001*"AAL" + 0.001*"ALA" + 0.001*"AKA"
0.002*"GAG" + 0.002*"AAA" + 0.002*"VGE" + 0.001*"VGL" + 0.001*"KVI" + 0.001*"KAL" + 0.001*"AAV" + 0.001*"GLD" + 0.001*"VLI" + 0.001*"GLG"
0.003*"AAA" + 0.001*"VGL" + 0.001*"VGE" + 0.001*"EVA" + 0.001*"AVA" + 0.001*"GDR" + 0.001*"ELG" + 0.001*"ALS" + 0.001*"VIL" + 0.001*"SGL"
0.077*"PWW" + 0.002*"GGA" + 0.001*"GAG" + 0.001*"AGV" + 0.001*"LVG" + 0.001*"GLG" + 0.001*"GVV" + 0.001*"APL" + 0.001*"GIG" + 0.001*"ALS"
0.002*"AAA" + 0.001*"LAA" + 0.001*"TGA" + 0.001*"AAG" + 0.001*"ALA" + 0.001*"VTG" + 0.001*"AAV" + 0.001*"AVA" + 0.001*"VAL" + 0.001*"EVA"
0.007*"TCK" + 0.006*"WGT" + 0.005*"HKG" + 0.005*"ECK" + 0.004*"AGA" + 0.004*"GVA" + 0.004*"IST" + 0.004*"GTS" + 0.004*"HEV" + 0.004*"PEG"
0.002*"AAA" + 0.002*"KAA" + 0.002*"GLD" + 0.002*"ADA" + 0.001*"VGL" + 0.001*"ALA" + 0.001*"GAG" + 0.001*"AAL" + 0.001*"AAG" + 0.001*"NAA"
0.002*"MGF" + 0.002*"FVN" + 0.002*"GGG" + 0.002*"TDV" + 0.002*"AFN" + 0.002*"LGA" + 0.001*"DKA" + 0.001*"GKD" + 0.001*"LDK" + 0.001*"GAG"
0.009*"NKC" + 0.009*"QYD" + 0.007*"SHW" + 0.007*"YEW" + 0.006*"NNS" + 0.005*"AWY" + 0.004*"FAP" + 0.004*"YAG" + 0.004*"FNN" + 0.004*"TNE"
0.002*"LIA" + 0.002*"VDL" + 0.002*"LIY" + 0.002*"PWI" + 0.002*"LKK" + 0.002*"IVD" + 0.002*"VIG" + 0.002*"YLK" + 0.002*"GFN" + 0.002*"GVP"
0.002*"EET" + 0.001*"GGG" + 0.001*"WFT" + 0.001*"ALK" + 0.001*"FGV" + 0.001*"AAL" + 0.001*"VIG" + 0.001*"AVL" + 0.001*"FET" + 0.001*"AGV"
```

Fig. 2. Example of 20 topics produced by an amino acid topic model. Note similarities between amino acid trigrams grouped in the same topic, where each row is a list of amino acid trigrams in the same topic.

III. Results

This section will present the results of both the evaluation of trained topic models and the analysis of the topic distributions of oxidoreductase and RNA helicase sequences.

A. Model evaluation

Generally, trained topic models with varying topic number parameters performed poorly in terms of UMass coherence and log perplexity. Within this generally negative performance, there was a slight but observable inverse correlation between coherence and perplexity scores and topic number, suggesting models generating fewer individual topics performed better as topic models for this dataset. The relationship between topic number and perplexity and coherence scores is described in Figure 1. Perplexity, as described in the previous section, is an evaluation of probability distributions, and as a result is less useful in this context since Gensim's implementation and reported results generally masks the construction and use of these probability distributions. Coherence is a measure of how interrelated the constituent words of topics are, which is more readily comparable with Gensim's outputs, specifically the display of topic constituents as shown in Figure 2. The generally poor UMass coherence metrics of the tested models (including the 20-topic model whose generated topics are displayed in Figure 2) was surprising, given the apparent relationships between amino acid trigrams in the generated topics; most topics in Figure 2 contain pairs of amino acid trigrams that could obviously overlap as trigrams in a full sequence (e.g. "ALL" and "LLG").

