# Assessment of Dense Word Representations for Text Classification in Biocuration of Infectious Disease

Daniel E. Sullivan

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
In
Genetics, Bioinformatics, and Computational Biology

Alice R. Wattam
David R. Bevan
Stefan Hoops
Madhav Marathe

April 28, 2016
Blacksburg, Virginia

Keywords: text mining, machine learning, biocuration, linguistics, natural language processing

Assessment of Dense Word Representations for
Text Classification in Biocuration of Infectious Disease
Daniel E. Sullivan

ABSTRACT (Academic)

This research addresses the problem, can unsupervised learning generate a representation that improves on the commonly used term frequency-inverse document frequency (TF-IDF ) representation by capturing semantic relations? The analysis measures the quality of sentence classification using term TF-IDF representations, and finds a practical upper limit to precision and recall in a biomedical text classification task (F1-score of 0.85). Arguably, one could use ontologies to supplement TF-IDF, but ontologies are sparse in coverage and costly to create. This prompts a correlated question: can unsupervised learning capture semantic relations at least as well as existing ontologies, and thus supplement existing sparse ontologies? A shallow neural network implementing the Skip-Gram algorithm is used to generate semantic vectors using a corpus of approximately 2.4 billion words. The ability to capture meaning is assessed by comparing semantic vectors generated with MESH. Results indicate that semantic vectors trained by unsupervised methods capture comparable levels of semantic features in some cases, such as amino acid (92% of similarity represented in MESH), but perform substantially poorer in more expansive topics, such as pathogenic bacteria (37.8% similarity represented in MESH). Possible explanations for this difference in performance are proposed along with a method to combine manually curated ontologies with semantic vector spaces to produce a more comprehensive representation than either alone. Semantic vectors are also used as representations for paragraphs, which, when used for classification, achieve an F1-score of 0.92. The results of classification and analogical reasoning tasks are promising but a formal model of semantic vectors, subject to the constraints of known linguistic phenomenon, is needed. This research includes initial steps for developing a formal model of semantic vectors based on a combination of linear algebra and fuzzy set theory subject to the semantic molecularism linguistic model. This research is novel in its analysis of semantic vectors applied to the biomedical domain, analysis of different performance characteristics in biomedical analogical reasoning tasks, comparison semantic relations captured by between vectors and MESH, and the initial development of a formal model of semantic vectors.

Assessment of Dense Word Representations for
Text Classification in Biocuration of Infectious Disease

Daniel E. Sullivan

ABSTRACT (Public)

Advances in life sciences, including genomics and related fields, are detailed in the scientific literature. An unfortunate consequence of having a large amount of biomedical literature, is that it is difficult for researchers to find specific information without spending extended periods of time search in databases such as PubMed and PubMed Central. This research evaluates commonly used techniques based on simple word statistics as well as a new from of word and document representation known as dense word representations. The latter are particularly well suited to capturing the meanings of words, phrases and paragraphs. The research demonstrates that dense word representations perform better than simple statistical techniques and complement other resources, such as biomedical ontologies. The dissertation also discusses an approach to developing a mathematical formalism that captures both linguistic properties of language and structural properties of dense word representations.

To Katherine

Table of Contents

# Chapter 1: Introduction

The focus of the research described here is to evaluate the use of dense word representations derived by unsupervised learning algorithms. The evaluation will consist of measuring the quality of sentence classification using term frequency-inverse document frequency (TF-IDF) representations, a statistical approach that does not take into account the semantics of word or phrases. [Salton, 1975] An alternative representation scheme using dense vectors, referred to as semantic vectors in this research, is evaluated. Semantic vectors are designed to capture some semantics of words by using vectors which are in close proximity to the vectors of semantically related words. The ability to capture meaning is assessed by comparing semantic vectors generated using unsupervised learning techniques with manually curated ontologies. Semantic vectors are also used to generate vector representations for paragraphs. These vector representations are evaluated using a classification task. Furthermore, this research includes initial steps for developing a formal model of semantic vectors based on a combination of linear algebra and fuzzy set theory. This research is novel in its objective of focusing on biocuration for infectious bacterial diseases and evaluating dense word representations for each task and comparing results to commonly used existing techniques. This proposed research has three specific aims.

## Aims of Research

There are three primary aims of this research:

.Evaluate text classifiers using TF-IDF representation and multiple machine learning algorithms.

.Evaluate semantic vector representations and their ability to capture word meaning, which is measured by comparing automatically generated word vectors with comparable terms in manually curated ontologies.

.Evaluate the quality of paragraph classifiers using semantic vectors for paragraphs.

In addition the three specific aims, this research begins to outline a formal model of semantic vectors using linear algebra and fuzzy set theory. This is a novel approach to semantic vectors; the author is unaware of any similar attempts to combine these two formalisms and apply them to linguistic phenomenon.

This research also provides an analysis of characteristics of dense word vectors and clusters of dense

word vectors and their relation to the quality of classification and information extraction. Together, these specific aims will enhance the understanding of dense word vectors as a representation scheme for improving the quality of classification and information extraction.

The motivation for this research is the hypothesis for the need of broad, semantic representations of biomedical terms, which has not yet been met. As noted by one researcher:

> Broad coverage semantic taxonomies such as WordNet (Felbaum, 1998) and CYC (Lenat, 1995) have been constructed by hand at great cost; while a crucial source of knowledge about the relation between words, these taxonomies still suffer from sparse coverage.[Snow, 2006]

Many text mining applications use the term frequency inverse document frequency (TF-IDF) document model. While useful in many areas, the conventional implementation of TF-IDF based classifiers fail to take into account semantic information contained with a text. The overall objective of this research is to assess the value of representations that include semantic features when applied to word and paragraph representations.

## Contents of this Dissertation

The remainder of this dissertation consists of five chapters.

Chapter 2 examines the limits of TF-IDF representation schemes. Specifically, the experiments described in that chapter demonstrate that TF-IDF can be used with multiple types of supervised machine learning algorithms but there is a limit to the quality of classification results, measured in terms of accuracy and recall. After training with a sufficiently large training set, adding training examples does not improve the quality of classification results regardless of the algorithm used. This finding motivates the exploration of representation schemes that capture semantic properties.

Chapter 3 describes semantic vectors generated with the Skip-gram algorithm. This algorithm assumes the Distribution Hypothesis, which is the concept that the meaning of words is defined and can be inferred from the other words used in its proximity. The chapter includes an evaluation of semantic properties captured by semantic vectors relative to semantic properties captured in ontologies. The experiments in this chapter demonstrate that semantic vectors and ontologies capture overlapping information when entities have relatively simple properties (e.g. amino acids) and capture complementary features when more complex entities (e.g. bacteria) are compared.

Chapter 4 examines the use of the Skip-gram algorithm for generating semantic vectors for paragraphs. One of the insights gleaned from conducting the experiment in Chapter 2 is that classifying text at the sentence level may not be the optimal approach. Researchers often need multiple sentences to explain complex phenomenon, such as virulence factors or pathogenic features of an organism. In such cases, paragraph level classification may be more appropriate. The experiments in this chapter demonstrate that semantic vectors can produce high quality classification results (precision and recall greater than 0.90).

Chapter 5 outlines a proposed approach to developing a formal model of semantic vectors based on a combination of linear algebra and fuzzy set theory. Arguments are presented that demonstrate the proposed formalism addresses both complex linguistic phenomenon, i.e. the "family resemblance" feature of word definitions, as well as the imprecision introduced as an artifact of computing semantic vectors.

Chapter 6 discusses five topics of further research:

1. Assessing Ensemble methods combining TF-IDF and semantic vector methods.
2. Representing knowledge represented in ontologies in semantic vectors.
3. Examining the macro and meso-structures of vector space.
4. Developing a fuzzy linear algebra of semantic vectors.
5. Evaluate user interface design considerations for building biocuration systems based semantic vectors.

# References


Aziz RK, B.D., Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O., *The RAST Server: Rapid Annotations using Subsystems Technology*. BMC Genomics, 2008.

Cohen, A.M. *An effective general purpose approach for automated biomedical document classification*. in *AMIA Annual Symposium Proceedings*. 2006. American Medical Informatics Association.

Gene Ontology, C., *The Gene Ontology (GO) database and informatics resource*. Nucleic acids research, 2004. **32**(suppl 1): p. D258-D261.

Goldberg, Y. and O. Levy, *word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method*. arXiv preprint arXiv:1402.3722, 2014.

Hobbs, J.R., *Information extraction from biomedical text*. Journal of Biomedical Informatics, 2002. **35**(4): p. 260-264.

Jimeno-Yepes, A.J., et al., *GeneRIF indexing: sentence selection based on machine learning*. BMC bioinformatics, 2013. **14**(1): p. 171.

Kim, J.-D. and S. Pyysalo, *BioNLP Shared Task*, in *Encyclopedia of Systems Biology*. 2013, Springer. p. 138-141.

Mao C, Abraham D, Wattam AR, et al. Curation, integration and visualization of bacterial virulence factors in PATRIC. *Bioinformatics*. 2015;31(2):252-258. doi:10.1093/bioinformatics/btu631.

Mikolov, T., et al., *Efficient estimation of word representations in vector space*. arXiv preprint arXiv:1301.3781, 2013.

Moen, S.P.F.G.H. and T.S.S. Ananiadou, *Distributional Semantics Resources for Biomedical Text Processing*.

Pyysalo S, O.T., Cho H-, Sullivan D, Mao C, Sobral B, et al. , *Towards Event Extraction from Full Texts on Infectious Diseases. p. 132-40*. Proceedings of the 2010 Workshop on Biomedical Natural Language Processing. Uppsala, Sweden: Association for Computational Linguistics, 2010: p. 132-140.

Salton, G., A. Wong, and C.-S. Yang, *A vector space model for automatic indexing*. Communications of the ACM, 1975. **18**(11): p. 613-620.

Settles, B. *Biomedical named entity recognition using conditional random fields and rich feature sets*. in *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*. 2004. Association for Computational Linguistics.

Snow, R., D. Jurafsky, and A.Y. Ng. *Semantic taxonomy induction from heterogenous evidence*. in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual*

*meeting of the Association for Computational Linguistics*. 2006. Association for Computational Linguistics.

Tsuruoka, Y., et al., *Developing a robust part-of-speech tagger for biomedical text*, in *Advances in informatics*. 2005, Springer. p. 382-392.

Wattam, A.R., et al., *PATRIC, the bacterial bioinformatics database and analysis resource*. Nucleic acids research, 2013: p. gkt1099.

Müller, H.-M., E.E. Kenny, and P.W. Sternberg, *Textpresso: an ontology-based information retrieval and extraction system for biological literature*. PLoS biology, 2004. **2**(11): p. e309.

# Chapter 2: Limitations of TF-IDF for Classifying Sentences in Biocuration Tasks


## Abstract

The sequencing of large numbers of bacterial genomes has enabled the development of automated annotations systems but the curation of specialty gene sets, such as virulence factors and antibiotic resistance genes, still largely depend upon manual curation. This study evaluates the use of machine learning algorithms and the term frequency inverse document frequency (TF-IDF) representation to support classification of sentences from the biomedical literature. The study uses 13,978 sentences extracted from 1,127 papers on bacterial virulence factors. Nine supervised classification algorithms are used: support vector machines (SVMs), ridge classifier, perceptron, kNN, SGD, nearest centroid, Naïve Bayes, random forest, and ADABoost. The semi-supervised classification algorithm, label spreading, is also evaluated. SVMs achieve the best classification performance in these experiments (F1-score = 0.85).. Applying alternative supervised and semi-supervised algorithms did not improve results, as measured by F-score. In addition, analysis of error curves indicates that training on additional data is not likely to improve performance. If additional data and changes to algorithms will not lead to improved performance, we concluded that an alternative representation is required to realize substantial improvement in biocuration classification tasks.


## Introduction

**T**he overall biocuration process consists of three steps: identify papers relevant to the biocuration task; identify sentences describing the functional characteristics or biological process of the genes in the specialty gene set; and extract information from those sentences and map it to a structured format. In this study, we examine techniques for identifying assertion sentences about virulence within papers that have been previously classified as relevant to the biocuration task. There has been significant research on document classification. [Griffiths, 1984], [Willett 1988], [Han and Karypis 2000], and [Turney 2002]) Automatically extracting structured information from sentences is a more challenging problem. Although there have been some success, the precision and recall of current event extraction systems applied to biological processes, such as signaling pathways, are inadequate to support automated biocuration [Pyysalo, 2010].

## Methods

The data for this study consists of 13,978 sentences extracted from 1,127 papers on bacterial virulence factors. 4,696 sentences are positive examples of virulence factor assertion sentences manually identified by PATRIC curators [Mao, 2015].

| Genera | Virulence Factor Genes | Positive Examples of Virulence Factor Assertion Sentence | Unique Publications |
|---|---|---|---|
| *Escherichia* | 312 | 474 | 175 |
| *Listeria* | 122 | 608 | 170 |
| *Mycobacterium* | 345 | 965 | 202 |
| *Salmonella* | 327 | 2097 | 417 |
| *Shigella* | 112 | 552 | 163 |
| **Total** | **1,218** | **4,696** | **1,127** |

**Table 2.1. Composition of training set by genera.**

Positive examples of virulence factor assertion sentences typically include reference to a gene and a phenotype or molecular function such as:

“Mutations in the fimH gene of *Salmonella typhimurium* result in a non-fimbriate, non-adhesive phenotype.” [Hancox, Yeh et al. 1997]

“Unexpectedly, here we find that nonacylated LprG retains TLR2 activity.”[Drage, 2010]

“The autolysin Ami contributes to the adhesion of Listeria. “ [Milohanic, Jonquieres. 2001]

Negative examples were randomly selected from sentences from the same set of publications. Potential negative examples were compared to positive examples. Any potential negative example with an edit distance of less than 25 to a positive example were manually reviewed to avoid including the same sentence in both positive and negative example sets.

Sentences are represented a weighted vectors of terms using the term frequency inverse document frequency (TF-IDF) representation. TF-IDF is a bag of words model that represents the relative frequency of words within a document relative to their frequency across an entire corpus. In this experiment, sentences are the units of texts and the corpus is the set of all sentences. TF-IDF ignores

the syntactic structure of and semantic relations within sentences. TF-IDF representations have been successfully used in text classification tasks, including short text classification. [Robertson, 2004]

Three sets of experiments are described: an initial classification experiment using Support Vector Machine (SVM) to evaluate the impact of training data sized on classification; a second set of experiments using alternative supervised machine learning algorithms to assess the impact of machine learning algorithm on the quality of classification; and experiment using a semi-supervised learning algorithm to assess the impact of using non-labeled training data to train a classifier.

**Experiment 1: Support Vector Machines**

The Support Vector Machine (SVM) machine learning algorithm is used to classify sentences into one of two categories: virulence factor assertion sentences or non-assertion sentences. SVM is a large margin classifier that has been successfully used in text classification tasks. [Joachims 1998] The Scikit-Learn Python implementation of SVM was used in these experiments [Pedregosa, 2011). The C value parameter of SVM controls the level of tolerated misclassification. For this experiment the C value was set to 0.25 due to marginally improved performance over the default setting.

This classification problem entails an imbalanced data set with many more negative instances than positive instances. Too few negative instances risks under-representing the space of negative instances; too many negative instances can decrease the performance of the classifier. To determine a the number of negative instances relative to positive instances, a set of four data sets was constructed for each genera. The four had 1x, 2x, 5x and 10x as many negative instances as positive instances. Each data set was split into training and test subsets with 80% and 20% of instances, respectively. The 1x data set consistently performed better than larger numbers of negative instances when measuring area under the curve, precision, recall and F-score. The remaining experiments are preformed using 1x data sets.

| Data Set | Negative Example Factor | AUC | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| All | 1 | 0.86 | 0.86 | 0.89 | 0.87 |
| All | 2 | 0.84 | 0.79 | 0.79 | 0.79 |
| All | 5 | 0.75 | 0.72 | 0.55 | 0.62 |
| All | 10 | 0.68 | 0.69 | 0.37 | 0.48 |
| *Escherichia* | 1 | 0.83 | 0.79 | 0.87 | 0.83 |

| | | | | | |
|---|---|---|---|---|---|
| *Escherichia* | 2 | 0.79 | 0.68 | 0.72 | 0.70 |
| *Escherichia* | 5 | 0.70 | 0.75 | 0.43 | 0.55 |
| *Escherichia* | 10 | 0.58 | 0.88 | 0.16 | 0.27 |
| *Listeria* | 1 | 0.89 | 0.89 | 0.89 | 0.89 |
| *Listeria* | 2 | 0.86 | 0.78 | 0.82 | 0.80 |
| *Listeria* | 5 | 0.71 | 0.69 | 0.46 | 0.55 |
| *Listeria* | 10 | 0.68 | 0.90 | 0.36 | 0.51 |
| *Mycobacterium* | 1 | 0.82 | 0.75 | 0.87 | 0.81 |
| *Mycobacterium* | 2 | 0.79 | 0.74 | 0.71 | 0.73 |
| *Mycobacterium* | 5 | 0.69 | 0.70 | 0.42 | 0.52 |
| *Mycobacterium* | 10 | 0.60 | 0.68 | 0.20 | 0.32 |
| *Salmonella* | 1 | 0.84 | 0.82 | 0.88 | 0.85 |
| *Salmonella* | 2 | 0.83 | 0.79 | 0.77 | 0.78 |
| *Salmonella* | 5 | 0.75 | 0.77 | 0.54 | 0.64 |
| *Salmonella* | 10 | 0.65 | 0.70 | 0.32 | 0.44 |
| *Shigella* | 1 | 0.89 | 0.91 | 0.86 | 0.89 |
| *Shigella* | 2 | 0.84 | 0.74 | 0.78 | 0.76 |
| *Shigella* | 5 | 0.66 | 0.71 | 0.35 | 0.47 |
| *Shigella* | 10 | 0.62 | 0.78 | 0.25 | 0.37 |

**Table 2.2. Results with SVM classifier and alternate ratios of positive to negative training examples.**

**Experiment 2: Alternative Supervised Machine Learning Algorithms**

In the second experiment, 8 additional machine learning algorithms were employed: ridge classifier, perceptron, kNN, SGD, nearest centroid, Naïve Bayes, random forest, and ADABoost. This set of algorithms was selected to evaluate different types of machine learning algorithms, including linear classifiers, probabilistic classifiers, ensemble algorithms, and boosting techniques.

The same data set used in the SVM experiment was used with the additional eight supervised machine learning algorithms.

**Experiment 3: Semi-Supervised Machine Learning Algorithm**

The third experiment employed the label spreading algorithm. [Zhou, 2004] Semi-supervised learning algorithms use both labeled and unlabeled training data. Semi-supervised techniques are promising because they might lead to higher quality classifications than supervised techniques without requiring additional labeled data, which is costly and time consuming to collect in biocuration tasks.

For this experiment, the data consisted of 842 labeled sentences and 4,346 randomly selected unlabeled sentences.

# Results

### SVM Only Experiment

5-fold validation was used to train and evaluate the model, i.e. the model was trained with 80% of the example set and tested with the remaining 20%. The measures in Table 2 are the average of five training/validation iterations. The results are measured using precision, recall, F-score and area under the curve (AUC).

| Data Set | Precision | Recall | F1-score | AUC |
|---|---|---|---|---|
| All | 0.91 | 0.91 | 0.91 | 0.91 |
| *Escherichia* | 0.81 | 0.89 | 0.85 | 0.84 |
| *Listeria* | 0.85 | 0.88 | 0.86 | 0.86 |
| *Mycobacterium* | 0.80 | 0.86 | 0.83 | 0.82 |
| *Salmonella* | 0.83 | 0.89 | 0.86 | 0.85 |
| *Shigella* | 0.83 | 0.89 | 0.86 | 0.86 |
| Mean | 0.83 | 0.88 | 0.85 | 0.85 |

**Table 2.3. Results of SVM classifier evaluation.**

One way to improve model performance is to train the model with additional examples. This method can improve performance unless there is a bias in the model.[Haussler, 1988] In such cases, the features and representation of the model is insufficient to capture characteristics of the training set that are required to improve model performance. An indication of bias in the model is a lack of improvement in model performance with the addition of more examples.

Learning curves indicate that additional training examples will not improve metrics. Learning curves are graphs of training error and validation error for a set of models built using training sets of increasing size. Training error is defined as the rate of incorrectly classified examples when the same examples are used for both training the model and applying the model. Validation error is defined as the rate of incorrectly classified examples when different examples are used for training and classifying.

For each data set, an SVM model was trained with an increasing number of training examples using 5% increments of the total training set. The training error and validation error for each data set are shown

in Figure 1. As expected, the training error is consistently lower than validation error. If the use of additional training examples improves model performance, the validation error will approach the training error.

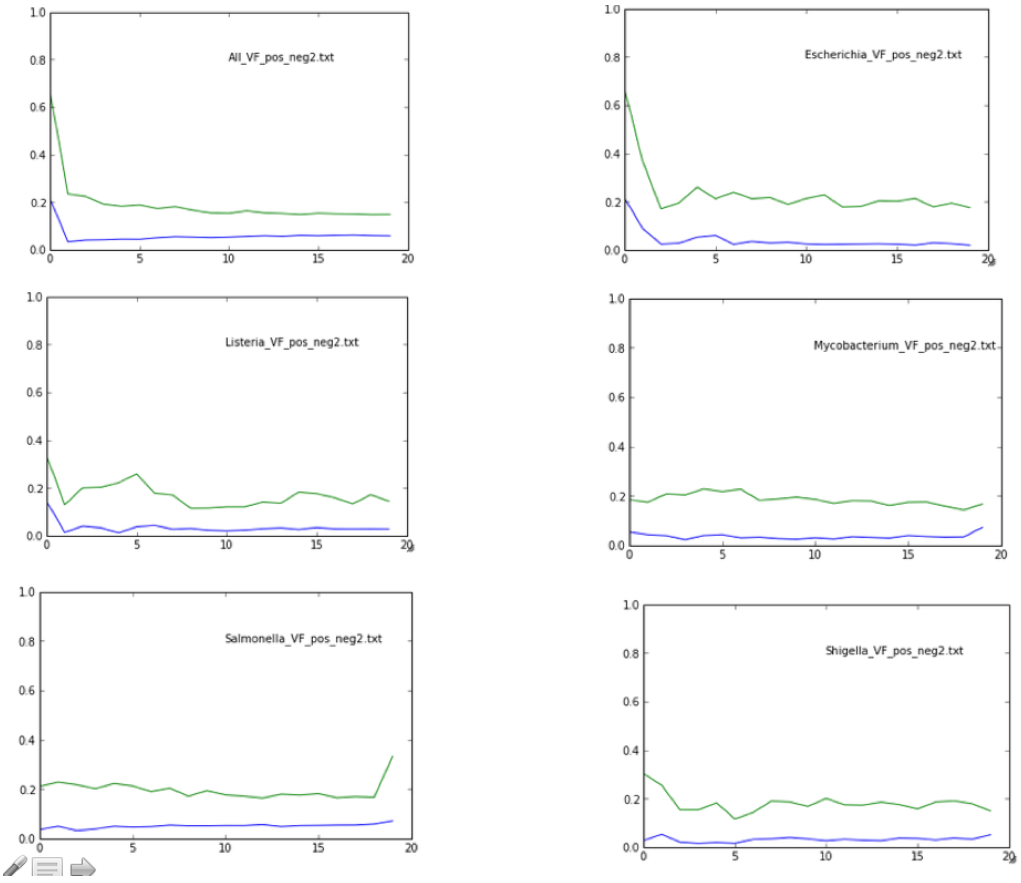


**Figure 2.1. Error rates do not necessarily improve with ddditional training examples. X axis represents the number of training examples and the Y-axis represents the error rate.**

### Alternate Supervised Machine Learning Algorithm Experiments

The eight alternative machine learning algorithms did not significantly improve upon the performance of the SVM classifier. The ridge classifier, another linear classifier, and Naïve Bayes performed the best.

| Algorithm | Precision | Recall | F1-score |
|---|---|---|---|
| Ridge Classifier | 0.88 | 0.88 | 0.88 |

| | | | |
|---|---|---|---|
| Perceptron | 0.86 | 0.86 | 0.86 |
| kNN | 0.30 | 0.55 | 0.39 |
| SGD | 0.87 | 0.87 | 0.87 |
| Nearest Centroid | 0.84 | 0.84 | 0.84 |
| Naïve Bayes | 0.88 | 0.88 | 0.88 |
| Random Forest | 0.71 | 0.70 | 0.70 |
| AdaBoost | 0.78 | 0.78 | 0.77 |

**Table 2.4. List of alternative machine learning algorithms and their preformance measured by precision, recall and F1-score.**

**Semi-Supervised Learning**

The label spreading algorithm using both labeled and unlabeled data did not perform as well as an SVM trained using only the labeled data. Results are shown in Table 2.5. Although semi-supervised learning algorithms are resilient to noise, sparse, high-dimensional data sets may require significantly larger labeled and unlabeled training sets to sufficiently specify the structure of clusters within high-dimensional data sets.

| Algorithm | Precision | Recall | F1-score |
|---|---|---|---|
| SVM | 0.80 | 0.79 | 0.79 |
| Label Spreading | 0.77 | 0.69 | 0.73 |

**Table 2.5. Training classifiers with a combination of labeled and unlabeled data can sometimes improve precision, recall and F1-score. This experiment shows that the semi-supervised Label-Spreading algorithm does note perform better than SVM.**

## Discussion

SVMs achieve the best classification performance in these experiments. Applying alternative supervised and semi-supervised algorithms did not improve results, as measured by F-score. In addition, analysis of error curves indicates that training on additional data is not likely to improve performance. If additional data and changes to algorithms will not lead to improved performance, it indicates that an alternative representation is required to realize substantial improvement in biocuration classification tasks.