B. Testing on enzyme groups

Testing sequences from the two selected EC classifications on trained topic models and comparing resulting topic distributions of constituent amino acid sequences resulted in some variability among topic distributions of same-class sequences, but certain topics stood out as predominant among sequences in each group. Figure 3 displays stacked bar graphs of topic distributions of EC oxidoreductase sequences extracted from 10- and 100-topic models as a comparison of both extremes of the range topic numbers tested. This analysis shows that topic distributions extracted from both models show a few significant topics that occur in the group's sequences; this is more significant in the case of the 100-topic model, as the number of significant topics (as shown by Figure 3b) is significantly less than the total topic number of 100. In addition to intra-EC class topic distribution variation by topic number, inter-EC class topic distribution variation was analyzed while keeping the number of generated topics constant. (In practice, a 20-topic model was used, in order to provide greater topic specificity while remaining on the better-performing end of the range of model evaluation results.) Figure 4 compares the 20-topic distributions of 109 oxidoreductase sequences and 1,549 RNA helicase sequences. This figure suggests that the same set of generated topics dominates among sequences in both EC classes (which are notably and intentionally different in function, and by biological reasoning, likely in structure). However, there are differences in the degree to which these dominant topics are most prevalent. Topic 1 dominated in roughly 61% of oxidoreductase sequences but only roughly 30% of RNA helicase sequences, while topic 14 was dominant in roughly 40% of RNA helicase sequences but only a quarter of oxidoreductase sequences (Figure 5).

IV. Discussion

These results suggest mixed potential for topic modeling to enable standardized enzyme (and other amino acid sequence) fingerprinting, but do demonstrate promising ability of topic modeling to act as a method of amino acid sequence similarity.

Of greatest concern is the relatively poor performance of the trained topic models in terms of coherence and perplexity measures, fairly standard and reliable metrics common in the world of topic modeling. This performance can be understood as poor performance of the overall topic model, but they may also imply that such metrics are not useful when building topic models on datasets of amino acid sequences as processed and studied in this project. The low coherence scores generated by the tested models contrasts with the observable similarity (and significant potential for overlap in full sequences) of trigrams composing generated topics as shown in Figure 2.

More promising results were generated by the analysis of topic distributions of sets of oxidoreductase and RNA helicase sequences. These sequence sets were chosen due to their significantly different taxonomic EC designations and biological functions; accordant to the structure-function principle in biology, these sets' constituent sequences may vary greatly in structure. Figure 3 demonstrates that even with varying numbers of topics generated, similar subsets of topics are emphasized in sequences across the oxidoreductase family studied. Figure 5 shows that while similar subsets of topics are emphasized across sequences in both the oxidoreductase and RNA helicase sequence sets (using the same topic model), there is noticeable variation in the degree to which such topics are emphasized in members of each sequence set, perhaps

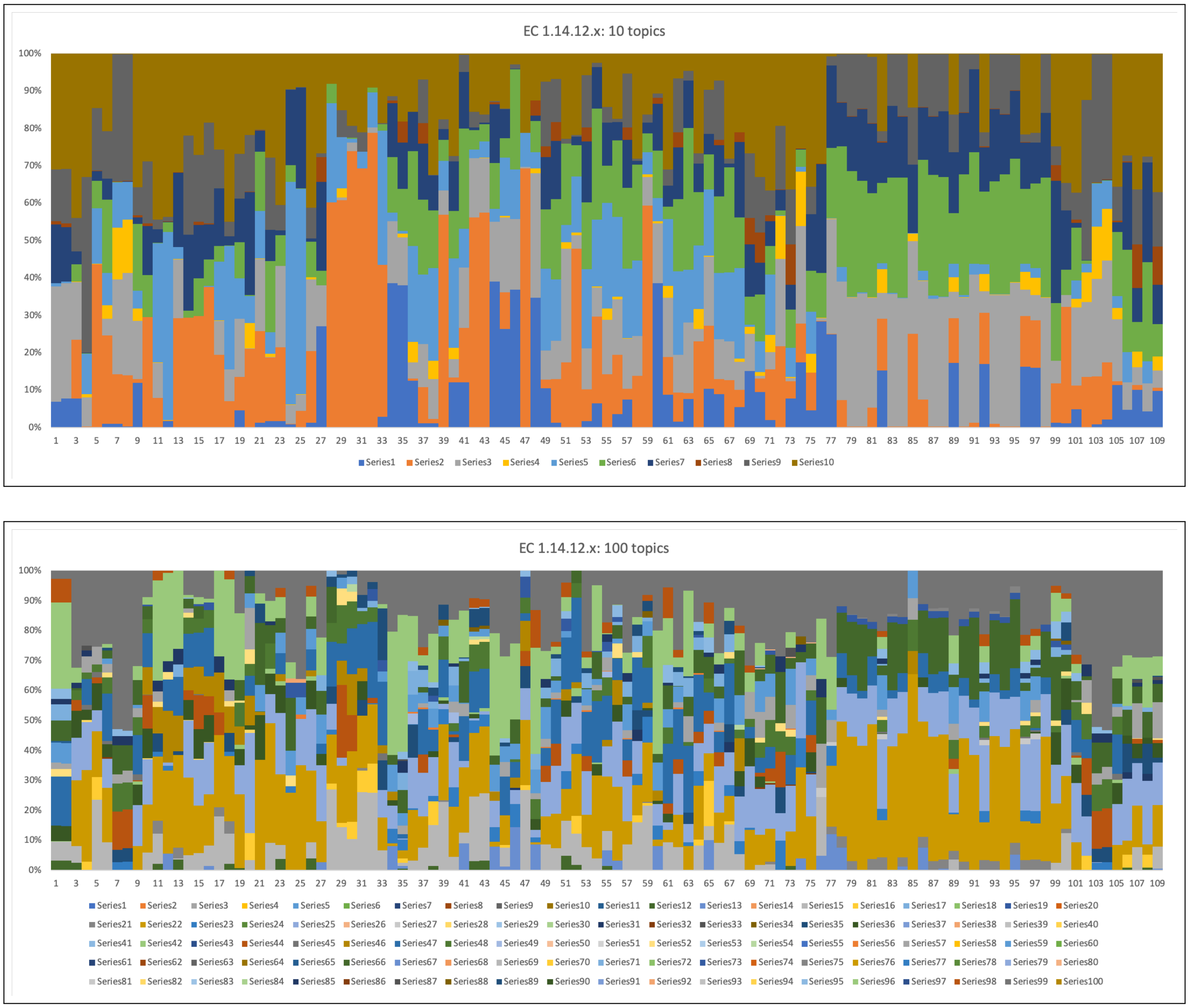


Fig. 3. (a) Topic distributions of 109 oxidoreductase amino acid sequences given by a 10-topic model (top), and (b) topic distributions of the same set of amino acid sequences given by a 100-topic model (bottom).

reflecting intra-family similarities between oxidoreductase and RNA helicase sequences that differ between the two enzyme families. This suggests that the topic model's generated topics and topic distributions may be reflecting different amino acid sequence-level signatures or commonalities shared by sequences in these two families. In addition, clusters of highly similar topic distributions are visible in the oxidoreductase distribution sets in figures 3 and 4, which correlated with highly similar oxidoreductase sequences in the dataset. These datasets show promise in this method of topic modeling's ability to reflect amino acid sequence similarity, a desired feature and widely-used application of fingerprinting techniques that have been developed for small organic molecules.