TF-IDF representations do not capture structural or semantic attributes of text. Misclassification

examples highlight the need for at least some ability to capture basic word semantics and syntactic relations. Examples of non-virulence factor sentences that were classified as virulence factor sentences by SVM algorithm using TF-IDF representation include:

"Collectively, these data suggest that EPEC 30-5-1(3) translocates reduced levels of EspB into the host cell." [Devinney, 2001]

"Data were log-transformed to correct for heterogeneity of the variances where necessary." [Kostakioti, 2004]

"Subsequently, the kanamycin resistance cassette from pVK4 was cloned into the PstI site of pMP3, and the resulting plasmid pMP4 was used to target a disruption in the cesF region of EHEC strain 85-170." [Viswanathan, 2004]

- Virulence factor sentences that were misclassified by TF-IDF as non-virulence factor sentences include:

"Here, it is reported that the pO157-encoded Type V-secreted serine protease EspP influences the intestinal colonization of calves. "[Dziva, 2007]

"Here, we report that intragastric inoculation of a Shiga toxin 2 (Stx2)-producing E. coli O157:H7 clinical isolate into infant rabbits led to severe diarrhea and intestinal inflammation but no signs of HUS. " [Ritchie,. 2003]

"The DsbLI system also comprises a functional redox pair" [Totsika, 2009]

One approach to improving performance is to explicitly model relations between variables or explicitly represent information not already in a representation scheme. For example, one could replace gene/protein names with term GENE_PROTEIN to allow TF-IDF to capture statistics on the concept of gene/protein rather than on individual gene/protein. Taxonomies, such as the Unified Medical Language System (UMLS) could be used to derive semantic similarity measures between terms. These features could be used to augment the features captured by TF-IDF. Based on our experience with the PATRIC project [Mao, 2015], we hypothesize that manual feature engineering is not a viable option give our goal to create a platform that will serve as a topic-independent support tool for biocuration.

The high quality curation of specialty gene sets is imperative for any biocuration effort that is directed toward linking gene names with specified functions; however, manual methods will not scale to the volume of literature that must be reviewed. Text classification can aid the process. An important task for curators and text miners is determining the point at which the marginal improvement of a classifier gained by training with more examples is outweighed by the cost of creating additional training data

Bag of word models such as TF-IDF are limited in their ability to capture characteristics of sentences needed to distinguish virulence factor assertion sentences from other sentences. Recent advances in distributed word representations (Wang 2014) and convolutional neural (Collobert and Weston 2008) have demonstrated impressive results. Future research will focus on creating dense word representations and deep learning neural networks to mitigate the weaknesses of the TF-IDF representation scheme

# References


Aziz, Ramy K., et al. "The RAST Server: rapid annotations using subsystems technology." *BMC genomics* 9.1 (2008): 75.

Collobert, Ronan, and Jason Weston. "A unified architecture for natural language processing: Deep neural networks with multitask learning."*Proceedings of the 25th international conference on Machine learning*. ACM, 2008.

Devinney, Rebekah, et al. "Tir tyrosine phosphorylation and pedestal formation are delayed in enteropathogenic Escherichia coli sepZ:: TnphoA mutant 30-5-1 (3)." *Infection and immunity* 69.1 (2001): 559-563.

Drage, Michael G., et al. "Mycobacterium tuberculosis lipoprotein LprG (Rv1411c) binds triacylated glycolipid agonists of Toll-like receptor 2." *Nature structural & molecular biology* 17.99 (2010): 1088-1095.

Dziva, Francis, et al. "EspP, a Type V-secreted serine protease of enterohaemorrhagic Escherichia coli O157: H7, influences intestinal colonization of calves and adherence to bovine primary intestinal epithelial cells." *FEMS microbiology letters* 271.2 (2007): 258-264

GriGriffiths, Alan, Lesley A. Robinson, and Peter Willett. "Hierarchic agglomerative clustering methods for automatic document classification."*Journal of Documentation* 40.3 (1984): 175-205.

HHan, Eui-Hong Sam, and George Karypis. *Centroid-based document classification: Analysis and experimental results*. Springer Berlin Heidelberg, 2000.

Hancox, Lisa S., Kuang-Sheng Yeh, and Steven Clegg. "Construction and characterization of type 1 non-fimbriate and non-adhesive mutants of Salmonella typhimurium." *FEMS Immunology & Medical Microbiology* 19.4 (1997): 289-296.

Haussler, David. "Quantifying inductive bias: AI learning algorithms and Valiant's learning framework." *Artificial intelligence* 36.2 (1988): 177-221.

Joachims, Thorsten. *Text categorization with support vector machines: Learning with many relevant features*. Springer Berlin Heidelberg, 1998.

Gillespie, Joseph J., et al. "PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species." *Infection and immunity* 79.11 (2011): 4286-4298.

Kostakioti, Maria, and Christos Stathopoulos. "Functional analysis of the Tsh autotransporter from an avian pathogenic Escherichia coli strain." *Infection and immunity* 72.10 (2004): 5548-5554.

Mao C, Abraham D, Wattam AR, et al. Curation, integration and visualization of bacterial virulence factors in PATRIC. *Bioinformatics*. 2015;31(2):252-258. doi:10.1093/bioinformatics/btu631.

Milohanic, Eliane, et al. "The autolysin Ami contributes to the adhesion of Listeria monocytogenes to eukaryotic cells via its cell wall anchor." *Molecular microbiology* 39.5 (2001): 1212-1224.

Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." *The Journal of Machine Learning Research* 12 (2011): 2825-2830.

Pyysalo, Sampo, et al. "Towards event extraction from full texts on infectious diseases." *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*. Association for Computational Linguistics, 2010.

Ritchie, Jennifer M., et al. "Critical roles for stx2, eae, and tir in enterohemorrhagic Escherichia coli-induced diarrhea and intestinal inflammation in infant rabbits." *Infection and immunity* 71.12 (2003): 7129-7139.

Robertson, Stephen. "Understanding inverse document frequency: on theoretical arguments for IDF." *Journal of documentation* 60.5 (2004): 503-520.

Totsika, Makrina, et al. "Characterization of two homologous disulfide bond systems involved in virulence factor biogenesis in uropathogenic Escherichia coli CFT073." *Journal of bacteriology* 191.12 (2009): 3901-3908.

Turney, Peter D. "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews." *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002.

Viswanathan, V. K., et al. "Comparative analysis of EspF from enteropathogenic and enterohemorrhagic Escherichia coli in alteration of epithelial barrier function." *Infection and immunity* 72.6 (2004): 3218-3227.

Wang, Huizhen. "Introduction to Word2vec and its application to find predominant word senses." (2014).

Willett, Peter. "Recent trends in hierarchic document clustering: a critical review." *Information Processing & Management* 24.5 (1988): 577-597.

Zhou, Dengyong, et al. "Learning with local and global consistency."*Advances in neural information processing systems* 16.16 (2004): 321-328.

# Chapter 3: Evaluation of Semantic Properties of Dense Word Representations Generated by Word2Vec Skip-Gram


## Abstract

Biocuration of genetic and proteomic data is increasingly important for infectious disease research. Biocuration is a largely manual process that depends on information retrieval and text mining techniques to effectively search and extract information from the biomedical literature. Statistical techniques and term frequency-inverse document frequency (TF-IDG) representations have been used as building blocks for automated tools supporting biocuration. Recent advances in dense word representations, also known as word embeddings, demonstrate the ability to capture a wide range of term features, including semantic and syntactic properties. [Mikolov, 2013] This study evaluates the ability of dense word representations to capture semantic properties as compared to ontologies and organism taxonomies. It also evaluates properties of word graphs based on word similarity. Results demonstrate that dense word representations complement ontologies and taxonomies by capturing some of the same information while incorporating information represented in the biomedical literature but not explicit in ontologies.


## Introduction

The corpus of biomedical literature, as indexed by PubMed in late 2015, includes over 23.3 million publications, of which, approximately 1.1 million full-text papers are freely available in PubMed Central. The size of the biomedical corpus offers opportunities to supplement the growing collections of structured data about genomics, transcriptomics, proteomics, and other areas of interest to human health. Biocurators annotate structured data sources, such as model organism databases and specialty databases, such a PATRIC [Wattam, 2013], UniProt [UniProt Consortium, 2008], and EupathDB [Aurrecoechea,2010]. Valuable information about the biological entities cataloged in these resources is available in the biomedical literature but is difficult to incorporate into structured databases for two reasons. First, the volume of literature makes it difficult to find specific details of interest without also retrieving irrelevant content. This is a common problem and not limited to biomedical research. The area of research known as information retrieval addresses this problem. [Frakes, 1992] The second difficulty is due to the fact that, once relevant content is found, it is difficult to identify and extract facts of interest within scientific publications. Natural language processing and text mining research attempts to address this problem.

A common approach to working with natural language text is to map the contents of documents to a

vector representation known as a term TF-IDF representation. [Salton, 1975]. This representation measures the frequency of each word in a document relative to the word's frequency in corpus, such as all Medline abstracts or PubMed Central open access papers. A TF-IDF vector has a size equal to the size of the corpus vocabulary. For any document, each word in the document is assigned a TF-IDF score; since most words in the vocabulary are not used in single document, the TF-IDF vector is sparse. An advantage of this representation is that it enables similarity measures between documents using the geometric distance between points in a vector space or, more commonly, using the cosine between two document vectors.

The TF-IDF representation works well in many applications but has limitations, especially when applied to curating genes of interest in bacterial infectious diseases. [Sullivan and Wattam, 2016; under review] Previous research has demonstrated that TF-IDF used with a variety of classification algorithms can classify sentences making assertions about virulence factors with reasonable precision and recall (F1-score approximately 0.85) [Sullivan and Wattam, 2016 under review] That research also showed that increasing the number of training examples and varying the types of classification algorithms used could not improve the precision and recall beyond an upper bound. Since increasing training set sizes and varying the classification algorithm did not improve performance, we hypothesize that a change in feature representation is required to achieve higher precision and recall. In particular, since TF-IDF representations do not capture semantic relations between words, we further hypothesize that a more semantic-rich representation can improve the quality of document classification.

This paper presents the results of evaluating an alternative representation scheme for words known as dense word embeddings, which have been shown to capture semantic relationships between words. [Le and Mikolov, 2014] The objective of the evaluation is to determine how dense word representations generated using the skip-gram algorithm [Mikolov *et. al.* 2013] capture: (1) the semantics of words when compared to biomedical ontologies, (2) the chemical properties of amino acids, and (3) taxonomic relations between bacteria.

## Methods

Distributed word representations are derived from an analysis of the distributional properties of words in large corpus. Properties of distributions, including syntactic and semantic properties, are captured in high dimension vectors (typically on the order of 50 to several hundred). The Skip-gram algorithm learns word representations by optimizing an objective function that predicts nearby words. The results described here use the Skip-gram algorithm implemented in Milokov's word2vec program. [Le and Mikolov, 2014]

This research uses a shallow neural network that implements the Skip-Gram algorithm to learn semantic vectors for words and paragraphs. Recent advances in machine learning have repeatedly demonstrated that deep, multi-layer networks can achieve high, and sometimes state of the art, levels of performance. [Le, 2013], [Glorot, 2011], [Collobert, 2008]. There are, however, drawbacks to using multi-layer networks and the decision to use a two layer network was based on a consideration of the benefits and costs of two layer vs multi-layer models.

The shallow Skip-Gram model has several advantages.

First, Skip-Gram, as implemented in the word2vec program [Mikolov, 2013a], has been widely successfully used in the text processing community for tasks including: sentiment analysis [Zhang, 2015], bilingual semantic representations [Wolf, 2014], and extracting quantitative variables from unstructured text [Amunategui, 2015]. The experiments described in this dissertation demonstrate high levels of performance in text classification tasks (F1-score > 0.9). In addition, the original Skip-Gram algorithm has been improved by introducing negative sampling, increasing the performance of the algorithms without increasing the time to train the network. [Mikolov, 2013b] The high levels of performance in a range of text analysis areas support the hypothesis that Skip-Gram is robust and applicable to a range of text analysis problems.

Deep neural networks have been used to train semantic vectors ([Collobert, 2004], [Turian, 20100], [Huang, 2013]) but they are computationally more expensive than the Skip-Gram algorithm. [Mikolov, 2013a]. Computational complexity is an important consideration given the size of trainings sets, which in the case of this study, included training the network on a corpus of over 2 billion words. There is the potential for a deep network to improve the performance of a classifier based on performance of other deep learning networks. However, one would also have to investigate if similar or better performance increases could not be achieved by increasing the size of the training corpus or number of training epochs while still training the shallow network in less time than required to train a deep network.

Deep learning networks do not always find optimal solutions and sometimes yield sub-optimal results. [Pandey, 2014]. Heuristics are used to tune a number of configuration parameters, such as the number of training epochs, learning rate, momentum, and batch size but some researchers have found that poor choices for these hyper-parameters can lead to decreased performance. [Bergstra, 2011]. Skip Gram requires hyper-parameter tuning as well but given the lower computational complexity, it is more efficient to apply grid search to find optimal hyper-parameter configurations.

Finally, it has been demonstrated that wide, shallow networks can perform better at some classification

tasks than deep networks. [Pandey, 2014] Specifically,

> Using a single RBM [Restricted Boltzman Machine] to learn a wide layer, we are able to obtain better results for many classification tasks than obtained by multi-layer neural network initialized using a deep belief network and fine-tuned using backpropagation. [Pandey, 2014]

Pandey's work does not compare Skip-Gram to a deep network but it does demonstrate that deep neural network are not inherently better at classification than shallow networks. More research is needed in this area to identify what kinds of classification problems are well suited to shallow but wide neural networks, such as used by Skip-Gram, and which are better suited for deep neural networks.

In this study, a collection of approximately 23 million biomedical abstracts from Medline and approximately 1.1 million full length open access papers from PubMed Central constitute the corpus for this study. The corpus includes 2.4 billion words. The text was preprocessed to remove punctuation and lowercase all words. Frequently occurring two and three word phrases, bigrams and trigrams, were concatenated to create a single word. Both the Word2Vec and bigram/trigram generation program were downloaded from https://code.google.com/p/word2vec/.

This research includes a comparative analysis of several methods for representing text and comparing semantic representations, including: using term frequency-inverse document frequency (TF-IDF) and semantic vectors for text representations, using machine learning algorithms, and ontology comparisons specifically developed for this research. The justification of each choice follows below.

TF-IDF representations was described by Salton in 1975 and has since been cited over 6,300 times according to Google Scholar. [Salton, 1975] The representation has been successfully used in text classification tasks, including short text classification [Robertson, 2004], news filtering [Lang, 2002], and spam filtering [Drucker, 1999]. These tasks are sufficiently similar to biomedical classification to warrant the use of TF-IDF. In addition, the combination TF-IDF combined with support vector machines has been studied as well. [Joachims, 1998] [Leopold, 2002].

Semantic vectors can potentially improve on TF-IDF as a feature representation scheme. Specifically, semantic vectors capture semantic relations between words. It is hypothesized that using a representation that captures semantic relations will improve the quality of classification. Results of experiments described in this dissertation support that hypothesis. Semantic vectors have used for

sentiment analysis [dos Santos, 2014], part of speech tagging [Tsuboi, 2014], and word sense disambiguation [Chen, 2014]. This breadth of task indicates that semantic vectors are sufficiently robust to support a range of text analysis tasks.

Several machine learning algorithms are used in this study. Algorithms were chosen to represent several widely used classes of learners. Linear learners include support vector machines (SVMs), Percptron, and Linear Regression. SVMs are wide margin classifiers widely used in text classification tasks. [Hearst, 1998]. Classification tree methods used include decision trees and random forests. Naive Bayes was the one probabilistic learning algorith used. The ensemble algorithm, Adaboost, was used as well. These algorithms were selected to help identify the optimal combination of representation (TF-IDF or semantic vector) and classification algorithm. It is not clear, *a priori*, that any one type of classification algorithm would work best with semantic vectors. The fact that SVMs with a linear kernel perform well on TF-IDF classifications support the hypothesis that text representations based on words are linearly separable.

Ontology comparisons are used to evaluate the ability of semantic vectors to capture semantic relations. Previous studies have depended on common knowledge analogies, such as countries and their capitals, to evaluate the ability of semantic vectors to capture useful semantic relations. One of the objectives of this research is to evaluate the quality of semantic relations captured by semantic vectors trained on a large, biomedical corpus, with particular emphasis on infectious disease topics. Biomedical ontologies such as MESH, represent important biomedical concepts and relations as defined by expert curators. [Lipscomb, 2000] Designers of MESH note that the resources complements automatic text analysis methods by providing precision not typically found in automatic information retrieval methods:

> Even with advances in automation and resulting changes in the capabilities of indexing and searching, an important role remains for MeSH in organizing information in a way that provides precision and power in retrieval. [Lipscomb, 2000]

For this study, MESH is considered the "gold standard" of semantic representations. MeSH has existed since 1960 and has been described as "one of the most sophisticated thesauri in existence today." [Nelson, 2001] MeSH terms represent single concepts within the biomedical field. MeSH terms comprise 15 hierarchies, know as the MeSH Tree Structure. [Lowe, 1994] Each hierarchy constitutes a a structured set of increasingly specific terms. In addition to biomedical terms, there are additional terms to support search, such as publication types. These additional terms are not relevant to the research at hand.

MESH is a thesaurus that encompasses features of an ontology, including the ability to reason about the ontology as a graph. Nodes of the graph represent terms and measures such as the number of edges between nodes can be uses as a measure of semantic similarity. Semantic vectors, however, exist in a linear space and similarity is typically measured using cosine or Euclidean distance. Obviously, direct comparisons between the two representations are not possible. Instead, custom methods were developed to compare similarity measures in MESH and semantic vectors.

Computing semantic similarity between semantic vectors is straight forward: the cosine of two vectors is used as a measure of similarity. Computing the semantic similarity between terms in an ontology such as the MeSH hierarchies, requires a different metric. The evaluations described in this dissertation used the Sanchez information content method. [Sanchez, 2011]. Information content (IC) of a term is defined as the amount of information it conveys in a context; it has been widely used as a measure of semantic similarity. Semantic similarity, in turn, is understood as a degree of taxonomical resemblance. [Goldstone, 1994] The Sanchez method for computing IC is applied to MeSH because it is both scalable and produces high quality results, as reported in [Sanchez, 2011]. The Semantic Similarity Toolkit implementation of the Sanchez measure is used in this research. [Harispe, 2014].

# Results

### Amino Acids Evaluation

The first evaluation assess the ability of dense word representations to capture similarity between amino acids. Two similarity measures are used. An ontology-based semantic similarity measure compares similarity of amino acids using MESH with similarity scores derived using word2vec. The Semantic Measures Toolkit [Harispe, et. al. 2014] is used to measure similarity of amino acids relative to the MESH ontology. The pairwise measure proposed by Lin [Lin, 1998] using information content measure proposed by Sanchez [Batet, et. al. 2010] are used for all ontology evaluations described here. The second measure counts the number of common characteristics between similar amino acids; the four common characteristics used are: pH, hydrophobicity, structure, and essentiality.

Within the top 100 most similar pairs of amino acids ranked by MESH similarity, the pairs average 2.02 common features. The word2vec ranking of top 100 pairs average 1.88 common features. By this measure, the dense word representations are able to capture approximately 93% of the semantic similarity captured in a manually curated ontology.

**Infectious Disease Evaluation**

Infectious diseases may be categorized in a variety of ways, including the taxonomic relation of the infectious agents, virulence mechanisms, or host response to infections. The following infectious agents are compared using MESH similarity and word2vec similarity:

*Bacillus anthracis*
*Brucella*
*Burkholderia mallei*
*Burkholderia mallei*
*Burkholderia pseudomallei*
*Campylobacter jejuni*
*Chlamydia psittaci*
*Clostridium botulinum*
*Clostridium perfringens*
*Coxiella burnetii*
*Entamoeba histolytica*
*Francisella tularensis*
*Giardia lamblia*
*Listeria monocytogenes*
*Rickettsia prowazekii*
*Rickettsia prowazekii*
*Salmonella*
*Shigella*
*Staphylococcus enterotoxin*
*Toxoplasma gondii*
*Yersinia pestis*

Table 3.3 lists the MESH and word2vec similarity scores for each pair of infectious disease pathogens. For each pathogen, the top five most similar other pathogens are compared. A mean of 1.89 pathogens are included in the top five most similar pathogens in both MESH and word2vec rankings. Word2Vec representations capture 37.8% of the similar rankings of MESH. Although this is a low percentage, especially when compared to the 93% similarity between word2vec and MESH amino acid similarity, it can be explained by the multiple dimensions along which similarity can be measured.

For example, *Bacillus anthracis* is ranked similar to *Listeria monocytogenes* and *Clostridium perfringens* in both MESH and word2vec rankings. *Bacillus anthracis* is the etiologic agent of anthrax

and is a Gram-positive, rod shaped bacterium that can grow under both areobic and anerobic conditions [Inglesby, 1999]. *Listeria monocytogenes,* a Gram-positive, facilitative anaerobe, is the cause of listeriosis, a food-borne disease that can lead, in some cases, to brain infection and. [Ramaswamy, 2007] *Bacillus anthracis* has also been associated with gastrointestinal disease. [Beatty, 2003] *Clostridum perfingens*, formally known as *Bacillus welchii,*is also Gram-positive, rod shaped, anerobic bacterium that causes food-borne disease. [Immerseel, 2004] *Bacillus anthracis , Listeria monocytogenes* and *Clostridium perfingens* are associated with disease in livestock as well. [D'Amelio,2015] [Niilo, 1980] [Immerseel, 2004] Ramaswamy, 2007] *Bacillus anthracis, Listeria monocytogenes* and *Clostridium perfingens* are all members of the phylum Firmicutes.

By the word2vec similarity, *Bacillus anthracis* is most similar to *Yersina pestis*, which is ranked 7th by MESH similarity. *Yersinia pestis* is a Gram-negative, rod shaped facultative anerobic bacterium. It is the etiologic agent of the Bubonic plague and infects both humans and animals. [Pechos, 2015] *Bacillus anthrasis*, as noted above, is a member of the phylum Firmicutes while *Yersina pestis* is a member of the phylum Proteobacteria.

One explanation for the difference in rankings is that the MESH ontology weighs taxonomic lineage more heavily than word2vec. Word2vec calculates word vectors based on the context of words. It is reasonable to expect the biomedical literature to focus less on describing commonly understood taxonomic relations than new research discoveries.

Ontologies are designed to capture important relations between entities in a highly structured manner. GO, for example, organizes relations along the dimensions of biological processes, cellular components, and molecular function. Organism taxonomies organize relations along lines of evolutionary decent (issues of horizontal gene transfer notwithstanding). The scientific literature about infectious disease is diverse. While it certainly includes descriptions of relations found in ontologies, it also incorporates other information as well. The textual context in which the names of pathogens appear varies widely and can include discussions of taxonomy, molecular function of virulence factors, antibiotic resistance, epidemiology, treatments, and public health. Amino acids, in contrast, are considerably simpler in structure and function and thus the textual context in which the names of amino acids appear is more homogeneous.

**Taxonomic Similarity**

A third measure of the semantic features of dense word embeddings is based on the ability to determine analogies across taxonomic categories. Dense word embeddings have demonstrated the ability to capture analogical relationships.

[Le and Mikolov, 2014] demonstrate that dense word vectors trained on news stories can be used to determine the missing term in a four term analogy, such as Paris is to France as Rome is to X, where X is determined to be Italy. Similar analogical reasoning was used with famous individuals and their professions, companies and their products, and countries and popular foods. [Mikolov, 2013]

For this study, three types of taxonomic analogies are used: genus to phylum, genus to class, and genus to family. The genus to phylum test poses analogies such as:

**staphylococcus:firmicutes as brucella:** X

where X should evaluate to proteobacteria. Similarly, genus to class and genus to family examples are:

**staphylococcus:bacilli as burcella: alphaproteobacteria**, and

**staphylococcus:staphylococcaceae as brucella:brucellaceae**,

respectively.

A total of 447 analogical tests were evaluated; 137 instances included the correct taxonomic analog in the top ten most analogous terms for an accuracy rate of 30.6%. The phylum to genus analogies had the most correct analogs with 59 correct analogs in 149 tests, for an accuracy rate of 39.6%. The order to genus tests yield correct results in 33 instances, for an accuracy rate of 22.1%. The family to genus tests yield correct results in 32 instances, for an accuracy rate of 21.4%.