V. Future work

Due to the exploratory nature and scope of this project, significant future steps can be taken to further develop topic modeling techniques for (and adjust them to) amino acid datasets. This project adapted its topic modeling pipeline from a purely text-mining application; features of text-mining topic modeling may not be applicable to datasets of amino acid sequences (e.g. coherence and perplexity evaluation metrics). In order to assess the validity of the direct analogy established in this project, further analysis of amino acid datasets should be conducted. This project's direct application of text-mining topic modeling relies on the assumption that co-occurence patterns between words in natural languages is similar to co-occurence between small n-grams of amino acids, and that the role of word co-occurrence in determining the semantic profile of a whole document is analogous to the role of amino acid n-grams (and the primary structure of enzymes more broadly) plays in determining enzyme function. The latter of these assumptions is especially important to validate when using a hypothetical amino acid fingerprinting technique to make inferences about enzyme function. Analysis of patterns in primary structures of enzymes should be conducted in order

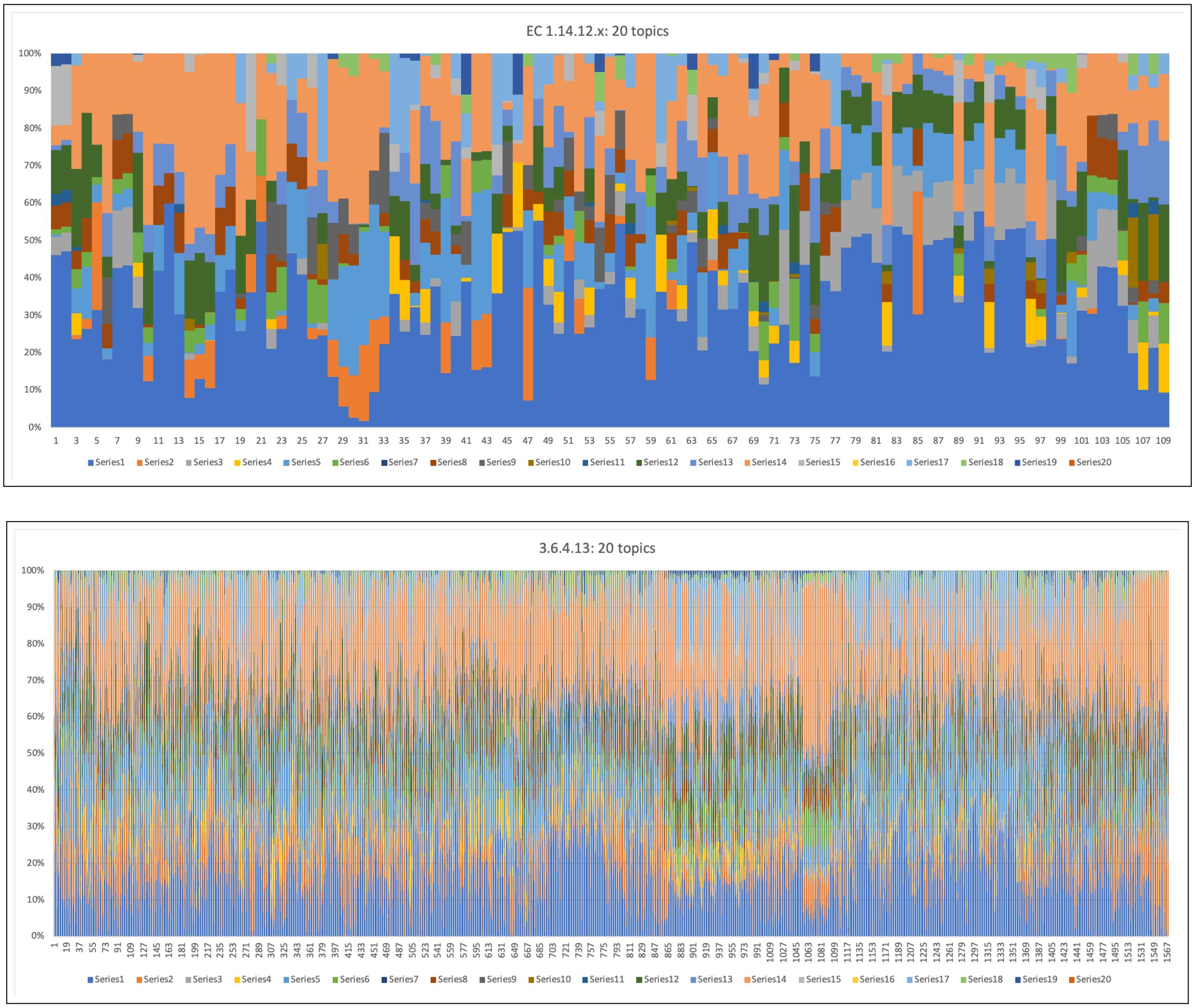


Fig. 4. (a) Topic distributions of 109 oxidoreductase sequences as given by a 20-topic model (top), and (b) topic distributions of 1,569 RNA helicase sequences as given by the same 20-topic model (bottom).