Semantic accuracy of dense word vectors generated using the Skip-gram algorithm range from 45.6% to 56.7% when trained on news stories. [Mikolov 2013]. We hypothesize that the analogy evaluation results are lower in this case because the corpus does not contain sufficient instances that include descriptions of taxonomic relations. This is not surprising given that taxonomic relations are well defined and foundational knowledge in life sciences. This kind of foundational knowledge is not often described in papers presenting the results of research. It is, however, the kind of information captured in biomedical ontologies. This is just one example demonstrating that biomedical ontologies and dense word representations derived from the biomedical literature complement, but do not replace, each other.

**Evaluation of Similarity Network**

In addition to evaluating the semantic properties of dense word representations at the level of

individual terms and their relation to each other, this study evaluated meso-scale properties of a similarity graph of a subset of terms.

A graph representing the relation of similar terms was created using the vocabulary of infectious disease research. The vocabulary is defined as the set of terms used in open access journals articles drawn from journals publishing papers on infectious diseases and related research. 14,136 papers from 24 journals constitute the corpus for the infectious disease terminology evaluation. The vocabulary includes 601,491 distinct terms. Commonly occurring words, such as "the", "of", and "in" are removed prior to analysis.

A graph is constructed with a node representing a term in the infectious disease vocabulary. A link is added between nodes when their similarity score is greater than 0.7. The resulting graph is not fully connected.

Several examples of clusters of similar terms further illustrate the ability of dense word representations to capture semantic similarity. See Appendix A for example clusters related to:

Pathogens
Agricultural plants
Carbohydrates
Lyme disease
Viruses
Adhesion related virulence factors
Genetic mutations
Antiseptics
Taxonomy
Immunology
Synonyms for "show", "demonstrate", etc.

## Discussion

Comparing the semantic properties of the MESH ontology with those of dense word representations generated by word2vec highlight the complementary nature of the two types of knowledge structures. Manually curated ontologies are well suited for capturing and expressing fundamental and widely used relations between biological entities. The biomedical research literature is the primary mechanism for sharing discoveries and unique analysis results.

The semantics of relatively simple entities, such as amino acids, are well represented in both MESH and in dense word representations. Since there are relatively few dimensions along with amino acids vary, both manually curated ontologies and automatically derived word representations capture a wide range of amino acid semantic properties. More complex entities, such as infectious bacteria vary by many dimensions. Ontologies and taxonomies capture vital information about these entities but the breadth of information is limited. The biomedical literature describes a wide array of characteristics of pathogenic bacteria ranging from low level genomic detail to population scale epidemiological properties.

Ontologies and dense word representations may be combined in ensemble techniques to improve the efficiency of biocuration. MESH is used to specify metadata tags of papers indexed in PubMed. These tags can provide coarse-grained classification of documents while additional information derived from dense word representations can provide finer grained information that is more semantically informed than possible with TF-IDF representations.

Additional research is needed to further elucidate the characteristics of dense word representations and methods for tuning results for application specific requirements. For example, the addition of structured text with detailed lineage information derived from taxonomies could be used to supplement the biomedical literature to improve the quality of taxonomy analogies. A detailed analysis of the textual context of key terms may provide insights into how manually crafted, supplement texts can help tune dense word vector representations.

## References


Amunategui, Manuel, Tristan Markwell, and Yelena Rozenfeld. "Prediction Using Note Text: Synthetic Feature Creation with word2vec." *arXiv preprint arXiv:1503.05123* (2015).

Aurrecoechea, Cristina, et al. "EuPathDB: a portal to eukaryotic pathogen databases." *Nucleic acids research* 38.suppl 1 (2010): D415-D419.

Batet M, Sanchez D, Valls A: An ontology-based measure to compute semantic similarity in biomedicine. Journal of biomedical informatics 2010, 44:118-125.

Beatty, Mark E., et al. "Gastrointestinal anthrax: review of the literature."*Archives of internal medicine* 163.20 (2003): 2527

Bergstra, James S., et al. "Algorithms for hyper-parameter optimization."*Advances in Neural Information Processing Systems*. 2011.

Chen, Xinxiong, Zhiyuan Liu, and Maosong Sun. "A Unified Model for Word Sense Representation and Disambiguation." *EMNLP*. 2014.

Coecke, Bob, Mehrnoosh Sadrzadeh, and Stephen Clark. "Mathematical foundations for a compositional distributional model of meaning." *arXiv preprint arXiv:1003.4394* (2010).

Collobert, Ronan, and Jason Weston. "A unified architecture for natural language processing: Deep neural networks with multitask learning."*Proceedings of the 25th international conference on Machine learning*. ACM, 2008.

D'Amelio, Enrico, et al. "Historical evolution of human anthrax from occupational disease to potentially global threat as bioweapon." *Environment international* 85 (2015): 133-146.

dos Santos, Cícero Nogueira, and Maira Gatti. "Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts." *COLING*. 2014

Drucker, Harris, Donghui Wu, and Vladimir N. Vapnik. "Support vector machines for spam categorization." *Neural Networks, IEEE Transactions on*10.5 (1999): 1048-1054.

Freeman, Linton C. "A set of measures of centrality based on betweenness."Sociometry (1977): 35-41.

Frakes, William B., and Ricardo Baeza-Yates. "Information retrieval: data structures and algorithms." (1992).

Glorot, Xavier, Antoine Bordes, and Yoshua Bengio. "Domain adaptation for large-scale sentiment classification: A deep learning approach." *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. 2011.

Goldstone, Robert L. "Similarity, interactive activation, and mapping." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20.1 (1994): 3.

Harispe, Sébastien, et al. "The semantic measures library and toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies." Bioinformatics 30.5 (2014): 740-742.

Harispe, Sébastien, et al. "A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain." Journal of biomedical informatics 48 (2014): 38-53.

Harispe, Sébastien, et al. "Semantic similarity from natural language and ontology analysis." Synthesis Lectures on Human Language Technologies8.1 (2015): 1-254.

Harispe, Sébastien, et al. "Semantic similarity from natural language and ontology analysis." Synthesis Lectures on Human Language Technologies8.1 (2015): 1-254.

Hearst, Marti A., et al. "Support vector machines." *Intelligent Systems and their Applications, IEEE* 13.4 (1998): 18-28.

Huang, Eric H., et al. "Improving word representations via global context and multiple word prototypes." *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 2012.

Immerseel, Filip Van, et al. "Clostridium perfringens in poultry: an emerging threat for animal and public health." *Avian pathology* 33.6 (2004): 537-549.

Inglesby, Thomas V., et al. "Anthrax as a biological weapon: medical and public health management." *Jama* 281.18 (1999): 1735-1745.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features, Springer.

Kiela, Douwe, and Stephen Clark. "A systematic study of semantic vector space model parameters." Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC) at EACL. 2014.

Lang, Ken. "Newsweeder: Learning to filter netnews." *Proceedings of the 12th international conference on machine learning*. 1995.

Le, Quoc V. "Building high-level features using large scale unsupervised learning." *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013.

Le, Quoc V., and Tomas Mikolov. "Distributed representations of sentences and documents." arXiv preprint arXiv:1405.4053 (2014).

Lee, Wei-Nchih, et al. "Comparison of ontology-based semantic-similarity measures." AMIA. 2008.

Leopold, Edda, and Jörg Kindermann. "Text categorization with support vector machines. How to represent texts in input space?." *Machine Learning*46.1-3 (2002): 423-444.

Lin D: An Information-Theoretic Definition of Similarity. In 15th International Conference of Machine Learning. Madison,WI: 1998:296-304.

Lipscomb, Carolyn E. "Medical subject headings (MeSH)." *Bulletin of the Medical Library Association* 88.3 (2000): 265.

Lowe, Henry J., and G. Octo Barnett. "Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches." *Jama*271.14 (1994): 1103-1108.

Maas, Andrew L., et al. "Learning word vectors for sentiment analysis."Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011.

Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." *arXiv preprint*

*arXiv:1301.3781* (2013a).

Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." *Advances in neural information processing systems*. 2013b.

Mnih, Andriy, and Geoffrey E. Hinton. "A scalable hierarchical distributed language model." *Advances in neural information processing systems*. 2009.

Nelson, Stuart J., W. Douglas Johnston, and Betsy L. Humphreys. "Relationships in medical subject headings (MeSH)." *Relationships in the Organization of Knowledge*. Springer Netherlands, 2001. 171-184.

Niilo, L. "Clostridium perfringens in animal disease: a review of current knowledge." *The Canadian Veterinary Journal* 21.5 (1980): 141.

Pandey, Gaurav, and Ambedkar Dukkipati. "To go deep or wide in learning?."*arXiv preprint arXiv:1402.5634* (2014).

Pechous, Roger D., et al. "Pneumonic Plague: The Darker Side of Yersinia pestis." *Trends in microbiology* (2015).

Ramaswamy, Vidhya, et al. "Listeria-review of epidemiology and pathogenesis." *Journal of Microbiology Immunology and Infection* 40.1 (2007): 4.

Robertson, S. (2004). "Understanding inverse document frequency: on theoretical arguments for IDF." Journal of Documentation **60**(5): 503-520.

Salton, Gerard, Anita Wong, and Chung-Shu Yang. "A vector space model for automatic indexing." Communications of the ACM 18.11 (1975): 613-620.

Sánchez, David, Montserrat Batet, and David Isern. "Ontology-based information content computation." *Knowledge-Based Systems* 24.2 (2011): 297-303.

Socher, Richard, et al. "Semantic compositionality through recursive matrix-vector spaces." Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics, 2012.

Tsuboi, Yuta. "Neural Networks Leverage Corpus-wide Information for Part-of-speech Tagging." *EMNLP*. 2014.

Turian, Joseph, Lev Ratinov, and Yoshua Bengio. "Word representations: a simple and general method for semi-supervised learning." *Proceedings of the 48th annual meeting of the association for computational linguistics*. Association for Computational Linguistics, 2010.

UniProt Consortium. "The universal protein resource (UniProt)." *Nucleic acids research* 36.suppl 1 (2008): D190-D195.

Wallach, Hanna M., et al. "Evaluation methods for topic models."Proceedings of the 26th Annual International Conference on Machine Learning. ACM, 2009.

Wattam, A.R., et al., *PATRIC, the bacterial bioinformatics database and analysis resource*. Nucleic acids research, 2013: p. gkt1099.

Wolf, Lior, et al. "Joint word2vec networks for bilingual semantic representations." *International Journal of Computational Linguistics and Applications* 5.1 (2014): 27-44

Yetisgen-Yildiz, Meliha, and Wanda Pratt. "A new evaluation methodology for literature-based discovery systems." Journal of biomedical informatics 42.4 (2009): 633-643.

Zager, Laura A., and George C. Verghese. "Graph similarity scoring and matching." Applied mathematics letters 21.1 (2008): 86-94.

Zhang, Dongwen, et al. "Chinese comments sentiment classification based on word2vec and SVM perf." *Expert Systems with Applications* 42.4 (2015): 1857-1863.

Zhang, Yuanzhe, et al. "Ontology Matching with Word Embeddings." Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. Springer International Publishing, 2014. 34-45

# Chapter 4: Evaluation of Semantic Paragraph Vectors For Supervised Learning of Classifiers for Biocuration Tasks


## Abstract

Biocuration of genetic and proteomic data is increasingly important for infectious disease research. Biocuration is a largely manual process that depends on information retrieval and text mining techniques to effectively search and extract information from the biomedical literature. Statistical techniques and term frequency-inverse document frequency (TF-IDG) representations have been used as building blocks for automated tools supporting biocuration. Recent advances in dense word representations, also known as word embeddings, demonstrate the ability to capture a wide range of term features, including semantic and syntactic properties. [Mikolov, 2013] This study evaluates the ability of dense word representations to capture semantic properties as compared to ontologies and organism taxonomies. It also evaluates properties of word graphs based on word similarity. Results demonstrate that dense word representations complement ontologies and taxonomies by capturing some of the same information while incorporating information represented in the biomedical literature but not explicit in ontologies.


## Introduction

Bag of words [Harris, 1954] and term frequency-inverse document frequency (TF-IDF) do not account for information latent in word order and do not capture semantic relationships between words. The arbitrary ordering of terms in a bag of words model means the words 'virulence', 'pathogenic' and 'glucose' could all be equidistant in a similarity measure. Le and Mikolov [Le and Mikolov, 2014] propose and unsupervised learning algorithm that maps texts of arbitrary length to a dense vector representation. The authors call the representation a paragraph vector and train it to predict words in a text, such as a sentence, paragraph or document.

Dense vector representations have been successfully used as representations that capture semantic properties of words. [Mikolov, 2013]. The success with word level semantics have led to studies to extend the application of dense vector representations to longer texts. [Mitchell and Lapta, 2010] use additive and multiplicative functions of vectors to combine semantic vectors. [Zonatto, et. al. 2010] estimate additive compositional semantics using vector sums using both positive and negative examples. Socher [Socher, 2011] used a combination of word vectors and syntax tree structures to generate sentence-level semantic vectors; this approach, however is restricted to sentences because of the need for a parse tree. Word vectors are often trained by neural networks using stochastic gradient decent and back-propagation. [Le and Mikolov, 2014]. An advantage of the training method developed

by Le and Mikolov is that at the end of training, words with similar meanings are located close to each other in the semantic vector space. This allows for determining the similarity of words using distance measures in a vector space. Although the Euclidean distance between points in a vector space is an obvious option, the cosine between vectors is often used to measure similarity in information retrieval applications.

Paragraph vectors are extensions of word vectors. Word vectors are dense vector representation of words. The values of vectors are learned using an iterative training method that optimizes the prediction of words in context. For example, the phrase "*Salmonella typhimurium* causes enteric ..." has a high probability of being followed by the word "disease." The vectors associated with each of the four words in the example phrase are concatenated or added and applied to a classifier to predict the fifth word "disease."

The Paragraph Vector model uses two types of vectors: word and paragraph. Both are dense vectors that start with arbitrary initial values and are trained to maximize the likelihood of predicting next words in a sequence. Unlike dense vector models that train only word vectors, the Paragraph Vector model trains paragraph and word vectors. During the prediction operation, the paragraph vector and word vectors are used to predict the next word in the text stream.

The focus of this study is to assess how well paragraph semantic vectors can be applied to biocuration tasks.

## Methods

This study consists of two experiments. The first experiment builds and evaluates a set of classifiers trained to identify paragraphs describing the virulence of a pathogenic bacteria. The second experiment evaluate the distance between paragraph semantic vectors as a method for retrieving similar paragraphs from a large corpus of paragraphs.

The results presented here are influenced by a combination of the data used to train classifiers, the representation of that data, and the machine learning algorithms used to create classifiers.

The data set used to train the classification algorithm constrain the vocabulary that can be learned by semantic vector training algorithms. Words that do not appear in the training corpus will not have semantic vectors generated. This effectively limits the usefulness of the semantic vectors to the domain of the training corpus. In this research, a combination of Medline abstracts and full text PubMed Central papers are used to train semantic word vectors. The corpus included over 2 billion words in context. The corpus includes commonly used English words as well as biomedical and biological terms

that are not likely to be found in large number in non-biomedical corpora, such as Wikipedia or Google News. Biocuration in the area of infectious disease research was the initial motivation for this research so the training corpus should be sufficient to create semantic vectors that can be used in application designed to support biocuration and related biomedical tasks.

Representations are used to make explicit features of texts that can be used by machine learning algorithms to create classifiers. TF-IDF captures features of words relative to their distribution in a document and across an entire corpus. Semantic vectors capture semantic and syntactic features of words based on the context in which those words are used. The latter representation present the opportunity to use semantic features which are not captured by TF-IDF. Arguably, one could supplement TF-IDF representations with features derived from biomedical ontologies but such ontologies are sparse and difficult to build. [REF] In some cases, representations cannot manifest a sufficient set of features to allow machine learning algorithms to discover a model that effectively classifies input text into a set of categories. As results from the TF-IDF experiment in this study show, additional training instances cannot overcome insufficient feature representation. Some practitioners resort to the practice of feature engineering, which is the process of creating derived or other explicit features on a case-by-case basis. This can be an effective way to develop a high quality classifier but it does not scale as a general solution.

Machine learning algorithms also impact the outcome of the experiments described here. In general, supervised machine learning programs attempt to maximize some objective function, such as maximizing the distance from margins in a data set, or creating a set of decision trees that together minimize the error rate on predictions over a training set. Machine learning algorithms sometimes have parameters that are not learn; these are referred to as hyper-parameters. Hyper-parameters include learning rates, training set sizes, number of training epochs, and momentum (a parameter that influences the ability of algorithms to avoid local minima).

**Experiment 1: Pathogenic Bacteria Descriptions**

In the first experiment, paragraphs are drawn from two sources: the Pubmed Central Open Access Data Subset (http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/), and PathInfo documents [He, 2005]. The Pubmed Central Open Access Data Set consists of full length journal articles that are made available

under a Creative Commons (https://creativecommons.org/about/license/) or similar license. Such licenses allow for more liberal reuse, such for use in text mining research. PathInfo documents are XML documents that have been manually curated and constructed to include information about pathogenic organisms, their interactions with hosts, epidemiology information, lab work results and other text-based data. For the purposes of this study, 117 paragraphs describing *Clostridium botulinum* and *Mycobacterium tuberculosis* are drawn from PathInfo documents. (See Supplemental Materials for contents of the two PathInfo documents). An equal number of randomly selected paragraphs are drawn from the PubMed Open Access subset as negative examples for the supervised learning task.

All text was preprocessed to convert case to lower case, remove punctuation, and replace number with the special symbol <NUM>. An example of a preprocessed paragraph on *C. botulinum* is [adapted from Shapario, 1998]:

> foodborne botulism is caused by ingestion of preformed toxin produced in food by c botulinum the most frequent source is home-canned foods in which spores that survive an inadequate cooking and canning process germinate reproduce and produce toxin spores of c botulinum are ubiquitous in the environment but growth and elaboration of toxin occur only under particular conditions that include an anaerobic low-salt low-acid environment he canning and fermentation of foods are particularly conducive to creating anaerobic conditions that allow c botulinum spores to germinate foodborne botulism while rare remains a public health emergency because of its severity and epidemic potential home-canned foods and alaska native foods remain the leading causes in the united states and restaurant-associated outbreaks continue to account for a disproportionate number of illnesses

One of the corresponding paragraph semantic vectors for this paragraphs is:

```
[ 0.04036991, -0.0213439 ,  0.03758252, -0.01621538, -0.01302081,
| -0.03816462, -0.01245213,  0.01613608, -0.02411742, -0.00271982,
  -0.00928097,  0.0059393 ,  0.00894287,  0.02785545,  0.01564969,
   0.0187883 , -0.03640994, -0.00724633, -0.03164659, -0.00412445,
  -0.04230138,  0.05019621, -0.03769375, -0.01958211,  0.04533418,
  -0.0026016 ,  0.01310881,  0.03167495,  0.04806663, -0.0350046 ,
   0.00675273, -0.0193779 ,  0.0200923 , -0.00411423,  0.02115137,
  -0.01893738, -0.03826446, -0.02828489,  0.0192324 ,  0.0396947 ,
   0.05046411, -0.08964178, -0.01509753, -0.06072177, -0.01098882,
  -0.05530918, -0.03268972, -0.05439398,  0.00715863, -0.00965075,
   0.0102442 ,  0.01930598, -0.05859354, -0.02759661,  0.01052014,
  -0.03944234, -0.0955945 , -0.0538061 ,  0.03224102,  0.0617632 ,
   0.01899998, -0.03722899,  0.01426818,  0.01609885, -0.01506272,
```

-0.02877383, -0.03547035, 0.0195378 , -0.05715351, -0.00824382,
0.00822788, -0.06070956, 0.010936 , -0.02600233, 0.06307442,
-0.00132673, 0.03335633, -0.00082981, 0.0159568 , -0.00101377,
-0.02791164, -0.05866118, 0.06919421, -0.03699533, 0.0023931 ,
-0.04201317, 0.0163089 , 0.01964355, 0.06147308, -0.03381988,
-0.06411717, 0.03068291, -0.02932315, -0.00809922, -0.06668216,
-0.02342989, -0.01985727, 0.04836439, 0.01894626, 0.04969702]

Paragraph vectors were generated using the doc2vec functions of the Gensim Python package. [Řehůřek and Sojka, 2010]. A number of parameters are required when using the doc2vec package; the most important for performance are the window size and the number of epochs. The window parameter specifies the maximum distance between the predicted word and the farthest word of context text used in training. The number of epochs specifies the number of times the vectors are trained using the full corpus. Setting the epoch to one would limit training to a single pass through the training corpus. This study examines windows of size 10 and 25, and epochs of 20 and 50. All paragraph semantic vectors are length 100. A label is assigned to each paragraph vector indicating whether the vector is a positive or negative example. The labels and vectors are then used to train classification algorithms.

9 supervised machine learning algorithms were applied to the labeled using 5-fold validation, and are defined in Table XX below. All algorithms are implemented in the Scikit-Learn machine learning package. [Pedregosa, 2011]

| **Algorithm** | **Description** |
|---|---|
| Decision Tree | A non-parametric machine learning algorithm that uses a series of decisions about attribute values to classify an entity. [Quinlan, 1986] |
| Adaboost | A meta-algorithm implementing the Stagewise Additive Modeling using a Multi-class Exponential loss function [Zhu, 2009] |
| Random Forest | An ensemble of random decision trees. [Breiman, 2001] |
| Extra Trees | Variation on decision tree algorithm. [Pedregosa, 2011 and Scikit-Learn, 2016a] |
| SVC | Support vector classification based on libsvm. [Chang, 2011] |

| Algorithm | Description |
|---|---|
| Linear SVC | Support vector classification based on liblinear. [Fan, 2008] |
| Perceptron | An early form of neural network that learns binary classification of linearly separable data sets. [Rosenblatt, 1958] |
| Linear Regression | Ordinary least squares linear regression. [Scikit-Learn, 2016b] |
| Naive Bayes | Application of Bayes theorem under the assumption of independence between all pairs of features. [Zhang, 2004] |

**Table 4.1 A list of 9 supervised machine learning algorithms used to train text classifiers using paragraph vectors.**

In addition to measuring results using individual algorithms, an ensemble of four algorithms (Logistic Regression LinearSVC, Naive Bayes, Perceptron) was used. Each algorithm in the ensemble was equally weighted in voting. [Dietterich,2000]

All algorithms and the ensemble voting algorithm are evaluated based on accuracy and F1-score. The F1-score is the harmonic mean of precision and recall.

**Experiment 2: Similarity by Distance Evaluation**

A second experiment attempted to assess the utility of distance measures as means using example paragraphs to find other similar paragraphs. The experiment consisted of building paragraph semantic vectors for 100,000 randomly selected paragraphs from the PubMed Central Open Access Subset and comparing a subset of the 117 paragraph semantic vectors generated from the PathInfo documents on *Clostridium botulinum* and *Mycobacterium tuberculosis*. The experiment entails generating the top ten most similar paragraphs for a set of *Clostridium botulinum* and *Mycobacterium tuberculosis* paragraphs. The most similar paragraphs are those closest in distance to the paragraph semantic vector of the *Clostridium botulinum* and *Mycobacterium tuberculosis* paragraphs. For each of the 10 closest paragraphs to the example *Clostridium botulinum* or *Mycobacterium tuberculosis* paragraph, a human expert would categorize each of the 10 paragraphs selected by distance.

# Results

**Experiment 1: Classification of Paragraphs**

The results of Experiment 1 are shown in Tables 4.2 and 4.3. Each table shows three sets of results, each with different parameter settings. The experiment was conducted three times with different window size and epoch sizes. The window size is the number of words that constitute the context around the target word in the Skip-gram algorithm. The epoch size is the number of iterations over the training set used to train the classifier.

| Individual Algorithms | | Window Size:10 Epoch size: 20 | | Window Size:25 Epoch size: 20 | | Window Size:25 Epoch size: 50 | |
|---|---|---|---|---|---|---|---|
| | | F1-score | Accuracy | F1-score | Accuracy | F1-score | Accuracy |
| | Decision Tree | 0.82 | 0.82 | 0.67 | 0.67 | 0.55 | 0.56 |
| | AdaBoost | 0.86 | 0.86 | 0.87 | 0.87 | 0.57 | 0.57 |
| | Random Forest | 0.80 | 0.80 | 0.80 | 0.80 | 0.51 | 0.52 |
| | Extra Trees | 0.81 | 0.81 | 0.78 | 0.79 | 0.52 | 0.53 |
| | SVC | 0.81 | 0.81 | 0.78 | 0.79 | 0.52 | 0.53 |
| | Linear SVC | 0.91 | 0.91 | 0.92 | 0.92 | 0.75 | 0.75 |
| | Perceptron | 0.83 | 0.84 | 0.82 | 0.82 | 0.61 | 0.61 |
| | Logistic Regression | 0.93 | 0.93 | 0.92 | 0.92 | 0.65 | 0.65 |
| | Naive Bayes | 0.91 | 0.91 | 0.90 | 0.91 | 0.65 | 0.50 |

**Table 4.2. Results of paragraph classification task using paragraph vectors and several supervise machine learning algorithms.**

| Ensemble (Equal Voting) | | Window Size:10 Epoch size: 20 | | Window Size:25 Epoch size: 20 | | Window Size:25 Epoch size: 20 | |
|---|---|---|---|---|---|---|---|
| | Individual Algorithms | F1-score | Accuracy | F1-score | Accuracy | F1-score | Accuracy |
| | Logistic Regression | 0.93 | 0.93 | 0.92 | 0.92 | 0.65 | 0.65 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Linear SVC | 0.91 | 0.91 | 0.90 | 0.91 | 0.74 | 0.74 |
| | Naive Bayes | 0.91 | 0.91 | 0.90 | 0.91 | 0.48 | 0.50 |
| | Perceptron | 0.88 | 0.99 | 0.82 | 0.82 | 0.60 | 0.61 |
| **Ensemble** | | **0.92** | **0.92** | **0.92** | **0.92** | **0.65** | **0.66** |

**Table 4.3. Results of paragraph classification task using paragraph vectors using an ensemble of machine learning algorithms. Each individual algorithm predicted a classification for an input test paragraph. A simple majority vote method was used to compute the ensemble predicted label.**

A number of factors are worth noting about the results.

The performance of decision trees get progressively worse as the window size and epoch increase. This may be due to over-fitting of the model as it is continually exposed to the same training set repeatedly. All algorithms perform poorly when the number of epochs is 50, again likely due to over-fitting. This would indicate that decision trees are especially vulnerable to over fitting when using only real valued features. Experiments with varying pruning parameters may lead to improved performance. [Esposito, 1997]

Linear SVC ad Logistic Regression consistently perform well. This implies the data sets are linearly separable. SVC implements a support vector machine but uses a Radial Basis Function (RBF) instead of a linear function and these have proven to be less accurate for text classification tasks.

The ensemble with hard voting method is the best except under the large epoch case (e=50), in which case Linear SVC is best. The results of the ensemble under the large epoch is poor because each algorithm is weighted equally and only Linear SVC performs about F1-score of 0.7.

The high quality results of these classifiers under low epoch training (e=10) indicates the data sets are linearly separable. This is not surprising; text classification tasks using term frequency inverse document frequency and related methods have achieved reasonable performance using linear classifiers such as support vector machines with linear kernels. [Yang, 1999] Achieving reasonably good performance (F1-score > 0.9) using paragraph semantic vectors would indicate that semantic representations perform as well or better than non-semantic representations, such as term frequency inverse document frequency representations. [Sullivan and Wattam, 2016, under review].

**Experiment 2: Similarity by Distance Evaluation**

The results of similarity by distance evaluation were unexpectedly poor. In no case was a paragraph in close proximity, as measured by cosine distance, to a *Clostridium botulinum* or *Mycobacterium tuberculosis* considered similar to the source paragraph. There are several possible reasons that contribute to the poor performance.

The set of 100,000 randomly selected paragraphs did not include a sufficient number of paragraphs on *Clostridium botulinum* or *Mycobacterium tuberculosis*. This would account for the fact that in some cases, the closest paragraph to a source paragraph was not similar as measured by cosine distance. For example, the paragraph above on *C. botulinum* [adapted from Shapario, 1998] is closest to the following paragraph:

> the mean vpt scores and standard deviations for the use of the art approach for operators 1 and 2 in the three independent studies are presented in table 4 children in group a had lower mean vpt scores than children in group b p 0 02 and group c p 0 00001

The cosine between the paragraph semantic vectors of the two paragraphs is 0.54, which under no circumstances can be considered a similar vector.

Another possibility is that additional training epochs are required for paragraph support vectors to reach more optimal configurations. The number of epochs must be chosen with care, however, since the classification experiment shows that an excessive number of training epochs can lead to a decrease in accuracy and F1-score.

A third approach to improving the quality of similarity by distance selection is to use more data. The experiment used on 100,117 paragraphs with a total of 11,912,102 words in context.

## Discussion

We must consider the apparent contradiction of high quality results from the classification experiment and the poor quality results from the similarity by distance experiment. If the quality of the semantic vectors themselves was poor, we would expect poor results when learning classifiers. This did not occur. One possible explanation is that some dimensions learned by the paragraph semantic vectors are more relevant to semantic similarity than others. Supervised classification algorithm can use positive and negative examples to discern which dimensions are most useful in distinguishing between

categories.

Measuring similarity by cosine is a distance measure that does not distinguish among dimensions. Treating all dimensions equally appears to lead to sub-optimal performance. Further investigation is needed to determine which dimensions are the most useful capturing semantic similarity.

Further research is needed on the geometry and topology of the semantic vector space. If the manifold hypothesis holds for this vector space, then paragraphs will cluster in a lower dimension space. [Narayanan and Mitter, 2010] If the semantic vectors do map to lower dimensions, this could contribute to the high quality classification results. It would appear, however, that the manifolds may still harbor distances along dimensions that are not relevant to semantic similarity. One approach is to select a set of paragraphs that are known to be semantically similar and identify the features/dimensions that most contribute to correct classification by machine learning algorithms. This type of analysis should be performed on a range of dimension sizes, training set sizes, and training epochs. If there is high variability between these different test scenarios it may be difficult to discern an underlying structure that can help tune similarity search algorithms.

The semantic vectors used in this research were developed using the Skip-Gram algorithm. The quality of semantic vectors may be improved by employing additional techniques.

As outlined in Chapter 6, section "Representing Knowledge from Ontologies Directly into Vector Space", the knowledge explicitly represented in ontologies can be used to initialize semantic vectors. This is one way to incorporate the declarative knowledge of ontologies into semantic vectors.

Alternatively, semantic vectors can be combined with classifiers that use TF-IDF representations supplemented with information from ontologies, such as the taxonomic lineage of terms defined in biomedical ontologies. This approach would not improve the quality of semantic vectors directly, but it could improve the overall performance of a classification program.

Advances in deep learning may also improve the quality of semantic vectors, although as noted earlier, there are demonstrated cases in which breadth of a single layer can outperform narrower but deeper neural networks. One weakness of the Skip-Gram algorithm is that a single word with multiple meanings are subject to training from multiple contexts. For example, "bank" can appear in contexts about financial institutions or about rivers. Context vectors have been used with TF-IDF and should be adaptable to semantic vector models as well. [Chen, 200] There has been some success in using context to derive syntactic properties that may be useful in improving the performance of semantic

vectors. [Erk, 2008] There has also been progress in capturing both global and local context with neural networks as in [Huang, 2012].

Deep learning techniques are perhaps the most promising avenue to improve the quality of representation. Recurrent neural networks are particularly promising because they can compress the history of previously encountered words in the training stream into a low dimensional space and have the potential to form short term memory, which can capture some aspects of context. [Mikolov, 2010] Convolutional neural networks have also performed well in natural language processing tasks, although these models are better known for their exemplary performance in image processing. [Kalchbrenner, 2014] Recursive neural network have used with morphological features as noted in [Loung, 2013]. Recursive neural networks are computationally more complex than Skip-Gram, but this can capture a broader range of features because it operates at the morphological and not word-based level. For example, "hydrate" and "dehydrate" are often used in similar contexts and therefore have close semantic vectors generated by Skip-Gram. The ability to work at a morphological level would allow a classifier to distinguish the antonym morpheme "de" in "dehydrate" and adjust the semantic vector of both forms of "hydrate."

In all cases, additional research is need to evaluate the trade-offs between the quality of representation and the computational complexity of the learning algorithm. Also, the literature is inconsistent in the use of some deep learning techniques. For example, [Kim, 2014] demonstrated significantly better results with convolution networks for a sentence classification task than [Kalchbrenner, 2014]. Additional formal analysis of deep learning techniques could help shed light on how to configure well established patterns, such as convolutional and recursive networks. Until then, one should expect extended periods of experimentation to find optimal numbers and types of layers in a deep neural network to solve a particular problem.

## References


Breiman, L. “Random Forests”, Machine Learning, 45(1), 5-32, 2001.

Chang, Chih-Chung, and Chih-Jen Lin. "LIBSVM: a library for support vector machines." *ACM Transactions on Intelligent Systems and Technology (TIST)* 2.3 (2011): 27.

Chen, Keh-Jiann, and Jia-Ming You. "A study on word similarity using context vector models." *Computational Linguistics and Chinese Language Processing* 7.2 (2002): 37-58.

Dietterich, Thomas G. "Ensemble methods in machine learning." *Multiple classifier systems*. Springer Berlin Heidelberg, 2000. 1-15.

Erk, Katrin, and Sebastian Padó. "A structured vector space model for word meaning in context." *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2008.

Esposito, Floriana, et al. "A comparative analysis of methods for pruning decision trees." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 19.5 (1997): 476-491.

Fan, Rong-En, et al. "LIBLINEAR: A library for large linear classification." *The Journal of Machine Learning Research* 9 (2008): 1871-1874.

Harris, Zellig S. "Distributional structure." *Word* 10.2-3 (1954): 146-162.

Huang, Eric H., et al. "Improving word representations via global context and multiple word prototypes." *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 2012.

He, Yongqun, et al. "PIML: the pathogen information markup language." *Bioinformatics* 21.1 (2005): 116-121.

Kalchbrenner, Nal, Edward Grefenstette, and Phil Blunsom. "A convolutional neural network for modelling sentences." *arXiv preprint arXiv:1404.2188*(2014).

Kim, Yoon. "Convolutional neural networks for sentence classification." *arXiv preprint arXiv:1408.5882* (2014).

Le, Quoc V., and Tomas Mikolov. "Distributed representations of sentences and documents." *arXiv preprint arXiv:1405.4053* (2014).

Luong, Thang, Richard Socher, and Christopher D. Manning. "Better Word Representations with Recursive Neural Networks for Morphology." *CoNLL*. 2013.

van der Maaten, L.J.P.; Hinton, G.E. Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research 9:2579-2605, 2008.

van der Maaten, L.J.P. t-Distributed Stochastic Neighbor Embedding http://homepage.tudelft.nl/19j49/t-SNE.html

L.J.P. van der Maaten. Accelerating t-SNE using Tree-Based Algorithms. Journal of Machine Learning Research 15(Oct):3221-3245, 2014. http://lvdmaaten.github.io/publications/papers/JMLR_2014.pdf

Mikolov, Tomas, et al. "Recurrent neural network based language model." *INTERSPEECH*. Vol. 2. 2010.

Mikolov, Tomas, Chen, Kai, Corrado, Greg, and Dean, Jeffrey. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013a

Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013b

Mitchell, Jeff and Lapata, Mirella. Composition in distributional models of semantics. Cognitive Science, 2010.

Narayanan, Hariharan, and Sanjoy Mitter. "Sample complexity of testing the manifold hypothesis." *Advances in Neural Information Processing Systems*. 2010.

Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." *The Journal of Machine Learning Research* 12 (2011): 2825-2830.

Quinlan, J. Ross. "Induction of decision trees." *Machine learning* 1.1 (1986): 81-106.

Řehůřek, Radim, and Petr Sojka. Sojka "Software framework for topic modelling with large corpora." *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. 2010.

Rosenblatt, Frank. "The perceptron: a probabilistic model for information storage and organization in the brain." *Psychological review* 65.6 (1958): 386.

Scikit-Learn, ExtraTreeClassifier. http://scikit-learn.org/stable/modules/generated/sklearn.tree.ExtraTreeClassifier.html. Accessed March 19, 2016a.

ScikitLearn, LinearRegression. http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html 2016b.

Shapiro, Roger L., Charles Hatheway, and David L. Swerdlow. "Botulism in the United States: a clinical and epidemiologic review." *Annals of internal medicine* 129.3 (1998): 221-228.

Socher, Richard, Lin, Cliff C, Ng, Andrew, and Manning, Chris. Parsing natural scenes and natural language with recursive neural networks. In Proceedings of the 28th International Conference on Machine Learning (ICML-11), pp. 129–136, 2011. Socher, Richard, Pennington

Sullivan, Daniel and Alice R. Wattam, Limitations of TF-IDF for Classifying Sentences in Biocuration Tasks, 2016. (under review).

Yang, Yiming, and Xin Liu. "A re-examination of text categorization methods." *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1999.

Zanzotto, Fabio, Korkontzelos, Ioannis, Fallucchi, Francesca, and Manandhar, Suresh. Estimating linear models for compositional distributional semantics. In COLING 2010.

Zhang, H. The optimality of Naive Bayes. Proc. FLAIRS. 2004.

Zhu, Ji, et al. "Multi-class adaboost." *Statistics and its Interface* 2.3 (2009): 349-360.

# Chapter 5: Toward a Formal Model of Semantic Vectors

The preceding chapters of this dissertation have demonstrated the limits of a commonly used text classification technique that do not employ semantic representations, evaluated the capacity of semantic vectors to capture relationships by comparing relations captured by semantic vector to those captured in manually defined ontologies, and evaluated the ability of semantic vectors to support classification of paragraph length texts. The results, although promising, highlight a number of shortcomings. Semantic vectors trained on the open access biomedical corpus, for example, are well suited to capturing properties of amino acids but are less successful at capturing taxonomic relations. Classification algorithms applied to semantic vectors perform well (F-1 scores above 0.9) but vectors in close proximity to each other do not always exhibit obvious semantic similarity. Future research in the area of semantic vectors should proceed along at least two fronts: additional experiments with large corpi and varying algorithms for generating semantic vectors, and a formalization of a linguistic-semantic vector model that describes the property of the semantic vector space and its relation to known linguistic phenomenon. This chapter addresses the latter.

## Language and Formal Models

The history of science is rich with examples of scientists making observations, formulating models to explain those observations, making predictions based on those models, and then revising models in light of those predictions and further observations. Ideally, models allow one to derive true statements about the phenomenon that is the subject of the model, but this is not always the case; useful models sometimes entail predictions that do are not supported by data. Useful models are those that make truthful propositions about crucial aspects of a phenomenon. [Bailer-Jones, 2003]. The goal of developing a formal model of semantic vectors should not be to create a model that generates only truthful propositions about linguistic phenomenon but to create a model that advances the effort at hand: to create computational tools to support biocuration.

The “crucial aspects” that [Bailer-Jones, 2003] calls for in this case would include:

Ability to represent semantic entities, such as words, sentences and paragraphs.

Capture important relations between semantic entities, such a similarity.

Model known linguistic phenomenon that are relevant to biocuration.

The semantic vectors examined in this dissertation meet the first crucial aspect. Experiments described here and elsewhere apply skip-gram [Mikolov, 2003], continuous bag of word [Mikolov, 2003], and GLOVE [Pennington, 2014] algorithms to generate semantic vectors or words and text. These vectors capture both semantic properties, such as similarity, and syntactic relations, such a singular-plural relations. Deciding which linguistic phenomenon are relevant to biocuration is a more subjective task. Although there are metrics such as Euclidean distance and cosine between points and vectors for measuring similarity, there are no comparable metrics for evaluating the relative importance of linguistic phenomenon to biocuration. Instead, a subjective defense is offered for defining "crucial aspects" of linguistic phenomenon that should be formally modeled.

**Choosing Linguistic Phenomenon to Model**

Linguists study a wide array of language characteristics, including syntax, semantics, phonology, morphology and pragmatics. Some linguists examine problems that span the constituent components of languages, e.g. syntax and semantics, to understand broader linguistic phenomena, such as language acquisition.

The formal model proposed in this chapter is limited to semantic representations. Vector space models do capture non-semantic features of words, such as some syntactic properties, but these are not considered here. Syntactic properties are ignored it the model because including them would not alter the representational or predictive capacity of the model with regard to semantics.

The model does not address multi-faceted linguistic phenomenon such as language acquisition. This requires an understanding of other disciplines, such as neurolinguistics, that are outside the scope of this research.

The mathematical formalisms proposed here follow from theories and models developed by linguists and cognitive scientists. This is analogous to developing models of metabolic pathways by building from the knowledge and theories created by the biochemists that study those pathways in organisms. Theories from biochemistry and physics should constrain what should be considered in formal models of metabolic pathways. A model of a metabolic pathway that precisely predicts the results of experiments but violates conservation of energy would be of no use to a biochemist. The predictions from such a model could not be accepted. At best, what would appear to be an accurate prediction

would be dismissed as a chance occurrence. Similarly, a formal model of semantic vectors should be constrained by linguistic theories to mitigate the risk of creating a formal model that does not provide a mechanism for making truthful inferences about crucial aspects of semantic vectors.

Linguistics and cognitive science provide us with a set abstractions that are the building blocks of a formal model. For example, the work of Chomsky and others [Chomsky, 2014] [Haegeman, 1991] provide us with grammars, which are useful for modeling syntax. Syntax is used to distinguish well structured from non-well structured, or ungrammatical, sentences. For example,

*S. Typhi* causes typhoid fever.

Is a grammatical sentence in English because it follows the grammar, or rules of ordering and composition, of English. The same words in a different order, however, produce an ungrammatical sentence in English:

causes *S. Typhi* typhoid fever.

Placing a verb before the subject noun phrase in an active voice sentence is grammatically incorrect in English, that is it does not follow the rules of structuring well formed sentences. It is important to note that well-formed sentences are not always semantically correct.

Typhoid fever causes *S. Typhi*.

Is grammatically correct, but factually incorrect. Semantics is the domain of linguistics that addresses meaning and, from that, the truthfulness of statements.

A semantic interpretation of a sentence can be represented or denoted by labeling words and phrases in a sentence with conceptual types, such as Situation, Event, State, Object, Place, Property, Etc. [Jackendoff, 2002]. For example, *S. Typhi* could be an object or agent that cause infectious state known as Typhoid fever. One could even put more of a veil of formality on this by using predicate logic notation such as:

CAUSE(*S. Typhi*, typhoid fever)

and specifying appropriate types constraints on the parameters of the predicate. This type of notation and rewriting is useful. It provides a standard way to represent the meaning of a statement regardless of the syntactic level representation, which could be either:

*S. Typhi* causes typhoid fever.

or

Typhoid fever is caused by *S. Typhi*.

Mapping natural language expressions to a logical form helps to distinguish the roles and functions of entities in an expression. It can also help clarify the meaning but it does not solve the fundamental problem of how to represent meaning of individual constituents of the logical form.

For example, the logical form:

CAUSE(*S. Typhi*, typhoid fever)

does not reflect the semantic relationship between *S. Typhi* and the genus to which it belongs, *Salmonella,* nor does it provide any indication of the characteristics that constitute typhoid fever. These are issues of lexical semantics, or specifying and representing what words (or in some cases phrases) mean. In an effort to begin to frame a formal model of semantic vectors, we must operationalize the term 'meaning' as it relates to semantics. Just a computational biologist will turn to biochemistry for insights about modeling metabolic pathways, a computational linguist turn to philosophy of language and linguistics for insights about modeling meaning.

## Model Foundations: Philosophy of Language and Linguistic Theories

The study of language has not yet produced a comprehensive, well agreed upon model such as the Standard Model of Physics. [Oerter, 2006] Rather than try to describe and assess the advantages and disadvantages of different models, this research works from models that best fit with the task at hand, supporting the automation of biocuration. In particular, this research begins with a philosophical approach to semantics known as semantic molecularism, operationalizes the definition of meaning in semantic molecularism using the Distributional Hypothesis, and links the Distributional Hypothesis to the generation of semantic vectors. Thus, the semantic vector model is based on a defined, although not necessarily universally accepted, model of language.

### Semantic Molecularism

The philosophy of language frames the scope of abstractions used to describe language and identifies important phenomenon that formal models should address. Just as importantly, it identifies the limits of what can be described with abstractions and formal systems. A particularly vexing question in the philosophy of language is, "what is meaning?" This question then leads to others, such as how is meaning shared and how does language related to truth and logical propositions.

One approach to address these questions is to work under the assumption that words have meaning

because their lexical representations, such as the five letter word "chair" corresponds to a set of objects in the world that are used by humans for sitting. This is a denotational semantics, often called Tarskian semantics. [McDermott, 1978] [Tarski, 1944]. The denotational approach models an object language, *L*, using a metalanguage, *M*, which is used to formalize what is said about *L*. Truth of statements is determined using syntactic and set-theoretic operations on expressions in the metalanguage with reference to objects denoted by symbols in formulas of the metalanguage. This approach does not offer useful abstractions for a computational model of semantics, especially since it depends on the ability to interpret the meaning of a metalanguage predicate with respect to denoted objects. For example, there is no obvious way to evaluate the truth of the predicate CHAIR(X) if the evaluation requires determining if the object referenced by variable X, is in fact, a chair. Tarskian semantics is a well established model of semantics but it has limitations, especially with regards to supporting the research at hand. For more on the limitations of this approach, see [Jackendoff, 2002], [Lyons, 2002] [Wittgenstein 1958].

Alternatively, rather than assuming words acquire meaning by reference to a set of objects that exist outside the language, some have argued that words acquire meaning by their relation to other words. The language philosopher, V.O. Quine argued that:

> "It is misleading to speak of the empirical content of an individual statement" (Quine 1951: 43),

and that:

> "the unit of empirical significance is the whole of science" (Quine 1951: 42)

Quine's stand is representative of the model of meaning known as semantic holism. A basic tenet of this approach is that the meaning of works can only be understood by taking into account all words in a language. This approach is too constraining. One does not need to know full language before understanding a subset of words within the language. This is due, at least in part, because some words are frequently used with others and rarely, if ever, used with others. This implies that the range of possible contexts in which a particular word is found is less than the set of all possible contexts of a given language. For example, consider the following passage:

> "A growing body of evidence suggests that Treg modulation could offer a new therapeutic strategy in RA and other autoimmune disorders." [König et. al. 2016]

The term Treg is not explicitly defined in this context but one does not need to know all of English to understand Treg. One can infer that since a Treg can be modulated, it is a biological object such as a

cell or chemical. Furthermore, it is likely part of the immune system given the reference to RA [rheumatoid arthritis] and autoimmune disorders. With additional context, it is likely that once could infer that "Treg" means "Regulatory T cell."

It is not necessary to know all of the words in language to begin to infer meaning. As the Treg example shows, it is necessary to understand the meaning of words around "Treg" to infer the meaning of "Treg." This type of approach to language meaning is known as semantic molecularism. [Block, 1996]

Semantic molecularism assumes word meaning is the function of some subset of words of a language. This is not universally agreed upon and arguments for and against a variety of models can be found in [Fodor and Leopore, 1992] [Dretske, 1981],[Kripke, 1972]. This research effort assumes semantic molecularism because it fits with the problem at hand: supporting the automation of biocuration tasks. There is no need to use a model that may be better at describing language learning, how meanings become shared, or other interesting questions that do not impact the engineering of a biocuration application. It is beyond the scope of this research to formulate a logical, formal definition of word meaning. That is not a practical from an engineering perspective, however, this research should be begin the steps to formalize a computable set of functions that can be applied to semantic representations to yield other semantic representations. The next step in this direction is to define a way to operationalize the concept of semantic molecularism in such as way that one can eventually formulate an algorithm to compute semantic vectors.

## Distributional Hypothesis of Word Meanings

To progress from the starting point of semantic molecularism toward the algorithms and data structures of semantics one needs a conceptual bridge that spans the vague but driving principles from the philosophy of language to a computational implementation. The needed bridge is found in linguistics and is known as the Distributional Hypothesis. [Harris, 1954] and [Sahlgren, 2008].

Under this hypothesis, words with similar meanings have similar distributions in a corpus. For example, the words "car" and "automobile" are synonyms and are likely to be found in similar contexts. *Salmonella typhimurium* and *Escherichia coli* are both enteric pathogens and are likely to found in somewhat similar contexts. There are limits to the Distributional Hypothesis and the example of the pathogens makes clear. While it is true that both bacteria mentioned can cause enteric disease, not all *E. coli* are pathogenic. Furthermore, since *E. coli* is a model organism, it is referenced in many biomedical publications on topics other than enteric diseases.

Another limitation of the distributional hypothesis is that it does not distinguish semantic relatedness from semantic similarity. The words 'hydrate' and 'dehydrate' are semantically related; they are

antonyms. They are not, however semantically similar in the sense of synonyms. Since semantically related terms may also be found within similar distributions of words it is quite possible to have both antonyms and synonyms in similar distributions.

In spite of the limitations of the Distributional Hypothesis, it provides a foundation for formulating algorithms that use words around a particular word to create a semantic vector. In this research the word2vec implementation of the skip-gram algorithm was used. [Mikolov, 2013] The skip-gram algorithm considers the context of a target word as a window of length n words around the target word. Skip-gram trains a vector to predict words in the context of the target word. Let context be $c$ and the target word be $w$, the skip-gram algorithm computes $p(c|w)$. Skip-gram modifies the context by deleting words within the context so as to create a set of n-grams with dropped words. This allows skip-gram to compute probabilities for a larger number of contexts than would otherwise be available if only n-grams were used.

The skip-gram algorithm is just one algorithm for computing semantic vectors based on the Distributional Hypothesis; Continuous Bag of Words [Mikolov, 2013] and Global Vectors for Word Representations (GLoVe) [Pennington, 2014] are alternatives. From the perspective of formalizing the model of semantic vector, the particular algorithm used to generate the vector is less important than the fact that the algorithm consider the context of word uses across a sufficient corpus to generate vectors that reflect semantic and other relations between words.