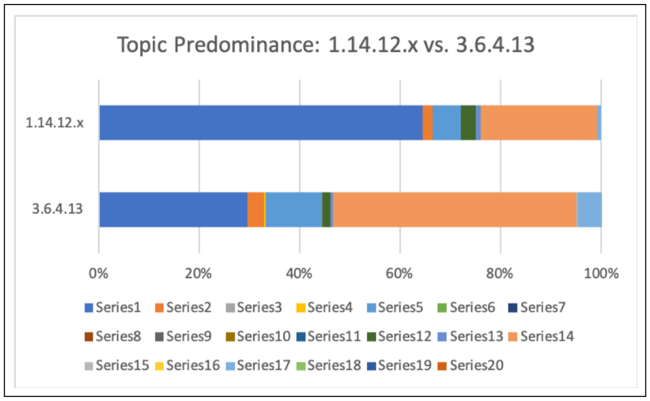


Fig. 5. Relative share of predominant topics across oxidoreductase and RNA helicase amino acid sequence datasets.

to answer this question.

If amino acid datasets like the one used in this study are determined to be suitable for topic modeling, other implementations of topic modeling. This project uses a specific topic model construction algorithm (latent Dirichlet allocation) and an even more specific implementation of this algorithm (as provided by Gensim). It is possible that other methods of topic modeling are more applicable to amino acid datasets and comparative analysis should be conducted in order to identify possibly better-informative candidate methods. In addition, this project uses a specific method of subdividing amino acid sequences. Trigrams could be too-specific units of amino acid sequences that carry litte “semantic” value, and there may be more nuanced (non n-gram) methods of subdividing amino acid sequences into structurally and functionally significant “words” that could inform a topic model. In a similar vein, topic modeling applied to natural language datasets usually preprocesses texts to prune universally-occurring words (in practice, word classes like pronouns, articles, prepositions, etc.) that carry little semantic value; collectively, such words are called “stopwords.” Further research on and analysis of amino acid sequences could be conducted to identify amino acid analogs of linguistic stopwords. Pruning stopwords from datasets serves to make resulting tpic models more semantically coherent, and a similar approach to amino acid datasets would result in a more nuanced language-amino acid sequence analogy and potentially more robust topic model.

Once a sufficiently informative topic model can be developed for amino acid sequences, many routes of future work exist beginning with the topic distributions the model would produce for given sequences. Ultimately, the goal of an amino acid sequence topic model would be to generate standard fingerprint representations of sequences in order to enable efficient and informative profiling and comparison of the enzymes that the sequences encode. Applications of such a representation are abundant, including the construction of fingerprint datasets to enable classification of novel amino acid sequences (and potential resulting functional characterization) and the use of fingerprints as feature vectors to construct machine learning classifiers that can further identify characteristics of and relationships between enzymes[3].

Acknowledgments

I’d first like to thank Professor Soha Hassoun for her guidance in developing the idea and development of this project, and for the computational systems biology course she conducted that enabled this project to come to fruition. I’d also like to thank Sara Amr Amin, a student of Prof. Hassoun, for her generosity in sharing the enzyme sequence dataset that this project relied on. Lastly, I’d like to recognize Andrew Valenti, a graduate student in Tufts’ Human-Robot Interaction Laboratory, for introducing me to and helping me develop my skillset in the field of natural language processing, a research passion of mine that allowed me to develop this project in the first place.

References

[1] Weininger, D. ”SMILES, a Chemical Language and Information System.” *J. Chem. Inf. Comp. Sci.* 1988, 28: 31-36.

[2] Blei, D. ”Probabilistic Topic Models.” *Communications of the ACM* 2012, 55(4): 77-84.

[3] Schneider, N., Fechner, N., Landrum, G., Stielf, N. ”Chemical Topic Modeling: Exploring Molecular Datasets Using a Common Text-Mining Approach.” *J. Chem. Inf. Model.* 2017, 57(8): 1816-1831.

[4] Rosner, F., Hinneburg, A., Röder, M., Nettling, M., Both, A. ”Evaluating topic coherence measures.” *Neural Information Processing Systems*, 2013.

[5] Amin, S.A., Endalur Gopinarayanan, V., Nair, N.U., Hassoun, S. ”Establishing synthesis pathway-host compatibility via enzyme solubility.” *Biotechnology and Bioengineering*, 2019, 116: 1405-1416.

[3] Another student in this semester’s CS150 class proposed and tested the use of n-gram frequencies as features to enable the machine learning classification of enzymes and non-enzymes; this is one example of an application where amino acid sequence topic vectors could be used to construct such classifiers.