## Towards an Algebra of Semantic Vectors

This dissertation has addressed issues around collecting corpi, representing paragraphs, sentences and word in ways that reflect statistical or semantic properties. Narrowly defined operations are performed to evaluate the utility of different algorithms and data structures. These are application specific tests that evaluate the data structures and representations that have been created. They do not, however, provide guidance on how to formally reason about meaning or perform algebraic operations on the data structures to yield additional insights about the semantic space. For this we need an algebric formulation.

### Linear Operations and Approximations

Since semantic vectors constitute a linear space, linear algebra is an obvious candidate for reasoning over objects in that space. This is the approach typically taken. For example, in [Mikolov 2013], linear operations are applied to vectors to perform analogical reasoning. Using a 2-dimensional principal component analysis (PCA) projection, [Mikolov, 2013] demonstrated similar vectors between countries and their capital cities. The same study was able to achieve 72% accuracy in analogy tests involving cities and newspapers; sports teams and cities; and companies and executives. In addition to

performing analogical reasoning, [Mikolov 2013] demonstrated that element-wise addition of vectors yields semantically related vectors. Examples include:

Czech + currency = koruna
Vietnam + capital = Hanoi
Airline + German = Airline Lufthansa

Each word or term represents the vector learned by the skip-gram algorithm, addition is element wise and the result is the closest word to the resulting vector. [Milokov, 2013] argues that additive compositionality is effective because words assigned high probabilities by both input words (i.e. words frequently in the context) vectors will have high probability assigned by the composition of the two words (e.g. 'koruna' is frequently in the context of both 'Czech' and 'currency.').

[Mikolov, 2013] and others apply linear operations, such as element wise addition of vectors, to produce semantically related words to input words. Typically, the closest words or terms to resulting vector are used as the result of the operation. The exact sum of two word vectors may not correlate with a word found in the training corpus. Rather than insists on equality of the form

$$Word_1 + Word_2 = Word_3$$

Researchers implicitly recognize that the relation is an approximate one

$$Word_1 + Word_2 \approx Word_3$$

To formalize the practice of working with approximate matches, an algebra of semantic vectors should accommodate some measure of approximation. Before addressing the need to formalize approximate matches, it is worth noting a linguistic phenomenon that also warrants recognition of the need to accommodate approximate word definitions.

**Wittgenstein's Family Resemblance Problem**

In some schools of linguistics and the philosophy of language, a word acquires its meaning by reference to an external object. According to this approach, the word 'Sun' has a meaning only in its relation to a large gaseous object at the center of our solar system. The language philosopher Ludwig Wittgenstein took this approach in his early influential treatise, *Tractatus Logico-Philosophicus* [Wittgenstein, 1922], but later abandoned it in favor of an understanding that words acquire meaning through their use with other words as described in *Philosophical Investigations*

[Wittgenstein, 1953].

Wittgenstein argues against the stand that words can be defined by their essential characteristics. Instead, he postulates that entities we categorize and label with a term have, at best, a set of overlapping characteristics that are not shared among all members of the category. He uses games as an example. Games include board games, card games, ball games, Olympic games, gambling games, etc. He notes:

> For if you look at them you will not see something that is common to *all*, but similarities, relationships and a whole series of them at that. … Look for example at board-games, with their multifarious relationships. Now pass to card-games; here you find many correspondences with the first group, but many common features drop out and others appear. When we pass next to ball games, much that is common is retained, but much is lost - Are they all 'amusing'? … Or is there always a winning and losing, or competition between players? Think of patience. In ball games there is winning and losing; but when a child throws a ball against a wall and catches it again this feature has disappeared. … And we can go through the many, many other groups of games in the same way; can see how similarities crop up and disappear. And the result of this examination is: we see a complicated network of similarities overlapping and criss-crossing: sometimes overall similarities, sometimes similarities of detail. [Wittgenstein, 1953 pp.31-32]

Wittgenstein calls these sets of overlapping similarities "family resemblances" A computational model of linguistics must accommodate some form of semantic theory. One option is an essentialist model that assumes words correspond to entities that exist outside of language and there exists decision procedures for determining which entities are members of a set that constitute the extension of a term (e.g. the set of all celestial bodies undergoing nuclear fusion are members of the category "star"). Another option is to follow Wittgenstein's line of reasoning in *Philosophical Investigations* and consider how words are used in the context of other words. The latter approach is clearly more aligned with the Distributional Hypothesis and the skip-gram algorithm for generating semantic vectors. Leaving philosophical and linguistic questions aside, the Wittgenstein approach is more pragmatically compatible with the research goals at hand.

## Two Types of Imprecision Related to Word Meanings

A formal model of semantic vectors should address the two types of imprecision identified here.

> The first is the imprecision inherent in language itself due to the fact that one cannot define the meaning of words using a list of essential characteristics. Wittgenstein has demonstrated that. One can see this constraint as analogous to a engineer having to account for entropy when designing mechanical systems. When one designs computational linguistic systems that include semantic representations, the system must account for the lack of crisp definitions of meaning.

The second type of imprecision is introduced by methods for computing semantic vectors. Variables such as the content of the training corpus and the order of training examples can result in variations in the final real values assigned to each dimension of each semantic vector. There is no external reference against which one can measure the quality of a semantic vector other than its relation to other vectors and the quality of semantic and other relations between words.

Linear algebra is well suited as the basis for an algebra of semantic vectors but it must be modified to accommodate the inherent imprecision and lack of crisp definitions known to exist with words. Linear algebra modified to accommodate fuzzy set theory is proposed as the foundation for a formal mathematical model of semantic vectors.

**Modeling Imprecision with Fuzzy Set Theory**

Fuzzy sets [Zadeh, 1965] are an extension of classical set theory, which assumes the principle of bivalence, by providing for degree of membership. In classical set theory, membership in a set is determined under the principle of bivalence, in which a proposition such as "object E is an element of set S" is either true or false. For example, the "Virginia" is a member of the set "States of the United States" while "British Columbia" is not a member of the set "States of the United States." Classical set theory sets are sometimes called crisp sets to distinguish them from fuzzy sets, which are not based on bivalent logic. Instead, membership is measured on a continuum, typically [0,1]. An element E is a member of a set S with a degree of membership. For example the temperature 39 degrees Celsius is a member of the set "Warm Temperatures" with membership 0.7 while the temperature 20 degrees Celsius is a member of the same set with membership of 0.3.

Fuzzy sets consist of a pair: a set *S* and a membership function *m*. The function *m* maps each element of set *S* to a value in the range [0,1]. For each object E:

if *m*(E)=0 then E is not included in *S*

if $m$(E)=1 then E is fully included in $S$

if 0 < $m$(E) < 1 then E is a fuzzy member of $S$

Fuzzy sets have been used to model imprecise linguistic statements. [De Cock, 2000]. De Cock [De Cock, 2000] focused on interpreting statements with modifiers that implied some degree of imprecision such as "brighter than average" and "at least middle age." This work on modeling linguistic phenomenon showed limitations of assuming algebraic properties over fuzzy operations. Specifically, transitivity does not hold in fuzzy interpretations of language.

> In every-day life we usually do not feel a difference in temperature between 0○ and 1○ , neither between 1○ and 2○ , between 35○ and 36○ , etc. For us, 0○ and 1○ are certainly approximately equal, and so are 1○ and 2○ , and 35○ and 36○ , etc. To formalize this, consider a universe X of temperatures and a T-equivalence relation E on X used to represent "approximately equal". We would thus expect, E(k, k + 1) = 1 for every k in N. By induction, it is easy to show that, for every k and n in N, E(k, k + n) = 1. This means, in turn, that all temperatures are approximately equal to the degree 1 — obviously, a completely counter-intuitive result. [De Cock, 2000]

De Cock conclude that fuzzy set relations applied to language are symmetric and reflexive but not transitive; the authors propose using a pseudo-metric relation called resemblance that takes into account the distance between two entities. [De Cock, 2000] This distance can be the basis for defining a membership function.

**Fuzzy Set Member Functions and Semantic Vectors**

Semantic vectors are imprecise representations of word meanings. Some imprecision is function of inherent imprecision of semantics, as demonstrated by Wittgenstein's formulation of family resemblance. The skip-gram algorithm, or other means of computing semantic vectors, introduce imprecision as well. The real values computed for each semantic vector is a function of the corpus used and the order in which sentences are presented. Fuzzy set theory was designed to address imprecision in a formal, mathematical way. It is only logical to inquire, could a fuzzy set variant of linear algebra provide a formal foundation for reasoning about semantic vector spaces?

Without delving into formal definitions and proofs, if we assume that a fuzzy set linear algebra such as defined in [Kandasamy, 2008] can provide both linear algebraic operations that have been used with semantic vectors (such as performing analogical reasoning) and accommodates imprecision in ways not currently supported by classical linear algebra, we can formulate an algebra that models more linguistic

phenomenon than currently available with classical linear algebra.

It remains to be demonstrated that fuzzy set linear algebra is a viable model for reasoning about semantic vector spaces; however, if fuzzy set linear algebra is in fact a reasonable formalism, one must define methods for computing the membership function.

Before defining a method of computing a membership function, one must define characteristics and properties of such a function. Just as choosing among different characteristics of algebras leads to different type of mathematical entities (e.g. groups, rings, and fields), varying decisions can lead to different types of membership functions.

**Word Membership Functions, Individual Words and Clusters of Words**

Given a word $w$, and a semantic vector of that word $SV(w)$, a membership function, $m$, may be defined such that $m(SV(w)) = 1$ by definition. In other words, a word is a fully included member of the fuzzy set defined by that word. The function $m$ may be further defined with a distribution such that $m(SV(w) + \Delta)$, where $\Delta$ is a vector, computes a membership value in the range [0, 1]. Options for specifying the distribution is discussed below.

When a number of semantic vectors are in close proximity, the membership function may be defined over a cluster of similar terms. For example, the Salmonella serovars *S. enterica* subsp.*salamae, S. enterica subsp.arizonae, S. enterica subsp .diarizonae,* and *S. enterica subsp. houtenae* may be in close proximity. In such cases, one could define a membership function $m(SW(w)) = 1$ when w = '*S. enterica*.' Alternatively, the centroid of the four semantic vectors of the serovar names could be defined as the point in the vector space designating the fully included member of the fuzzy set defined for the *S. enterica* serovars. For each serovar $w$, $m(SV(w))$ is in the range [0,1] and the value of the membership is a function of the semantic vector of w distance from the centroid.

**Distribution of Membership Functions**

Determining appropriate distributions of membership functions is challenging. Considerations include:

Mapping semantic similarity between concepts

Density of vectors in different regions of the vector space

Need for different types of membership functions for different types of entities

The distribution should be based, to some degree, on the distribution of semantic vectors in a vector space. Distribution should ideally, follow some concept of similarity and family resemblance. Words

and concepts with close family resemblance should have larger membership values than words or concepts that do not. For example, *Clostridium difficile* and *Salmonella enterica* would both have high membership values in the set of enteric pathogens while *Mycobacterium tuberculosis* would have a much lower membership value.

Membership function values should reflect the density of a region of the vector space. For example, the region of a vector space with term related to enteric pathogens may be densely populated with enteric pathogens, including a large number of serovars and pathovars. Assume a membership function based on Euclidean distance with the term 'enteric pathogen' as a fully included member of a fuzzy set EP. There may exist large number of other pathogens with high membership values in EP. Similarly, there may be a sparsely populated region of space around the semantic vector representing prions, misfolded protiens capable of acting as the etiol agent of some diseases. [Imran, 2011] There may be few pathogens with high membership values in EP if the same membership function is used for both enteric pathogens and prions. This results seems intuitively plausible since the object of the skip-gram algorithm is to generate vectors according to the Distributional Hypothesis.

Another factor to consider when formulating membership functions is the potential need for different types of membership functions for different types of semantic entities. One type of membership function may work well for biological concepts but less well for chemistry or physics concepts. Additional research is needed to evaluate a broad set of membership functions over a variety of corpi. Future research should also investigate the relation between the quality of a vector space, which is a function of the training corpus, vectorization algorithm, and parameters, and the ability of membership functions to capture semantic relatedness.

## Summary

Advances in the application of semantic vectors to biocuration will come from additional engineering efforts to algorithms and corpus construction as well as efforts to formalize the representation of meaning and formal operations on those representations. Just as computational models of metabolic pathways are informed and constrained by biochemistry, the computational models of language are constrained by linguistics. There are varied approaches to the philosophy of language and linguistics. This effort to begin to formalize a model of semantic vector works with the assumptions of semantic molecularism and the related operational form known as the Distributional Hypothesis. To ensure any formal model of semantic vectors is sufficiently descriptive, it is argued that the formal model should capture some aspects of the family resemblance phenomenon identified by Wittgenstein. Fuzzy set theory applied to linear algebra is proposed as a suitable mathematical foundation to begin a formal model of semantic vectors. There is still much work to be done with this formalization, especially with regard to defining membership functions.

## References


Bailer-Jones, Daniela M. "When scientific models represent." *International Studies in the Philosophy of Science* 17.1 (2003): 59-74.

Block, Ned. "Holism, mental and semantic." *The Routledge Encylopedia of Philosophy* (1996).

Chomsky, Noam. *Aspects of the Theory of Syntax*. Vol. 11. MIT press, 2014.

De Cock, Martine, Ulrich Bodenhofer, and Etienne E. Kerre. "Modelling linguistic expressions using fuzzy relations." *Proceedings of the 6th International Conference on Soft Computing', Iizuka*. 2000.

Dretske, F., 1981, *Knowledge and the Flow of Information*, Cambridge: MIT.

Fodor, Jerry A., and Ernest Lepore. "Holism: A shopper's guide." (1992).

Haegeman, Liliane. *Introduction to government and binding theory*. 1991.

Harris, Z. (1954). "Distributional structure". *Word* **10** (23): 146–162.

Imran, Muhammad, and Saqib Mahmood. "An overview of human prion diseases." *Virology journal* 8.1 (2011): 1.

Jakendoff, Ray, 2002, *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford: Oxford University press.

Kandasamy, WB Vasantha, Florentin Smarandache, and K. Ilanthenral. *Set Linear Algebra and Set Fuzzy Linear Algebra*. Infinite Study, 2008.

König, Martin, et al. "Tregalizumab–A Monoclonal Antibody to Target Regulatory T Cells." *Frontiers in immunology* 7 (2016).

Kripke, S., 1972, *Naming and Necessity*, Cambridge: Harvard University Press.

Lyons, John., 2002. *Linguistic Semantics: An Introduction*. Cambridge, UK: Cambridge University Press.

Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." *Advances in neural information processing systems*. 2013.

R. Oerter (2006). *The Theory of Almost Everything: The Standard Model, the Unsung Triumph of Modern Physics* (Kindle ed.). Penguin Group.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global Vectors for Word

Representation." *EMNLP*. Vol. 14. 2014.

Quine, W.V., 1951, “Two dogmas of empiricism”, reprinted in W.V. Quine, 1953, *From a logical point of view*, Cambridge: Harvard University Press.

Sahlgren, Magnus. "The distributional hypothesis." *Italian Journal of Linguistics* 20.1 (2008)

Wittgenstein, Ludwig. *Tractatus Logico-Philosophicus* [1922]. Translated by CK Ogden. (1983).

Wittgenstein, Ludwig 1953. *Philosophical Investigations*, 3rd. Edition. Eagle Wood Cliffs, New Jersey. Translated by G.E.M Anscombe. (1958)

Zadeh, Lotfi A. "Fuzzy sets." *Information and control* 8.3 (1965): 338-353.

Zadeh, Lotfi Asker. "Fuzzy sets as a basis for a theory of possibility." *Fuzzy sets and systems* 1.1 (1978): 3-28.

# Chapter 6: Conclusions

The research presented in this dissertation focuses on problems of information retrieval for biocuration. It includes an evaluation of commonly used term frequency-inverse document frequency approaches to document classification; the use of dense word vectors, or semantic vectors, for capturing semantic and related features of words, phrases and longer text structures; the complementary types of relations captured by semantic vectors and manually curated ontologies; and the utility of using paragraph-level semantic vectors for document classification. The research also exposed a number of additional potential areas of research related to automating tools for biocuration. This concluding chapter of the dissertation examines five areas that warrant additional research:

Assessing Ensemble methods combining TF-IDF and semantic vector methods
Representing knowledge represented in ontologies in semantic vectors
Examining the macro- and meso-structures of vector space
Developing a fuzzy linear algebra of semantic vectors
Evaluate user interface design considerations for building biocuration systems based on semantic vector.

Research in each of these areas could materially advance the application of semantic vector spaces to biocuration. The chapter concludes with a discussion of how this research generalizes to other text analysis tasks.

## Assessing Ensemble of TF-IDF and Semantic Vector Approaches to Text Classification

Ensemble learning methods [Dietterich, 2000] combine multiple machine learning algorithms to leverage the relative strengths of each algorithm while compensating for its weakness. TF-IDF representations have been successfully applied to a wide range of information retrieval tasks. [Ponte, 1998]. Research performed as part of this dissertation demonstrated the use of TF-IDF representations with a variety of machine learning algorithms. Since TF-IDF captures measures of word frequency directly from texts with minimal transformation, it can be implemented efficiently. TF-IDF representations used with linear classifiers, such as Support Vector Machines with linear kernels, perform well. There are, however, limitations to the set of features that can be represented in TF-IDF. Chapter 2 of this dissertation demonstrates that even with additional training examples, classification algorithms trained on TF-IDF representations did not improve. This is evidence for a bias in the representation that can not be overcome without modifying the features represented. Although it may be possible to manually engineer additional features, this is not a viable or scalable solution for

biocuration. Feature generation should be automated.

Semantic vectors capture features using unsupervised learning techniques such as the Skip-gram algorithm. [Mikolov, 2013]. The range of features is controlled by the size of the semantic vector. This allows developers to easily modify the set of features captured by varying the length of the semantic vector. Additional training data typically improves the quality of semantic vectors as measured by several metrics, however, overtraining on limited corpus can lead to a decrease in performance.

## Representing Knowledge from Ontologies Directly into Vector Space

Manually curated ontologies capture complementary relations to those captured by the Skip-gram algorithm. This presents an opportunity to combine the two methods of knowledge representation into a single representation scheme. From a pragmatic perspective, these are the kinds of relations captured by ontologies that are important to capture in semantic vector models. Unfortunately, the kinds of information captured in ontologies is not readily extracted from the biomedical corpus. Ontologies represent facts that one would expect to find in textbooks, they are foundational knowledge for practitioners in life sciences. This kind of information is not discussed extensively in the biomedical research literature since it is typically assumed that readers of the research literature are already familiar with this fundamental knowledge.

One possible solution to the problem of capturing foundational knowledge is to use textbooks as a training corpus. This is a reasonable approach, but there are limitations. Fundamental knowledge may be stated as facts. Consider the definition of a lipid bilayer:

> **Lipid bilayer** is a universal component of all cell membranes. The structure is called a "lipid bilayer" because it composed of two layers of fatty acids organized in two sheets. The lipid bilayer is typically about five nanometers to ten nanometers thick and surrounds all cells providing the cell membrane structure. With the hydrophobic tails of each individual sheet interacting with one another, a hydrophobic interior is formed and this acts as a permeability barrier. The hydrophilic head groups interact with the aqueous medium on both sides of the bilayer. The two opposing sheets are also known as leaflets. [Wikimedia Foundation, 2016]

A formal definition such as this includes a series of logical statements about the form and function of an entity. With sufficient example text such as this, a Skip-gram analysis would likely yield semantic vectors in which lipid bilayer, hydrophilic and hydrophobic are related by a vector. It may capture the fact that the lipid bilayer is a part of a cell membrane but may not capture the fact that a lipid bilayer is a supramolecular entity as defined by the NanoParticle Ontology [Thomas, 2011] or is semantically

related to sarcolemma, a plasma membrane which found in striated muscle fibers, according to the Computer of Retrieval of Information on Scientific Thesaurus ontology. [Bair, 1995]

An alternative approach is to build a set of semantic vectors using a biomedical corpus and then adjust the vectors of ontology terms to better reflect the distances found between such terms in ontologies. A high level, proposed approach includes:

Training a semantic vector space using Skip-gram and a large biomedical corpus, such as the set of PubMed Medline abstracts and the PubMed Central Open Access set of publications.

Identify an ontology and collect the set of all terms defined in that ontology.

For each pair of terms in an ontology, calculate distance between each pair using an ontology similarity metric, such as used in Chapter 4.

For each pair of terms in the ontology, calculate distance between each ontology term in the vector space using cosine similarity.

For each pair of terms, calculate the error of the pair as the difference between the ontology distances (Step 3) and vector space distance (Step 4)

Adjust semantic vectors to reduce the sum of errors (or sum of square of errors) and optimize to minimize the sum of errors (or sum of squared errors). This is fundamentally a search problem analogous to finding weights for neural networks. [Rumelhart, 1988] Genetic algorithms and stochastic gradient descent [Bottou, 2010] could be used to search the search space.

After adjusting to minimize the error in distance calculations between the ontology and the vector space, train the vector space again using the original training set to adjust non-ontology terms with respect to the adjusted values of ontology terms.

Note that this method is not guaranteed to produce a set of semantic vectors that reflect distance as measured by ontologies because step 7 will update all vectors, including the vectors associated with the ontology terms.

A number of parameters need to be explored for this approach. The optimal size of vectors needs to be determined. If there are a variety of semantic distance measures for ontologies, which should be used here? Semantic vectors will be modified to minimize the error calculation. A learning rate parameter should be used to control the size of changes to semantic vectors in single step. A learning rate the produces too large changes may miss the optimal configuration (minimized error) but a parameter that yields small changes can prevent the algorithm from finding an optimal solution in reasonable amount of time.

## Study the Structure of Vector Spaces Produced by the Skip-Gram Algorithm

Most of the analysis performed in this research has focused on local relations between words and paragraphs. Additional insights into how to apply semantic vectors space could follow from an investigation of the meso- and macro-scale properties of semantic vector spaces.

It is difficult to visualize large dimensional spaces. Commonly used techniques, such as principal component analysis (PCA) map data from high to low dimensional spaces. Such techniques may lose important information about the structure of vector spaces. Alternatively, differential topology may offer useful techniques. Saeki proposes to use differential topology techniques to hierarchically analyze data in large dimension spaces and notes that it may be difficult to interpret differential topological features so this approach may trade one difficult interpretation problem for another. [Saeki, 2014] Zomorodian proposes topological data analysis for discerning the shape of data using topological techniques. [Zomorodian, 2007]

Questions to consider in this area include, can topology inform understanding of macro- and meso-scale structures of vector space. For example, are there lower dimensional manifolds? If so, do they correspond to some discernible semantic structure, such as terms about a particular domain such as infectious disease, virulence factors, macromolecules, etc?

## Developing a Fuzzy Linear Algebra of Semantic Vectors

This research proposes the combination of fuzzy set theory and linear algebra as a formal foundation for modeling semantic vector spaces. There is much work to be done in this area.

One area of research needs to address predictions of the formal model that do not correspond to the properties of computed semantic vector spaces. For example, an algebraic group includes an inverse function. In the case of linear algebra, the inverse of a vector can be computed by scalar multiplying a vector by -1. Under this model, if the vector for “hydrate” is V, then the vector for “dehydrate” should be in close proximity to -1V. Skip-gram, however, places “dehydrate” in close proximity to “hydrate” as measured by cosine. This is understandable because ‘hydrate’ and ‘dehydrate’ appear in similar contexts in the biomedical literature. The naive interpretation of the inverse operation in semantic vector spaces does not hold. Further analysis is needed to understand how this aspect of the inverse property affects other potential operations. Is this discrepancy between the formal model prediction and the actual output of the Skip-gram algorithm mean linear algebras or groups are not the appropriate mathematical structure for modeling semantics?

Another area of research is defining other operations on the semantic vector space in addition to analogical reasoning and composition. Further work is needed to understand how fuzzy sets apply to Wittgenstein's problem of family resemblance. Additional experiments and analysis are needed to better understand how to define fuzzy set membership functions.

## Evaluate User Interface Design Considerations for Biocuration Systems Based on Semantic Vectors

This research is motivated by the pragmatic needs of biocurators. It is reasonable to consider how this research directly influences the way biocuration systems are designed.

The experiments conducted as part of this project focused primarily on classifying and identifying texts that would be useful to biocurators performing specific tasks, e.g. identifying facts about virulence factors. This research is also applicable to improving the human factors aspects of building biocuration applications.

When engineering a biocuration interface, it may be possible to capture feedback from user with minimal impact on a user's workflow. For example, a user interface could track the navigation path followed by a user by recording all interactions, such a entering a query, selecting a document from a query result set, or following a link from one document to a set of similar documents. A user may start with a document, review similar documents, follow links to some subset, and repeat the process. This data can be used to answer multiple questions, such as:

What patterns can be inferred from the user's choices? Are all documents on a single vector?

Are there other documents on that vector that may be of interest?

Are some dimensions more important than others? If so, can those dimensions be weighted more in ranking of similar documents to present documents more likely to be of interest the user? If so, how long should past interactions be considered. The weighting scheme should have decay rate parameter on the function over previous selections. Fast decay rate would more heavily weight more recent searches while smaller decay rate would include the influence of older selections.

User interface and human factors engineering are integral to designing an effective biocuration application. Further research should examine the overlap between using semantic vectors for

representing meaning and capturing user interaction data to help understand and support user's tasks.

## Generalizing to Other Text Analysis Tasks

Semantic vectors are a fundamental representation scheme that captures, to some degree, the meaning of words, sentences, and paragraphs. The experiments in this dissertation provide evidence for the hypothesis that a semantically rich representation can improve upon the performance of a text representation scheme that focuses on non-semantic features, such as word frequency. It is reasonable to propose that, as a fundamental representation scheme, semantic vectors should generalize to other tasks. One researcher has clearly argued for the importance of semantic vectors to deep learning applied to natural language processing: "Our results add to the well-established evidence that unsupervised pre-training of word vectors is an important ingredient in deep learning for NLP." [Kim, 2014]

In addition to text classification, semantic vectors can be applied to other text mining tasks, including part of speech tagging, named entity recognition, and sentence parsing. Researchers were recently able to obtain state of the art performance in part of speech tagging using neural networks and semantic vector representations. [Tsuboi, 2014] Named entity recognition is an especially difficult task in the biomedical domain. [Ananiadou, 2011]. A recent application of semantic vectors to named entity recognition (NER) demonstrated promising results but also found that performance increased with the size of the training set but only up to a certain point. [Siencnik, 2015]. More experimentation is required to understand impact of training set size on NER performance. A recent paper on sentence parsing using convolutional networks, semantic vectors, and variable size convolutional filters demonstrate favorable results with multi-class sentiment prediction. [Yin, 2016]

Just as additional techniques can help improve the quality of semantic vectors, semantic vectors can improve the quality of performance on a number of text mining related tasks. Some experimentation is required to find optimal numbers and types of layers in deep neural networks, but as the theory of deep learning advances, we should have a more solid, formal foundation for reasoning about deep learning network design.

## Summary

The research described in this dissertation identified limitations in term frequency-inverse document frequency representations for common biocuration tasks, such as classifying texts. The experiments described early demonstrate that semantic word vectors capturing semantic relations that complement

those found in manually curated ontologies. They also demonstrate that semantic vector representations can be used effectively with longer texts, such as paragraphs. In addition to revealing important properties of semantic vectors and offering insight into improving the quality of biocuration support tools, these experiments revealed a number of questions that require further research.

Open issues include: evaluating the potential improvements that might arise by using ensemble of TF-IDF and semantic vector methods; developing methods for capturing relations defined in ontologies but not captured by Skip-gram or related algorithms; analyzing the meso- and macro-scale properties of vector spaces and relation of those properties to semantic phenomenon; developing a formal model of semantic vector spaces that address both linguistic phenomena, such as family resemblance, and properties of vector spaces generated by Skip-gram; and finally evaluating how semantic vector spaces can be used to improve end user applications in biocuration.

# References


Ananiadou, Sophia, et al. "Named entity recognition for bacterial type IV secretion systems." *PLoS One* 6.3 (2011): e14780.

Bair, A. H., et al. "Taking a bite out of CRISP. Strategies on using and conducting searches in the Computer Retrieval of Information on Scientific Projects database." *Computers in nursing* 14.4 (1995): 218-24.

Bottou, Léon. "Large-scale machine learning with stochastic gradient descent." *Proceedings of COMPSTAT'2010*. Physica-Verlag HD, 2010. 177-186.

Dietterich, Thomas G. "Ensemble methods in machine learning." *Multiple classifier systems*. Springer Berlin Heidelberg, 2000. 1-15.

Kim, Yoon. "Convolutional neural networks for sentence classification." *arXiv preprint arXiv:1408.5882* (2014).

Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." *Advances in neural information processing systems*. 2013.

Ponte, Jay M., and W. Bruce Croft. "A language modeling approach to information retrieval." *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1998.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. "Learning representations by back-propagating errors." *Cognitive modeling* 5.3 (1988): 1.

Saeki, Osamu, and Shigeo Takahashi. "Visual data mining based on differential topology: a survey." *Pacific Journal of Mathematics for Industry*6.1 (2014): 1-10.

Siencnik, Scharolta Katharina. "Adapting word2vec to named entity recognition." *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*. 2015.

Thomas, Dennis G., Rohit V. Pappu, and Nathan A. Baker. "NanoParticle Ontology for cancer nanotechnology research." *Journal of biomedical informatics* 44.1 (2011): 59-74.

Tsuboi, Yuta. "Neural Networks Leverage Corpus-wide Information for Part-of-speech Tagging." *EMNLP*. 2014.

Wikimdeia Foundation, https://en.wikibooks.org/wiki/Structural_Biochemistry/Lipids/Lipid_Bilayer. Accessed April 6, 2016.

Yin, Wenpeng, and Hinrich Schütze. "Multichannel variable-size convolution for sentence classification." *arXiv preprint arXiv:1603.04513* (2016).

Zomorodian, Afra. "Topological data analysis." *Advances in Applied and Computational Topology. Proceedings of Symposia in Applied Mathematics*. Vol. 70. 2007.

# Appendix A: Chapter 3 Tables

**Table 1. Amino acid similarity measures ranked by MESH similarity.**

| MESH Sanchez Lin | | |
|---|---|---|
| Amino Acid 1 | Amino Acid 2 | Distance |
| alanine | tyrosine | 0.7157868686 |
| alanine | arginine | 0.7082664778 |
| alanine | glycine | 0.7082664778 |
| alanine | serine | 0.7049472391 |
| alanine | phenylalanine | 0.7031853627 |
| alanine | methionine | 0.6985206877 |
| alanine | aspartic_acid | 0.6974095605 |
| alanine | cysteine | 0.6962679332 |
| alanine | lysine | 0.6941910709 |
| alanine | histidine | 0.6917278676 |
| alanine | proline | 0.6913569185 |
| alanine | threonine | 0.68744862 |
| alanine | valine | 0.68744862 |
| alanine | glutamine | 0.6868804664 |
| alanine | tryptophan | 0.6868804664 |
| alanine | glutamic_acid | 0.6864329611 |
| alanine | asparagine | 0.6826608009 |
| alanine | isoleucine | 0.6826608009 |
| alanine | leucine | 0.6826608009 |
| arginine | lysine | 0.9109880404 |
| arginine | glutamine | 0.9012806442 |
| arginine | asparagine | 0.8956786701 |
| arginine | phenylalanine | 0.8564954157 |
| arginine | methionine | 0.8507452595 |
| arginine | histidine | 0.842373389 |
| arginine | threonine | 0.8371004003 |
| arginine | valine | 0.8371004003 |
| arginine | tryptophan | 0.8364003666 |
| arginine | isoleucine | 0.8312016605 |
| arginine | leucine | 0.8312016605 |
| arginine | tyrosine | 0.7246297412 |
| arginine | glycine | 0.7169233882 |
| arginine | serine | 0.7135227084 |
| arginine | alanine | 0.7082664778 |
| arginine | aspartic_acid | 0.7058015316 |
| arginine | cysteine | 0.7046322875 |

| | | |
|---|---|---|
| arginine | proline | 0.699602997 |
| arginine | glutamic_acid | 0.6945613075 |
| asparagine | arginine | 0.8956786701 |
| asparagine | lysine | 0.8783117384 |
| asparagine | glutamine | 0.8692847932 |
| asparagine | serine | 0.8670411215 |
| asparagine | methionine | 0.8593302798 |
| asparagine | cysteine | 0.8566265101 |
| asparagine | threonine | 0.8460374094 |
| asparagine | tyrosine | 0.6978496158 |
| asparagine | glycine | 0.6906995341 |
| asparagine | phenylalanine | 0.6858664844 |
| asparagine | alanine | 0.6826608009 |
| asparagine | aspartic_acid | 0.6803705733 |
| asparagine | histidine | 0.6749620406 |
| asparagine | proline | 0.6746088509 |
| asparagine | valine | 0.6708871056 |
| asparagine | tryptophan | 0.6703459865 |
| asparagine | glutamic_acid | 0.6699197598 |
| asparagine | isoleucine | 0.6663264318 |
| asparagine | leucine | 0.6663264318 |
| aspartic_acid | glutamic_acid | 0.9542405893 |
| aspartic_acid | tyrosine | 0.7132693918 |
| aspartic_acid | glycine | 0.7058015316 |
| aspartic_acid | arginine | 0.7058015316 |
| aspartic_acid | serine | 0.7025053025 |
| aspartic_acid | phenylalanine | 0.7007555961 |
| aspartic_acid | alanine | 0.6974095605 |
| aspartic_acid | methionine | 0.6961229956 |
| aspartic_acid | cysteine | 0.6938856551 |
| aspartic_acid | lysine | 0.6918229594 |
| aspartic_acid | histidine | 0.6893765033 |
| aspartic_acid | proline | 0.6890080713 |
| aspartic_acid | threonine | 0.6851262096 |
| aspartic_acid | valine | 0.6851262096 |
| aspartic_acid | glutamine | 0.6845618866 |
| aspartic_acid | tryptophan | 0.6845618866 |
| aspartic_acid | isoleucine | 0.6803705733 |
| aspartic_acid | leucine | 0.6803705733 |
| aspartic_acid | asparagine | 0.6803705733 |
| cysteine | methionine | 0.9138323301 |
| cysteine | serine | 0.8844494812 |
| cysteine | threonine | 0.8626045064 |
| cysteine | glutamine | 0.8618951403 |
| cysteine | asparagine | 0.8566265101 |

| | | |
|---|---|---|
| cysteine | tyrosine | 0.712075295 |
| cysteine | glycine | 0.7046322875 |
| cysteine | arginine | 0.7046322875 |
| cysteine | phenylalanine | 0.699602997 |
| cysteine | alanine | 0.6962679332 |
| cysteine | aspartic_acid | 0.6938856551 |
| cysteine | lysine | 0.6906995341 |
| cysteine | histidine | 0.688261003 |
| cysteine | proline | 0.687893762 |
| cysteine | valine | 0.6840244109 |
| cysteine | tryptophan | 0.6834619008 |
| cysteine | glutamic_acid | 0.6830188375 |
| cysteine | isoleucine | 0.6792840052 |
| cysteine | leucine | 0.6792840052 |
| glutamic_acid | aspartic_acid | 0.9542405893 |
| glutamic_acid | tyrosine | 0.7017919849 |
| glutamic_acid | glycine | 0.6945613075 |
| glutamic_acid | arginine | 0.6945613075 |
| glutamic_acid | serine | 0.691368993 |
| glutamic_acid | phenylalanine | 0.6896742537 |
| glutamic_acid | alanine | 0.6864329611 |
| glutamic_acid | methionine | 0.68518654 |
| glutamic_acid | cysteine | 0.6830188375 |
| glutamic_acid | lysine | 0.6810201498 |
| glutamic_acid | histidine | 0.6786493691 |
| glutamic_acid | proline | 0.6782923109 |
| glutamic_acid | threonine | 0.6745299255 |
| glutamic_acid | valine | 0.6745299255 |
| glutamic_acid | tryptophan | 0.6739829165 |
| glutamic_acid | glutamine | 0.6739829165 |
| glutamic_acid | isoleucine | 0.6699197598 |
| glutamic_acid | leucine | 0.6699197598 |
| glutamic_acid | asparagine | 0.6699197598 |
| glutamine | arginine | 0.9012806442 |
| glutamine | lysine | 0.8836979243 |
| glutamine | serine | 0.8724390429 |
| glutamine | asparagine | 0.8692847932 |
| glutamine | methionine | 0.8646323241 |
| glutamine | cysteine | 0.8618951403 |
| glutamine | threonine | 0.8511761986 |
| glutamine | tyrosine | 0.702259747 |
| glutamine | glycine | 0.6950194773 |
| glutamine | phenylalanine | 0.6901259965 |
| glutamine | alanine | 0.6868804664 |
| glutamine | aspartic_acid | 0.6845618866 |

| | | |
|---|---|---|
| glutamine | histidine | 0.67908678 |
| glutamine | proline | 0.6787292615 |
| glutamine | valine | 0.6749620406 |
| glutamine | tryptophan | 0.6744143308 |
| glutamine | glutamic_acid | 0.6739829165 |
| glutamine | isoleucine | 0.6703459865 |
| glutamine | leucine | 0.6703459865 |
| glycine | tyrosine | 0.7246297412 |
| glycine | arginine | 0.7169233882 |
| glycine | serine | 0.7135227084 |
| glycine | phenylalanine | 0.7117177608 |
| glycine | alanine | 0.7082664778 |
| glycine | methionine | 0.706939582 |
| glycine | aspartic_acid | 0.7058015316 |
| glycine | cysteine | 0.7046322875 |
| glycine | lysine | 0.7025053025 |
| glycine | histidine | 0.6999828501 |
| glycine | proline | 0.699602997 |
| glycine | threonine | 0.6956011808 |
| glycine | valine | 0.6956011808 |
| glycine | tryptophan | 0.6950194773 |
| glycine | glutamine | 0.6950194773 |
| glycine | glutamic_acid | 0.6945613075 |
| glycine | isoleucine | 0.6906995341 |
| glycine | leucine | 0.6906995341 |
| glycine | asparagine | 0.6906995341 |
| histidine | tyrosine | 0.8481672141 |
| histidine | arginine | 0.842373389 |
| histidine | phenylalanine | 0.8364003666 |
| histidine | methionine | 0.830916001 |
| histidine | lysine | 0.825824815 |
| histidine | proline | 0.819549672 |
| histidine | threonine | 0.8178949329 |
| histidine | valine | 0.8178949329 |
| histidine | tryptophan | 0.8172266394 |
| histidine | isoleucine | 0.8122628454 |
| histidine | leucine | 0.8122628454 |
| histidine | glycine | 0.6999828501 |
| histidine | serine | 0.6967406209 |
| histidine | alanine | 0.6917278676 |
| histidine | aspartic_acid | 0.6893765033 |
| histidine | cysteine | 0.688261003 |
| histidine | glutamine | 0.67908678 |
| histidine | glutamic_acid | 0.6786493691 |
| histidine | asparagine | 0.6749620406 |

| | | |
|---|---|---|
| isoleucine | valine | 0.9465873914 |
| isoleucine | leucine | 0.9401525139 |
| isoleucine | arginine | 0.8312016605 |
| isoleucine | phenylalanine | 0.8253854715 |
| isoleucine | methionine | 0.8200441452 |
| isoleucine | lysine | 0.815084918 |
| isoleucine | histidine | 0.8122628454 |
| isoleucine | threonine | 0.807358987 |
| isoleucine | tryptophan | 0.8067077934 |
| isoleucine | tyrosine | 0.6978496158 |
| isoleucine | glycine | 0.6906995341 |
| isoleucine | serine | 0.6875425389 |
| isoleucine | alanine | 0.6826608009 |
| isoleucine | aspartic_acid | 0.6803705733 |
| isoleucine | cysteine | 0.6792840052 |
| isoleucine | proline | 0.6746088509 |
| isoleucine | glutamine | 0.6703459865 |
| isoleucine | glutamic_acid | 0.6699197598 |
| isoleucine | asparagine | 0.6663264318 |
| leucine | valine | 0.9465873914 |
| leucine | isoleucine | 0.9401525139 |
| leucine | arginine | 0.8312016605 |
| leucine | phenylalanine | 0.8253854715 |
| leucine | methionine | 0.8200441452 |
| leucine | lysine | 0.815084918 |
| leucine | histidine | 0.8122628454 |
| leucine | threonine | 0.807358987 |
| leucine | tryptophan | 0.8067077934 |
| leucine | tyrosine | 0.6978496158 |
| leucine | glycine | 0.6906995341 |
| leucine | serine | 0.6875425389 |
| leucine | alanine | 0.6826608009 |
| leucine | aspartic_acid | 0.6803705733 |
| leucine | cysteine | 0.6792840052 |
| leucine | proline | 0.6746088509 |
| leucine | glutamine | 0.6703459865 |
| leucine | glutamic_acid | 0.6699197598 |
| leucine | asparagine | 0.6663264318 |
| lysine | arginine | 0.9109880404 |
| lysine | glutamine | 0.8836979243 |
| lysine | asparagine | 0.8783117384 |
| lysine | phenylalanine | 0.8393929639 |
| lysine | methionine | 0.8338694121 |
| lysine | histidine | 0.825824815 |
| lysine | threonine | 0.8207563456 |

| | | |
|---|---|---|
| lysine | valine | 0.8207563456 |
| lysine | tryptophan | 0.8200833698 |
| lysine | isoleucine | 0.815084918 |
| lysine | leucine | 0.815084918 |
| lysine | tyrosine | 0.7099032073 |
| lysine | glycine | 0.7025053025 |
| lysine | serine | 0.6992397183 |
| lysine | alanine | 0.6941910709 |
| lysine | aspartic_acid | 0.6918229594 |
| lysine | cysteine | 0.6906995341 |
| lysine | proline | 0.6858664844 |
| lysine | glutamic_acid | 0.6810201498 |
| methionine | cysteine | 0.9138323301 |
| methionine | serine | 0.887332034 |
| methionine | threonine | 0.8653462048 |
| methionine | glutamine | 0.8646323241 |
| methionine | asparagine | 0.8593302798 |
| methionine | arginine | 0.8507452595 |
| methionine | phenylalanine | 0.8446533516 |
| methionine | lysine | 0.8338694121 |
| methionine | histidine | 0.830916001 |
| methionine | valine | 0.8257850392 |
| methionine | tryptophan | 0.8251037951 |
| methionine | isoleucine | 0.8200441452 |
| methionine | leucine | 0.8200441452 |
| methionine | tyrosine | 0.7144316721 |
| methionine | glycine | 0.706939582 |
| methionine | alanine | 0.6985206877 |
| methionine | aspartic_acid | 0.6961229956 |
| methionine | proline | 0.690092568 |
| methionine | glutamic_acid | 0.68518654 |
| phenylalanine | tyrosine | 0.9271445905 |
| phenylalanine | tryptophan | 0.8895257847 |
| phenylalanine | arginine | 0.8564954157 |
| phenylalanine | methionine | 0.8446533516 |
| phenylalanine | lysine | 0.8393929639 |
| phenylalanine | histidine | 0.8364003666 |
| phenylalanine | proline | 0.8329594867 |
| phenylalanine | threonine | 0.8312016605 |
| phenylalanine | valine | 0.8312016605 |
| phenylalanine | isoleucine | 0.8253854715 |
| phenylalanine | leucine | 0.8253854715 |
| phenylalanine | glycine | 0.7117177608 |
| phenylalanine | serine | 0.7083661715 |
| phenylalanine | alanine | 0.7031853627 |

| | | |
|---|---|---|
| phenylalanine | aspartic_acid | 0.7007555961 |
| phenylalanine | cysteine | 0.699602997 |
| phenylalanine | glutamine | 0.6901259965 |
| phenylalanine | glutamic_acid | 0.6896742537 |
| phenylalanine | asparagine | 0.6858664844 |
| proline | tyrosine | 0.8477021204 |
| proline | phenylalanine | 0.8329594867 |
| proline | histidine | 0.819549672 |
| proline | tryptophan | 0.8138746914 |
| proline | arginine | 0.699602997 |
| proline | glycine | 0.699602997 |
| proline | serine | 0.6963642775 |
| proline | alanine | 0.6913569185 |
| proline | methionine | 0.690092568 |
| proline | aspartic_acid | 0.6890080713 |
| proline | cysteine | 0.687893762 |
| proline | lysine | 0.6858664844 |
| proline | threonine | 0.6792840052 |
| proline | valine | 0.6792840052 |
| proline | glutamine | 0.6787292615 |
| proline | glutamic_acid | 0.6782923109 |
| proline | asparagine | 0.6746088509 |
| proline | isoleucine | 0.6746088509 |
| proline | leucine | 0.6746088509 |
| serine | methionine | 0.887332034 |
| serine | cysteine | 0.8844494812 |
| serine | threonine | 0.8731658784 |
| serine | glutamine | 0.8724390429 |
| serine | asparagine | 0.8670411215 |
| serine | tyrosine | 0.7211557365 |
| serine | arginine | 0.7135227084 |
| serine | glycine | 0.7135227084 |
| serine | phenylalanine | 0.7083661715 |
| serine | alanine | 0.7049472391 |
| serine | aspartic_acid | 0.7025053025 |
| serine | lysine | 0.6992397183 |
| serine | histidine | 0.6967406209 |
| serine | proline | 0.6963642775 |
| serine | valine | 0.6923993223 |
| serine | tryptophan | 0.6918229594 |
| serine | glutamic_acid | 0.691368993 |
| serine | isoleucine | 0.6875425389 |
| serine | leucine | 0.6875425389 |
| threonine | serine | 0.8731658784 |
| threonine | methionine | 0.8653462048 |

| | | |
|---|---|---|
| threonine | cysteine | 0.8626045064 |
| threonine | glutamine | 0.8511761986 |
| threonine | asparagine | 0.8460374094 |
| threonine | arginine | 0.8371004003 |
| threonine | phenylalanine | 0.8312016605 |
| threonine | lysine | 0.8207563456 |
| threonine | histidine | 0.8178949329 |
| threonine | valine | 0.8129230418 |
| threonine | tryptophan | 0.8122628454 |
| threonine | isoleucine | 0.807358987 |
| threonine | leucine | 0.807358987 |
| threonine | tyrosine | 0.7028536384 |
| threonine | glycine | 0.6956011808 |
| threonine | alanine | 0.68744862 |
| threonine | aspartic_acid | 0.6851262096 |
| threonine | proline | 0.6792840052 |
| threonine | glutamic_acid | 0.6745299255 |
| tryptophan | tyrosine | 0.9051653694 |
| tryptophan | phenylalanine | 0.8895257847 |
| tryptophan | arginine | 0.8364003666 |
| tryptophan | methionine | 0.8251037951 |
| tryptophan | lysine | 0.8200833698 |
| tryptophan | histidine | 0.8172266394 |
| tryptophan | proline | 0.8138746914 |
| tryptophan | valine | 0.8122628454 |
| tryptophan | threonine | 0.8122628454 |
| tryptophan | isoleucine | 0.8067077934 |
| tryptophan | leucine | 0.8067077934 |
| tryptophan | glycine | 0.6950194773 |
| tryptophan | serine | 0.6918229594 |
| tryptophan | alanine | 0.6868804664 |
| tryptophan | aspartic_acid | 0.6845618866 |
| tryptophan | cysteine | 0.6834619008 |
| tryptophan | glutamine | 0.6744143308 |
| tryptophan | glutamic_acid | 0.6739829165 |
| tryptophan | asparagine | 0.6703459865 |
| tyrosine | phenylalanine | 0.9271445905 |
| tyrosine | tryptophan | 0.9051653694 |
| tyrosine | histidine | 0.8481672141 |
| tyrosine | proline | 0.8477021204 |
| tyrosine | arginine | 0.7246297412 |
| tyrosine | glycine | 0.7246297412 |
| tyrosine | serine | 0.7211557365 |
| tyrosine | alanine | 0.7157868686 |
| tyrosine | methionine | 0.7144316721 |

| | | |
|---|---|---|
| tyrosine | aspartic_acid | 0.7132693918 |
| tyrosine | cysteine | 0.712075295 |
| tyrosine | lysine | 0.7099032073 |
| tyrosine | valine | 0.7028536384 |
| tyrosine | threonine | 0.7028536384 |
| tyrosine | glutamine | 0.702259747 |
| tyrosine | glutamic_acid | 0.7017919849 |
| tyrosine | asparagine | 0.6978496158 |
| tyrosine | isoleucine | 0.6978496158 |
| tyrosine | leucine | 0.6978496158 |
| valine | isoleucine | 0.9465873914 |
| valine | leucine | 0.9465873914 |
| valine | arginine | 0.8371004003 |
| valine | phenylalanine | 0.8312016605 |
| valine | methionine | 0.8257850392 |
| valine | lysine | 0.8207563456 |
| valine | histidine | 0.8178949329 |
| valine | threonine | 0.8129230418 |
| valine | tryptophan | 0.8122628454 |
| valine | tyrosine | 0.7028536384 |
| valine | glycine | 0.6956011808 |
| valine | serine | 0.6923993223 |
| valine | alanine | 0.68744862 |
| valine | aspartic_acid | 0.6851262096 |
| valine | cysteine | 0.6840244109 |
| valine | proline | 0.6792840052 |
| valine | glutamine | 0.6749620406 |
| valine | glutamic_acid | 0.6745299255 |
| valine | asparagine | 0.6708871056 |

**Table 2. Amino acid similarity measures ranked by word2vec similarity using vectors of length 300.**

| Word2Vec Skip-Gram Vectors | | |
|---|---|---|
| Amino Acid 1 | Amino Acid 2 | Similarity |
| alanine | valine | 0.867003247 |
| alanine | phenylalanine | 0.8568250734 |
| alanine | glutamic_acid | 0.8525401159 |
| alanine | aspartic_acid | 0.848915265 |
| alanine | leucine | 0.8470100396 |
| alanine | isoleucine | 0.8362005956 |
| alanine | threonine | 0.8226329164 |
| alanine | asparagine | 0.8190050493 |
| alanine | arginine | 0.8179479198 |
| alanine | proline | 0.8136103686 |
| alanine | histidine | 0.8116708959 |
| alanine | serine | 0.796112217 |
| alanine | glycine | 0.7656688031 |
| alanine | tyrosine | 0.7550502758 |
| alanine | tryptophan | 0.7512805687 |
| alanine | methionine | 0.7509828744 |
| alanine | cysteine | 0.717941776 |
| alanine | lysine | 0.6932988688 |
| alanine | glutamine | 0.6806736532 |
| arginine | proline | 0.8217219769 |
| arginine | alanine | 0.8179479198 |
| arginine | phenylalanine | 0.811326929 |
| arginine | leucine | 0.8074111753 |
| arginine | glutamic_acid | 0.8057159534 |
| arginine | histidine | 0.8047430715 |
| arginine | isoleucine | 0.7918742532 |
| arginine | valine | 0.7843296917 |
| arginine | asparagine | 0.782311913 |
| arginine | tryptophan | 0.7795826129 |
| arginine | aspartic_acid | 0.7785523347 |
| arginine | methionine | 0.762185611 |
| arginine | threonine | 0.7471197236 |
| arginine | cysteine | 0.7465933775 |
| arginine | glycine | 0.7445819752 |
| arginine | lysine | 0.7443292082 |
| arginine | serine | 0.7178882726 |
| arginine | tyrosine | 0.7072041249 |
| arginine | glutamine | 0.681033236 |

| | | |
|---|---|---|
| asparagine | glutamic_acid | 0.8394453512 |
| asparagine | isoleucine | 0.837647272 |
| asparagine | aspartic_acid | 0.8289475088 |
| asparagine | phenylalanine | 0.8208708431 |
| asparagine | alanine | 0.8190050493 |
| asparagine | proline | 0.8170773228 |
| asparagine | histidine | 0.8158799002 |
| asparagine | valine | 0.8154084924 |
| asparagine | leucine | 0.8034815641 |
| asparagine | arginine | 0.782311913 |
| asparagine | threonine | 0.771159308 |
| asparagine | methionine | 0.7486257278 |
| asparagine | tryptophan | 0.7365658364 |
| asparagine | glycine | 0.7256812288 |
| asparagine | cysteine | 0.7249359209 |
| asparagine | glutamine | 0.7171100115 |
| asparagine | serine | 0.7110291605 |
| asparagine | tyrosine | 0.6866899578 |
| asparagine | lysine | 0.6701038321 |
| aspartic_acid | glutamic_acid | 0.8713766559 |
| aspartic_acid | valine | 0.8591524705 |
| aspartic_acid | alanine | 0.848915265 |
| aspartic_acid | asparagine | 0.8289475088 |
| aspartic_acid | phenylalanine | 0.8239387834 |
| aspartic_acid | isoleucine | 0.8209012515 |
| aspartic_acid | histidine | 0.8116528278 |
| aspartic_acid | threonine | 0.8021783033 |
| aspartic_acid | leucine | 0.7971397071 |
| aspartic_acid | proline | 0.7882071883 |
| aspartic_acid | glycine | 0.7815732041 |
| aspartic_acid | arginine | 0.7785523347 |
| aspartic_acid | serine | 0.7722169188 |
| aspartic_acid | tyrosine | 0.7453165661 |
| aspartic_acid | cysteine | 0.7231820334 |
| aspartic_acid | tryptophan | 0.7160212406 |
| aspartic_acid | methionine | 0.7124990562 |
| aspartic_acid | lysine | 0.6925704748 |
| aspartic_acid | glutamine | 0.6288964401 |
| cysteine | methionine | 0.8100333705 |
| cysteine | histidine | 0.8052759711 |
| cysteine | proline | 0.7627465635 |
| cysteine | glutamic_acid | 0.7520330764 |
| cysteine | tyrosine | 0.7483064414 |
| cysteine | arginine | 0.7465933775 |
| cysteine | glycine | 0.7308321281 |

| | | |
|---|---|---|
| cysteine | asparagine | 0.7249359209 |
| cysteine | aspartic_acid | 0.7231820334 |
| cysteine | alanine | 0.717941776 |
| cysteine | phenylalanine | 0.7136623066 |
| cysteine | leucine | 0.7124007957 |
| cysteine | valine | 0.7067593484 |
| cysteine | tryptophan | 0.704046007 |
| cysteine | serine | 0.698381192 |
| cysteine | isoleucine | 0.6933520752 |
| cysteine | lysine | 0.6783799583 |
| cysteine | threonine | 0.6690936284 |
| cysteine | glutamine | 0.6293909002 |
| glutamic_acid | aspartic_acid | 0.8713766559 |
| glutamic_acid | alanine | 0.8525401159 |
| glutamic_acid | valine | 0.8515985213 |
| glutamic_acid | asparagine | 0.8394453512 |
| glutamic_acid | proline | 0.8301983365 |
| glutamic_acid | phenylalanine | 0.8285873986 |
| glutamic_acid | isoleucine | 0.8247062108 |
| glutamic_acid | leucine | 0.8214863197 |
| glutamic_acid | arginine | 0.8057159534 |
| glutamic_acid | histidine | 0.8049838232 |
| glutamic_acid | threonine | 0.7810750959 |
| glutamic_acid | glycine | 0.7659421849 |
| glutamic_acid | serine | 0.7647822409 |
| glutamic_acid | tryptophan | 0.7559569626 |
| glutamic_acid | cysteine | 0.7520330764 |
| glutamic_acid | tyrosine | 0.7425242751 |
| glutamic_acid | methionine | 0.7338943892 |
| glutamic_acid | lysine | 0.7033134431 |
| glutamic_acid | glutamine | 0.6925645526 |
| glutamine | leucine | 0.7389217756 |
| glutamine | asparagine | 0.7171100115 |
| glutamine | glutamic_acid | 0.6925645526 |
| glutamine | glycine | 0.6844299456 |
| glutamine | proline | 0.6814647567 |
| glutamine | arginine | 0.681033236 |
| glutamine | alanine | 0.6806736532 |
| glutamine | methionine | 0.6671780258 |
| glutamine | histidine | 0.6641863967 |
| glutamine | isoleucine | 0.6474885276 |
| glutamine | valine | 0.6413913227 |
| glutamine | phenylalanine | 0.6370580788 |
| glutamine | cysteine | 0.6293909002 |
| glutamine | aspartic_acid | 0.6288964401 |

| | | |
|---|---|---|
| glutamine | tryptophan | 0.6176974703 |
| glutamine | threonine | 0.6145120833 |
| glutamine | tyrosine | 0.5877369796 |
| glutamine | lysine | 0.5797500116 |
| glutamine | serine | 0.5694687277 |
| glycine | aspartic_acid | 0.7815732041 |
| glycine | glutamic_acid | 0.7659421849 |
| glycine | alanine | 0.7656688031 |
| glycine | histidine | 0.7525963938 |
| glycine | arginine | 0.7445819752 |
| glycine | cysteine | 0.7308321281 |
| glycine | leucine | 0.7269395245 |
| glycine | proline | 0.726168955 |
| glycine | asparagine | 0.7256812288 |
| glycine | valine | 0.7237581323 |
| glycine | phenylalanine | 0.7223605625 |
| glycine | isoleucine | 0.7026985836 |
| glycine | methionine | 0.6945388429 |
| glycine | tryptophan | 0.6870523764 |
| glycine | glutamine | 0.6844299456 |
| glycine | serine | 0.6775571972 |
| glycine | threonine | 0.6614020648 |
| glycine | tyrosine | 0.6525265613 |
| glycine | lysine | 0.6464542484 |
| histidine | asparagine | 0.8158799002 |
| histidine | alanine | 0.8116708959 |
| histidine | aspartic_acid | 0.8116528278 |
| histidine | phenylalanine | 0.8109908222 |
| histidine | proline | 0.8096903522 |
| histidine | valine | 0.80910133 |
| histidine | cysteine | 0.8052759711 |
| histidine | glutamic_acid | 0.8049838232 |
| histidine | arginine | 0.8047430715 |
| histidine | isoleucine | 0.8005763614 |
| histidine | leucine | 0.7923991983 |
| histidine | tryptophan | 0.76889797 |
| histidine | threonine | 0.7561487142 |
| histidine | methionine | 0.7559796124 |
| histidine | tyrosine | 0.7534435318 |
| histidine | glycine | 0.7525963938 |
| histidine | serine | 0.7247864265 |
| histidine | lysine | 0.6996056871 |
| histidine | glutamine | 0.6641863967 |
| isoleucine | valine | 0.8963862845 |
| isoleucine | phenylalanine | 0.8475390048 |

| | | |
|---|---|---|
| isoleucine | leucine | 0.8452388878 |
| isoleucine | asparagine | 0.837647272 |
| isoleucine | alanine | 0.8362005956 |
| isoleucine | glutamic_acid | 0.8247062108 |
| isoleucine | proline | 0.8234492803 |
| isoleucine | aspartic_acid | 0.8209012515 |
| isoleucine | threonine | 0.8129920032 |
| isoleucine | histidine | 0.8005763614 |
| isoleucine | arginine | 0.7918742532 |
| isoleucine | methionine | 0.7696502343 |
| isoleucine | tryptophan | 0.7552613583 |
| isoleucine | tyrosine | 0.718026413 |
| isoleucine | serine | 0.7167155685 |
| isoleucine | glycine | 0.7026985836 |
| isoleucine | cysteine | 0.6933520752 |
| isoleucine | glutamine | 0.6474885276 |
| isoleucine | lysine | 0.6393536906 |
| leucine | valine | 0.8513632074 |
| leucine | alanine | 0.8470100396 |
| leucine | isoleucine | 0.8452388878 |
| leucine | phenylalanine | 0.8281130772 |
| leucine | proline | 0.8228100233 |
| leucine | glutamic_acid | 0.8214863197 |
| leucine | arginine | 0.8074111753 |
| leucine | threonine | 0.8037454137 |
| leucine | asparagine | 0.8034815641 |
| leucine | aspartic_acid | 0.7971397071 |
| leucine | histidine | 0.7923991983 |
| leucine | methionine | 0.7637636845 |
| leucine | tryptophan | 0.7615176113 |
| leucine | serine | 0.7468755877 |
| leucine | glutamine | 0.7389217756 |
| leucine | tyrosine | 0.7295778702 |
| leucine | glycine | 0.7269395245 |
| leucine | cysteine | 0.7124007957 |
| leucine | lysine | 0.6518053363 |
| lysine | arginine | 0.7443292082 |
| lysine | glutamic_acid | 0.7033134431 |
| lysine | histidine | 0.6996056871 |
| lysine | alanine | 0.6932988688 |
| lysine | aspartic_acid | 0.6925704748 |
| lysine | proline | 0.6814074204 |
| lysine | serine | 0.6813237391 |
| lysine | cysteine | 0.6783799583 |
| lysine | asparagine | 0.6701038321 |

| | | |
|---|---|---|
| lysine | methionine | 0.6674088757 |
| lysine | valine | 0.6673905375 |
| lysine | phenylalanine | 0.6636672651 |
| lysine | threonine | 0.66183477 |
| lysine | leucine | 0.6518053363 |
| lysine | glycine | 0.6464542484 |
| lysine | isoleucine | 0.6393536906 |
| lysine | tyrosine | 0.6329576495 |
| lysine | tryptophan | 0.6054281184 |
| lysine | glutamine | 0.5797500116 |
| methionine | cysteine | 0.8100333705 |
| methionine | isoleucine | 0.7696502343 |
| methionine | leucine | 0.7637636845 |
| methionine | arginine | 0.762185611 |
| methionine | valine | 0.7601785274 |
| methionine | histidine | 0.7559796124 |
| methionine | phenylalanine | 0.7555737465 |
| methionine | alanine | 0.7509828744 |
| methionine | asparagine | 0.7486257278 |
| methionine | tryptophan | 0.7483683151 |
| methionine | proline | 0.744774991 |
| methionine | glutamic_acid | 0.7338943892 |
| methionine | aspartic_acid | 0.7124990562 |
| methionine | tyrosine | 0.7081127726 |
| methionine | threonine | 0.7051660785 |
| methionine | glycine | 0.6945388429 |
| methionine | serine | 0.6857637641 |
| methionine | lysine | 0.6674088757 |
| methionine | glutamine | 0.6671780258 |
| phenylalanine | alanine | 0.8568250734 |
| phenylalanine | valine | 0.8565129039 |
| phenylalanine | isoleucine | 0.8475390048 |
| phenylalanine | tryptophan | 0.8415400774 |
| phenylalanine | proline | 0.8361830321 |
| phenylalanine | glutamic_acid | 0.8285873986 |
| phenylalanine | leucine | 0.8281130772 |
| phenylalanine | aspartic_acid | 0.8239387834 |
| phenylalanine | asparagine | 0.8208708431 |
| phenylalanine | arginine | 0.811326929 |
| phenylalanine | histidine | 0.8109908222 |
| phenylalanine | tyrosine | 0.7952590901 |
| phenylalanine | threonine | 0.7917952955 |
| phenylalanine | methionine | 0.7555737465 |
| phenylalanine | serine | 0.7350704826 |
| phenylalanine | glycine | 0.7223605625 |

| | | |
|---|---|---|
| phenylalanine | cysteine | 0.7136623066 |
| phenylalanine | lysine | 0.6636672651 |
| phenylalanine | glutamine | 0.6370580788 |
| proline | phenylalanine | 0.8361830321 |
| proline | glutamic_acid | 0.8301983365 |
| proline | valine | 0.82821438 |
| proline | isoleucine | 0.8234492803 |
| proline | leucine | 0.8228100233 |
| proline | arginine | 0.8217219769 |
| proline | asparagine | 0.8170773228 |
| proline | alanine | 0.8136103686 |
| proline | histidine | 0.8096903522 |
| proline | aspartic_acid | 0.7882071883 |
| proline | tryptophan | 0.7718880326 |
| proline | cysteine | 0.7627465635 |
| proline | threonine | 0.7536677004 |
| proline | methionine | 0.744774991 |
| proline | tyrosine | 0.73456413 |
| proline | glycine | 0.726168955 |
| proline | serine | 0.7249356355 |
| proline | glutamine | 0.6814647567 |
| proline | lysine | 0.6814074204 |
| serine | threonine | 0.8883705124 |
| serine | tyrosine | 0.7965899807 |
| serine | alanine | 0.796112217 |
| serine | aspartic_acid | 0.7722169188 |
| serine | glutamic_acid | 0.7647822409 |
| serine | valine | 0.7483581618 |
| serine | leucine | 0.7468755877 |
| serine | phenylalanine | 0.7350704826 |
| serine | proline | 0.7249356355 |
| serine | histidine | 0.7247864265 |
| serine | arginine | 0.7178882726 |
| serine | isoleucine | 0.7167155685 |
| serine | asparagine | 0.7110291605 |
| serine | cysteine | 0.698381192 |
| serine | methionine | 0.6857637641 |
| serine | lysine | 0.6813237391 |
| serine | glycine | 0.6775571972 |
| serine | tryptophan | 0.6537660092 |
| serine | glutamine | 0.5694687277 |
| threonine | serine | 0.8883705124 |
| threonine | alanine | 0.8226329164 |
| threonine | valine | 0.8156444032 |
| threonine | isoleucine | 0.8129920032 |

| | | |
|---|---|---|
| threonine | leucine | 0.8037454137 |
| threonine | aspartic_acid | 0.8021783033 |
| threonine | tyrosine | 0.7988339533 |
| threonine | phenylalanine | 0.7917952955 |
| threonine | glutamic_acid | 0.7810750959 |
| threonine | asparagine | 0.771159308 |
| threonine | histidine | 0.7561487142 |
| threonine | proline | 0.7536677004 |
| threonine | arginine | 0.7471197236 |
| threonine | methionine | 0.7051660785 |
| threonine | cysteine | 0.6690936284 |
| threonine | tryptophan | 0.6687310677 |
| threonine | lysine | 0.66183477 |
| threonine | glycine | 0.6614020648 |
| threonine | glutamine | 0.6145120833 |
| tryptophan | phenylalanine | 0.8415400774 |
| tryptophan | arginine | 0.7795826129 |
| tryptophan | proline | 0.7718880326 |
| tryptophan | histidine | 0.76889797 |
| tryptophan | tyrosine | 0.7687626032 |
| tryptophan | leucine | 0.7615176113 |
| tryptophan | valine | 0.756779598 |
| tryptophan | glutamic_acid | 0.7559569626 |
| tryptophan | isoleucine | 0.7552613583 |
| tryptophan | alanine | 0.7512805687 |
| tryptophan | methionine | 0.7483683151 |
| tryptophan | asparagine | 0.7365658364 |
| tryptophan | aspartic_acid | 0.7160212406 |
| tryptophan | cysteine | 0.704046007 |
| tryptophan | glycine | 0.6870523764 |
| tryptophan | threonine | 0.6687310677 |
| tryptophan | serine | 0.6537660092 |
| tryptophan | glutamine | 0.6176974703 |
| tryptophan | lysine | 0.6054281184 |
| tyrosine | threonine | 0.7988339533 |
| tyrosine | serine | 0.7965899807 |
| tyrosine | phenylalanine | 0.7952590901 |
| tyrosine | tryptophan | 0.7687626032 |
| tyrosine | alanine | 0.7550502758 |
| tyrosine | histidine | 0.7534435318 |
| tyrosine | cysteine | 0.7483064414 |
| tyrosine | aspartic_acid | 0.7453165661 |
| tyrosine | valine | 0.7451906472 |
| tyrosine | glutamic_acid | 0.7425242751 |
| tyrosine | proline | 0.73456413 |

| | | |
|---|---|---|
| tyrosine | leucine | 0.7295778702 |
| tyrosine | isoleucine | 0.718026413 |
| tyrosine | methionine | 0.7081127726 |
| tyrosine | arginine | 0.7072041249 |
| tyrosine | asparagine | 0.6866899578 |
| tyrosine | glycine | 0.6525265613 |
| tyrosine | lysine | 0.6329576495 |
| tyrosine | glutamine | 0.5877369796 |
| valine | isoleucine | 0.8963862845 |
| valine | alanine | 0.867003247 |
| valine | aspartic_acid | 0.8591524705 |
| valine | phenylalanine | 0.8565129039 |
| valine | glutamic_acid | 0.8515985213 |
| valine | leucine | 0.8513632074 |
| valine | proline | 0.82821438 |
| valine | threonine | 0.8156444032 |
| valine | asparagine | 0.8154084924 |
| valine | histidine | 0.80910133 |
| valine | arginine | 0.7843296917 |
| valine | methionine | 0.7601785274 |
| valine | tryptophan | 0.756779598 |
| valine | serine | 0.7483581618 |
| valine | tyrosine | 0.7451906472 |
| valine | glycine | 0.7237581323 |
| valine | cysteine | 0.7067593484 |
| valine | lysine | 0.6673905375 |
| valine | glutamine | 0.6413913227 |

**Table 3. MESH and word2vec similarity scores for pathogenic organisms.**

| Word2Vec 300 | | | MESH Sanchez-Lin Similarity | | |
|---|---|---|---|---|---|
| | | | | | |
| | | | | | |
| Bacteria 1 | Bacteria 2 | Similarity | Bacteria 1 | Bacteria 2 | Simialrity |
| bacillus_anthracis | yersinia_pestis | 0.758683807 | bacillus_anthracis | listeria_monocytogenes | 0.8109947456 |
| bacillus_anthracis | burkholderia_pseudomallei | 0.694882182 | bacillus_anthracis | clostridium_botulinum | 0.803816101 |
| bacillus_anthracis | francisella_tularensis | 0.682484291 | bacillus_anthracis | clostridium_perfringens | 0.7821019298 |
| bacillus_anthracis | listeria_monocytogenes | 0.677570833 | bacillus_anthracis | salmonella | 0.4886368013 |
| bacillus_anthracis | clostridium_perfringens | 0.669559982 | bacillus_anthracis | brucella | 0.484321295 |
| bacillus_anthracis | campylobacter_jejuni | 0.630192147 | bacillus_anthracis | shigella | 0.4833730016 |
| bacillus_anthracis | clostridium_botulinum | 0.615869355 | bacillus_anthracis | burkholderia_mallei | 0.4743239831 |
| bacillus_anthracis | burkholderia_mallei | 0.610976279 | bacillus_anthracis | burkholderia_pseudomallei | 0.4743239831 |
| bacillus_anthracis | shigella | 0.56467806 | bacillus_anthracis | campylobacter_jejuni | 0.4743239831 |
| bacillus_anthracis | rickettsia_prowazekii | 0.543889904 | bacillus_anthracis | chlamydia_psittaci | 0.4743239831 |
| bacillus_anthracis | salmonella | 0.536186579 | bacillus_anthracis | coxiella_burnetii | 0.4743239831 |
| bacillus_anthracis | coxiella_burnetii | 0.506694011 | bacillus_anthracis | francisella_tularensis | 0.4743239831 |
| bacillus_anthracis | brucella | 0.449096122 | bacillus_anthracis | rickettsia_prowazekii | 0.4743239831 |
| bacillus_anthracis | entamoeba_histolytica | 0.40738203 | bacillus_anthracis | yersinia_pestis | 0.4743239831 |
| bacillus_anthracis | chlamydia_psittaci | 0.378345303 | bacillus_anthracis | entamoeba_histolytica | 0.2703817266 |
| bacillus_anthracis | giardia_lamblia | 0.3584193 | bacillus_anthracis | giardia_lamblia | 0.2703817266 |
| bacillus_anthracis | toxoplasma_gondii | 0.357918893 | bacillus_anthracis | toxoplasma_gondii | 0.2703817266 |
| | | | | | |
| | | | | | |
| brucella | coxiella_burnetii | 0.652046325 | brucella | rickettsia_prowazekii | 0.7482048658 |
| brucella | salmonella | 0.557399382 | brucella | burkholderia_mallei | 0.6981551818 |
| brucella | chlamydia_psittaci | 0.551430199 | brucella | francisella_tularensis | 0.6981551818 |
| brucella | shigella | 0.550852083 | brucella | burkholderia_pseudo | 0.6981551818 |

| | | | | | |
|---|---|---|---|---|---|
| | | | | mallei | |
| brucella | listeria_monocytogenes | 0.547157641 | brucella | coxiella_burnetii | 0.6981551818 |
| brucella | francisella_tularensis | 0.533713552 | brucella | salmonella | 0.6007489119 |
| brucella | campylobacter_jejuni | 0.526233704 | brucella | shigella | 0.5941383356 |
| brucella | burkholderia_pseudomallei | 0.512132989 | brucella | campylobacter_jejuni | 0.5827812959 |
| brucella | yersinia_pestis | 0.474110716 | brucella | yersinia_pestis | 0.5827812959 |
| brucella | burkholderia_mallei | 0.469147095 | brucella | chlamydia_psittaci | 0.552394163 |
| brucella | clostridium_perfringens | 0.451108627 | brucella | clostridium_botulinum | 0.4980593787 |
| brucella | bacillus_anthracis | 0.449096122 | brucella | clostridium_perfringens | 0.484321295 |
| brucella | toxoplasma_gondii | 0.445962129 | brucella | listeria_monocytogenes | 0.484321295 |
| brucella | entamoeba_histolytica | 0.421155891 | brucella | bacillus_anthracis | 0.484321295 |
| brucella | rickettsia_prowazekii | 0.396640513 | brucella | giardia_lamblia | 0.2760805538 |
| brucella | giardia_lamblia | 0.360897599 | brucella | entamoeba_histolytica | 0.2760805538 |
| brucella | clostridium_botulinum | 0.302851924 | brucella | toxoplasma_gondii | 0.2760805538 |
| | | | | | |
| | | | | | |
| burkholderia_mallei | burkholderia_pseudomallei | 0.693596881 | burkholderia_mallei | burkholderia_pseudomallei | 0.9587161995 |
| burkholderia_mallei | yersinia_pestis | 0.658958001 | burkholderia_mallei | brucella | 0.6981551818 |
| burkholderia_mallei | francisella_tularensis | 0.656825591 | burkholderia_mallei | coxiella_burnetii | 0.6837439321 |
| burkholderia_mallei | bacillus_anthracis | 0.610976279 | burkholderia_mallei | francisella_tularensis | 0.6837439321 |
| burkholderia_mallei | rickettsia_prowazekii | 0.603512545 | burkholderia_mallei | salmonella | 0.5879741223 |
| burkholderia_mallei | coxiella_burnetii | 0.562032006 | burkholderia_mallei | shigella | 0.5816402194 |
| burkholderia_mallei | listeria_monocytogenes | 0.538628358 | burkholderia_mallei | campylobacter_jejuni | 0.5707515826 |
| burkholderia_mallei | campylobacter_jejuni | 0.527633613 | burkholderia_mallei | rickettsia_prowazekii | 0.5707515826 |
| burkholderia_mallei | clostridium_botulinum | 0.489437276 | burkholderia_mallei | yersinia_pestis | 0.5707515826 |
| burkholderia_mallei | clostridium_perfringens | 0.479844168 | burkholderia_mallei | chlamydia_psittaci | 0.5409916978 |
| burkholderia_mallei | brucella | 0.469147095 | burkholderia_mallei | clostridium_botulinum | 0.4874930494 |
| burkholderia_mallei | chlamydia_psittaci | 0.446314251 | burkholderia_mallei | bacillus_anthracis | 0.4743239831 |
| burkholderia_mallei | shigella | 0.423452341 | burkholderia_mallei | clostridium_perfringe | 0.4743239831 |

| | | | | | |
|---|---|---|---|---|---|
| | | | | ns | |
| burkholderia_mallei | toxoplasma_gondii | 0.415286353 | burkholderia_mallei | listeria_monocytogenes | 0.4743239831 |
| burkholderia_mallei | entamoeba_histolytica | 0.410001031 | burkholderia_mallei | entamoeba_histolytica | 0.2703817266 |
| burkholderia_mallei | salmonella | 0.391549924 | burkholderia_mallei | giardia_lamblia | 0.2703817266 |
| burkholderia_mallei | giardia_lamblia | 0.37962294 | burkholderia_mallei | toxoplasma_gondii | 0.2703817266 |
| | | | | | |
| | | | | | |
| burkholderia_pseudomallei | francisella_tularensis | 0.755478426 | burkholderia_pseudomallei | burkholderia_mallei | 0.9587161995 |
| burkholderia_pseudomallei | yersinia_pestis | 0.719093848 | burkholderia_pseudomallei | brucella | 0.6981551818 |
| burkholderia_pseudomallei | bacillus_anthracis | 0.694882182 | burkholderia_pseudomallei | coxiella_burnetii | 0.6837439321 |
| burkholderia_pseudomallei | burkholderia_mallei | 0.693596881 | burkholderia_pseudomallei | francisella_tularensis | 0.6837439321 |
| burkholderia_pseudomallei | listeria_monocytogenes | 0.61615112 | burkholderia_pseudomallei | salmonella | 0.5879741223 |
| burkholderia_pseudomallei | coxiella_burnetii | 0.585284131 | burkholderia_pseudomallei | shigella | 0.5816402194 |
| burkholderia_pseudomallei | campylobacter_jejuni | 0.563167255 | burkholderia_pseudomallei | rickettsia_prowazekii | 0.5707515826 |
| burkholderia_pseudomallei | clostridium_perfringens | 0.545520595 | burkholderia_pseudomallei | campylobacter_jejuni | 0.5707515826 |
| burkholderia_pseudomallei | rickettsia_prowazekii | 0.525044474 | burkholderia_pseudomallei | yersinia_pestis | 0.5707515826 |
| burkholderia_pseudomallei | shigella | 0.52229287 | burkholderia_pseudomallei | rickettsia_prowazekii | 0.5707515826 |
| burkholderia_pseudomallei | brucella | 0.512132989 | burkholderia_pseudomallei | chlamydia_psittaci | 0.5409916978 |
| burkholderia_pseudomallei | clostridium_botulinum | 0.468072435 | burkholderia_pseudomallei | clostridium_botulinum | 0.4874930494 |
| burkholderia_pseudomallei | entamoeba_histolytica | 0.461405609 | burkholderia_pseudomallei | clostridium_perfringens | 0.4743239831 |
| burkholderia_pseudomallei | toxoplasma_gondii | 0.448829873 | burkholderia_pseudomallei | listeria_monocytogenes | 0.4743239831 |
| burkholderia_pseudo | chlamydia_psittaci | 0.430852558 | burkholderia_pseudomallei | bacillus_anthracis | 0.4743239831 |

| | | | | | |
|---|---|---|---|---|---|
| mallei | | | | | |
| burkholderia_pseudomallei | giardia_lamblia | 0.428295047 | burkholderia_pseudomallei | giardia_lamblia | 0.2703817266 |
| burkholderia_pseudomallei | salmonella | 0.41936872 | burkholderia_pseudomallei | entamoeba_histolytica | 0.2703817266 |
| burkholderia_pseudomallei | staphylococcus_enterotoxin | 0.269030061 | burkholderia_pseudomallei | toxoplasma_gondii | 0.2703817266 |
| | | | | | |
| | | | | | |
| campylobacter_jejuni | salmonella | 0.658679635 | campylobacter_jejuni | salmonella | 0.5879741223 |
| campylobacter_jejuni | clostridium_perfringens | 0.646306096 | campylobacter_jejuni | brucella | 0.5827812959 |
| campylobacter_jejuni | listeria_monocytogenes | 0.635935657 | campylobacter_jejuni | shigella | 0.5816402194 |
| campylobacter_jejuni | bacillus_anthracis | 0.630192147 | campylobacter_jejuni | yersinia_pestis | 0.5707515826 |
| campylobacter_jejuni | shigella | 0.628804943 | campylobacter_jejuni | burkholderia_mallei | 0.5707515826 |
| campylobacter_jejuni | yersinia_pestis | 0.625828469 | campylobacter_jejuni | rickettsia_prowazekii | 0.5707515826 |
| campylobacter_jejuni | francisella_tularensis | 0.617449011 | campylobacter_jejuni | francisella_tularensis | 0.5707515826 |
| campylobacter_jejuni | burkholderia_pseudomallei | 0.563167255 | campylobacter_jejuni | burkholderia_pseudomallei | 0.5707515826 |
| campylobacter_jejuni | coxiella_burnetii | 0.537468792 | campylobacter_jejuni | coxiella_burnetii | 0.5707515826 |
| campylobacter_jejuni | burkholderia_mallei | 0.527633613 | campylobacter_jejuni | burkholderia_mallei | 0.5707515826 |
| campylobacter_jejuni | burkholderia_mallei | 0.527633613 | campylobacter_jejuni | chlamydia_psittaci | 0.5409916978 |
| campylobacter_jejuni | brucella | 0.526233704 | campylobacter_jejuni | clostridium_botulinum | 0.4874930494 |
| campylobacter_jejuni | chlamydia_psittaci | 0.494488507 | campylobacter_jejuni | bacillus_anthracis | 0.4743239831 |
| campylobacter_jejuni | clostridium_botulinum | 0.489194568 | campylobacter_jejuni | clostridium_perfringens | 0.4743239831 |
| campylobacter_jejuni | entamoeba_histolytica | 0.474002835 | campylobacter_jejuni | listeria_monocytogenes | 0.4743239831 |
| campylobacter_jejuni | giardia_lamblia | 0.447531708 | campylobacter_jejuni | giardia_lamblia | 0.2703817266 |
| campylobacter_jejuni | rickettsia_prowazekii | 0.44029191 | campylobacter_jejuni | entamoeba_histolytica | 0.2703817266 |
| campylobacter_jejuni | toxoplasma_gondii | 0.368116237 | campylobacter_jejuni | toxoplasma_gondii | 0.2703817266 |
| | | | | | |
| | | | | | |
| chlamydia_psittaci | coxiella_burnetii | 0.669150293 | chlamydia_psittaci | salmonella | 0.5573162272 |
| chlamydia_psittaci | francisella_tularensis | 0.560802847 | chlamydia_psittaci | brucella | 0.552394163 |

| | | | | | |
|---|---|---|---|---|---|
| chlamydia_psittaci | brucella | 0.551430199 | chlamydia_psittaci | shigella | 0.551312584 |
| chlamydia_psittaci | campylobacter_jejuni | 0.494488507 | chlamydia_psittaci | rickettsia_prowazekii | 0.5409916978 |
| chlamydia_psittaci | rickettsia_prowazekii | 0.469599243 | chlamydia_psittaci | campylobacter_jejuni | 0.5409916978 |
| chlamydia_psittaci | yersinia_pestis | 0.459664546 | chlamydia_psittaci | yersinia_pestis | 0.5409916978 |
| chlamydia_psittaci | toxoplasma_gondii | 0.458182392 | chlamydia_psittaci | burkholderia_mallei | 0.5409916978 |
| chlamydia_psittaci | burkholderia_mallei | 0.446314251 | chlamydia_psittaci | francisella_tularensis | 0.5409916978 |
| chlamydia_psittaci | burkholderia_mallei | 0.446314251 | chlamydia_psittaci | burkholderia_pseudomallei | 0.5409916978 |
| chlamydia_psittaci | burkholderia_pseudomallei | 0.430852558 | chlamydia_psittaci | coxiella_burnetii | 0.5409916978 |
| chlamydia_psittaci | entamoeba_histolytica | 0.401600245 | chlamydia_psittaci | burkholderia_mallei | 0.5409916978 |
| chlamydia_psittaci | clostridium_perfringens | 0.400515749 | chlamydia_psittaci | clostridium_botulinum | 0.4874930494 |
| chlamydia_psittaci | listeria_monocytogenes | 0.386162062 | chlamydia_psittaci | clostridium_perfringens | 0.4743239831 |
| chlamydia_psittaci | bacillus_anthracis | 0.378345303 | chlamydia_psittaci | listeria_monocytogenes | 0.4743239831 |
| chlamydia_psittaci | shigella | 0.36050759 | chlamydia_psittaci | bacillus_anthracis | 0.4743239831 |
| chlamydia_psittaci | giardia_lamblia | 0.356361647 | chlamydia_psittaci | giardia_lamblia | 0.2703817266 |
| chlamydia_psittaci | salmonella | 0.306779799 | chlamydia_psittaci | entamoeba_histolytica | 0.2703817266 |
| chlamydia_psittaci | clostridium_botulinum | 0.299672633 | chlamydia_psittaci | toxoplasma_gondii | 0.2703817266 |
| | | | | | |
| | | | | | |
| clostridium_botulinum | clostridium_perfringens | 0.635110685 | clostridium_botulinum | clostridium_perfringens | 0.8826947411 |
| clostridium_botulinum | bacillus_anthracis | 0.615869355 | clostridium_botulinum | bacillus_anthracis | 0.803816101 |
| clostridium_botulinum | yersinia_pestis | 0.493559377 | clostridium_botulinum | listeria_monocytogenes | 0.6951956203 |
| clostridium_botulinum | listeria_monocytogenes | 0.490690845 | clostridium_botulinum | salmonella | 0.5026243353 |
| clostridium_botulinum | burkholderia_mallei | 0.489437276 | clostridium_botulinum | brucella | 0.4980593787 |
| clostridium_botulinum | campylobacter_jejuni | 0.489194568 | clostridium_botulinum | shigella | 0.4970565801 |
| clostridium_botulinu | burkholderia_pseudom | 0.468072435 | clostridium_botulinum | francisella_tularensis | 0.4874930494 |

| | | | | | |
|---|---|---|---|---|---|
| m | allei | | | | |
| clostridium_botulinum | francisella_tularensis | 0.452875066 | clostridium_botulinum | burkholderia_pseudomallei | 0.4874930494 |
| clostridium_botulinum | rickettsia_prowazekii | 0.404366669 | clostridium_botulinum | coxiella_burnetii | 0.4874930494 |
| clostridium_botulinum | rickettsia_prowazekii | 0.404366669 | clostridium_botulinum | burkholderia_mallei | 0.4874930494 |
| clostridium_botulinum | shigella | 0.365614364 | clostridium_botulinum | chlamydia_psittaci | 0.4874930494 |
| clostridium_botulinum | salmonella | 0.358533515 | clostridium_botulinum | rickettsia_prowazekii | 0.4874930494 |
| clostridium_botulinum | coxiella_burnetii | 0.352984904 | clostridium_botulinum | campylobacter_jejuni | 0.4874930494 |
| clostridium_botulinum | brucella | 0.302851924 | clostridium_botulinum | yersinia_pestis | 0.4874930494 |
| clostridium_botulinum | chlamydia_psittaci | 0.299672633 | clostridium_botulinum | burkholderia_mallei | 0.4874930494 |
| clostridium_botulinum | giardia_lamblia | 0.297062759 | clostridium_botulinum | giardia_lamblia | 0.2778885679 |
| clostridium_botulinum | entamoeba_histolytica | 0.272100265 | clostridium_botulinum | entamoeba_histolytica | 0.2778885679 |
| clostridium_botulinum | toxoplasma_gondii | 0.203076924 | clostridium_botulinum | toxoplasma_gondii | 0.2778885679 |
| | | | | | |
| | | | | | |
| clostridium_perfringens | bacillus_anthracis | 0.669559982 | clostridium_perfringens | bacillus_anthracis | 0.7821019298 |
| clostridium_perfringens | toxoplasma_gondii | 0.262706126 | clostridium_perfringens | clostridium_botulinum | 0.8826947411 |
| clostridium_perfringens | giardia_lamblia | 0.347200657 | clostridium_perfringens | listeria_monocytogenes | 0.676415707 |
| clostridium_perfringens | rickettsia_prowazekii | 0.359320385 | clostridium_perfringens | salmonella | 0.4886368013 |
| clostridium_perfringens | entamoeba_histolytica | 0.393354502 | clostridium_perfringens | brucella | 0.484321295 |
| clostridium_perfringens | chlamydia_psittaci | 0.400515749 | clostridium_perfringens | shigella | 0.4833730016 |

| | | | | | |
|---|---|---|---|---|---|
| clostridium_perfringens | coxiella_burnetii | 0.415534492 | clostridium_perfringens | burkholderia_mallei | 0.4743239831 |
| clostridium_perfringens | brucella | 0.451108627 | clostridium_perfringens | burkholderia_pseudomallei | 0.4743239831 |
| clostridium_perfringens | burkholderia_mallei | 0.479844168 | clostridium_perfringens | campylobacter_jejuni | 0.4743239831 |
| clostridium_perfringens | francisella_tularensis | 0.515747131 | clostridium_perfringens | chlamydia_psittaci | 0.4743239831 |
| clostridium_perfringens | shigella | 0.518906715 | clostridium_perfringens | coxiella_burnetii | 0.4743239831 |
| clostridium_perfringens | yersinia_pestis | 0.521380081 | clostridium_perfringens | francisella_tularensis | 0.4743239831 |
| clostridium_perfringens | burkholderia_pseudomallei | 0.545520595 | clostridium_perfringens | rickettsia_prowazekii | 0.4743239831 |
| clostridium_perfringens | salmonella | 0.58739774 | clostridium_perfringens | yersinia_pestis | 0.4743239831 |
| clostridium_perfringens | listeria_monocytogenes | 0.603655117 | clostridium_perfringens | entamoeba_histolytica | 0.2703817266 |
| clostridium_perfringens | clostridium_botulinum | 0.635110685 | clostridium_perfringens | giardia_lamblia | 0.2703817266 |
| clostridium_perfringens | campylobacter_jejuni | 0.646306096 | clostridium_perfringens | toxoplasma_gondii | 0.2703817266 |
| | | | | | |
| | | | | | |
| coxiella_burnetii | chlamydia_psittaci | 0.669150293 | coxiella_burnetii | brucella | 0.6981551818 |
| coxiella_burnetii | brucella | 0.652046325 | coxiella_burnetii | burkholderia_mallei | 0.6837439321 |
| coxiella_burnetii | francisella_tularensis | 0.652032812 | coxiella_burnetii | francisella_tularensis | 0.6837439321 |
| coxiella_burnetii | yersinia_pestis | 0.605156222 | coxiella_burnetii | burkholderia_pseudomallei | 0.6837439321 |
| coxiella_burnetii | burkholderia_pseudomallei | 0.585284131 | coxiella_burnetii | salmonella | 0.6836982238 |
| coxiella_burnetii | rickettsia_prowazekii | 0.58367727 | coxiella_burnetii | shigella | 0.6763331409 |
| coxiella_burnetii | toxoplasma_gondii | 0.569772895 | coxiella_burnetii | yersinia_pestis | 0.6636717985 |
| coxiella_burnetii | burkholderia_mallei | 0.562032006 | coxiella_burnetii | rickettsia_prowazekii | 0.5707515826 |
| coxiella_burnetii | campylobacter_jejuni | 0.537468792 | coxiella_burnetii | campylobacter_jejuni | 0.5707515826 |
| coxiella_burnetii | shigella | 0.507879704 | coxiella_burnetii | chlamydia_psittaci | 0.5409916978 |

| | | | | | |
|---|---|---|---|---|---|
| coxiella_burnetii | bacillus_anthracis | 0.506694011 | coxiella_burnetii | clostridium_botulinum | 0.4874930494 |
| coxiella_burnetii | entamoeba_histolytica | 0.497119411 | coxiella_burnetii | clostridium_perfringens | 0.4743239831 |
| coxiella_burnetii | listeria_monocytogenes | 0.481744976 | coxiella_burnetii | listeria_monocytogenes | 0.4743239831 |
| coxiella_burnetii | giardia_lamblia | 0.459664601 | coxiella_burnetii | bacillus_anthracis | 0.4743239831 |
| coxiella_burnetii | salmonella | 0.422091432 | coxiella_burnetii | giardia_lamblia | 0.2703817266 |
| coxiella_burnetii | clostridium_perfringens | 0.415534492 | coxiella_burnetii | entamoeba_histolytica | 0.2703817266 |
| coxiella_burnetii | clostridium_botulinum | 0.352984904 | coxiella_burnetii | toxoplasma_gondii | 0.2703817266 |
| | | | | | |
| | | | | | |
| entamoeba_histolytica | giardia_lamblia | 0.796562048 | entamoeba_histolytica | giardia_lamblia | 0.3559424408 |
| entamoeba_histolytica | toxoplasma_gondii | 0.694059817 | entamoeba_histolytica | toxoplasma_gondii | 0.3559424408 |
| entamoeba_histolytica | coxiella_burnetii | 0.497119411 | entamoeba_histolytica | salmonella | 0.2785405476 |
| entamoeba_histolytica | campylobacter_jejuni | 0.474002835 | entamoeba_histolytica | clostridium_botulinum | 0.2778885679 |
| entamoeba_histolytica | rickettsia_prowazekii | 0.472009314 | entamoeba_histolytica | brucella | 0.2760805538 |
| entamoeba_histolytica | burkholderia_pseudomallei | 0.461405609 | entamoeba_histolytica | shigella | 0.2755399925 |
| entamoeba_histolytica | francisella_tularensis | 0.447512328 | entamoeba_histolytica | bacillus_anthracis | 0.2703817266 |
| entamoeba_histolytica | shigella | 0.447328837 | entamoeba_histolytica | burkholderia_mallei | 0.2703817266 |
| entamoeba_histolytica | listeria_monocytogenes | 0.440411903 | entamoeba_histolytica | burkholderia_pseudomallei | 0.2703817266 |
| entamoeba_histolytica | brucella | 0.421155891 | entamoeba_histolytica | campylobacter_jejuni | 0.2703817266 |
| entamoeba_histolytica | yersinia_pestis | 0.421061084 | entamoeba_histolytica | chlamydia_psittaci | 0.2703817266 |
| entamoeba_histolytica | burkholderia_mallei | 0.410001031 | entamoeba_histolytica | clostridium_perfringens | 0.2703817266 |
| entamoeba_histolytica | bacillus_anthracis | 0.40738203 | entamoeba_histolytica | coxiella_burnetii | 0.2703817266 |
| entamoeba_histolytica | chlamydia_psittaci | 0.401600245 | entamoeba_histolytica | francisella_tularensis | 0.2703817266 |
| entamoeba_histolytica | clostridium_perfringens | 0.393354502 | entamoeba_histolytica | listeria_monocytogenes | 0.2703817266 |
| entamoeba_histolytica | salmonella | 0.342054334 | entamoeba_histolytica | rickettsia_prowazekii | 0.2703817266 |
| entamoeba_histolytica | clostridium_botulinum | 0.272100265 | entamoeba_histolytica | yersinia_pestis | 0.2703817266 |
| | | | | | |

| | | | | | |
|---|---|---|---|---|---|
| francisella_tularensis | yersinia_pestis | 0.756898156 | francisella_tularensis | brucella | 0.6981551818 |
| francisella_tularensis | burkholderia_pseudomallei | 0.755478426 | francisella_tularensis | burkholderia_pseudomallei | 0.6837439321 |
| francisella_tularensis | listeria_monocytogenes | 0.699905915 | francisella_tularensis | coxiella_burnetii | 0.6837439321 |
| francisella_tularensis | bacillus_anthracis | 0.682484291 | francisella_tularensis | burkholderia_mallei | 0.6837439321 |
| francisella_tularensis | burkholderia_mallei | 0.656825591 | francisella_tularensis | salmonella | 0.6836982238 |
| francisella_tularensis | coxiella_burnetii | 0.652032812 | francisella_tularensis | shigella | 0.6763331409 |
| francisella_tularensis | campylobacter_jejuni | 0.617449011 | francisella_tularensis | yersinia_pestis | 0.6636717985 |
| francisella_tularensis | rickettsia_prowazekii | 0.593690513 | francisella_tularensis | campylobacter_jejuni | 0.5707515826 |
| francisella_tularensis | shigella | 0.560840049 | francisella_tularensis | rickettsia_prowazekii | 0.5707515826 |
| francisella_tularensis | chlamydia_psittaci | 0.560802847 | francisella_tularensis | chlamydia_psittaci | 0.5409916978 |
| francisella_tularensis | brucella | 0.533713552 | francisella_tularensis | clostridium_botulinum | 0.4874930494 |
| francisella_tularensis | salmonella | 0.520818212 | francisella_tularensis | clostridium_perfringens | 0.4743239831 |
| francisella_tularensis | clostridium_perfringens | 0.515747131 | francisella_tularensis | listeria_monocytogenes | 0.4743239831 |
| francisella_tularensis | toxoplasma_gondii | 0.486882499 | francisella_tularensis | bacillus_anthracis | 0.4743239831 |
| francisella_tularensis | clostridium_botulinum | 0.452875066 | francisella_tularensis | giardia_lamblia | 0.2703817266 |
| francisella_tularensis | entamoeba_histolytica | 0.447512328 | francisella_tularensis | entamoeba_histolytica | 0.2703817266 |
| francisella_tularensis | giardia_lamblia | 0.358153497 | francisella_tularensis | toxoplasma_gondii | 0.2703817266 |
| | | | | | |
| | | | | | |
| giardia_lamblia | entamoeba_histolytica | 0.796562048 | giardia_lamblia | entamoeba_histolytica | 0.3559424408 |
| giardia_lamblia | toxoplasma_gondii | 0.660509753 | giardia_lamblia | toxoplasma_gondii | 0.3559424408 |
| giardia_lamblia | rickettsia_prowazekii | 0.476872575 | giardia_lamblia | salmonella | 0.2785405476 |
| giardia_lamblia | coxiella_burnetii | 0.459664601 | giardia_lamblia | clostridium_botulinum | 0.2778885679 |
| giardia_lamblia | campylobacter_jejuni | 0.447531708 | giardia_lamblia | brucella | 0.2760805538 |
| giardia_lamblia | burkholderia_pseudomallei | 0.428295047 | giardia_lamblia | shigella | 0.2755399925 |
| giardia_lamblia | shigella | 0.402907011 | giardia_lamblia | burkholderia_mallei | 0.2703817266 |
| giardia_lamblia | burkholderia_mallei | 0.37962294 | giardia_lamblia | rickettsia_prowazekii | 0.2703817266 |
| giardia_lamblia | yersinia_pestis | 0.375460944 | giardia_lamblia | bacillus_anthracis | 0.2703817266 |
| giardia_lamblia | listeria_monocytogenes | 0.374727152 | giardia_lamblia | francisella_tularensis | 0.2703817266 |

| | | | | | |
|---|---|---|---|---|---|
| giardia_lamblia | brucella | 0.360897599 | giardia_lamblia | burkholderia_pseudomallei | 0.2703817266 |
| giardia_lamblia | bacillus_anthracis | 0.3584193 | giardia_lamblia | coxiella_burnetii | 0.2703817266 |
| giardia_lamblia | francisella_tularensis | 0.358153497 | giardia_lamblia | burkholderia_mallei | 0.2703817266 |
| giardia_lamblia | chlamydia_psittaci | 0.356361647 | giardia_lamblia | chlamydia_psittaci | 0.2703817266 |
| giardia_lamblia | clostridium_perfringens | 0.347200657 | giardia_lamblia | clostridium_perfringens | 0.2703817266 |
| giardia_lamblia | clostridium_botulinum | 0.297062759 | giardia_lamblia | listeria_monocytogenes | 0.2703817266 |
| giardia_lamblia | salmonella | 0.284989382 | giardia_lamblia | campylobacter_jejuni | 0.2703817266 |
| giardia_lamblia | staphylococcus_enterotoxin | 0.120978123 | giardia_lamblia | yersinia_pestis | 0.2703817266 |
| | | | | | |
| | | | | | |
| listeria_monocytogenes | francisella_tularensis | 0.699905915 | listeria_monocytogenes | bacillus_anthracis | 0.8109947456 |
| listeria_monocytogenes | salmonella | 0.690297156 | listeria_monocytogenes | clostridium_botulinum | 0.6951956203 |
| listeria_monocytogenes | bacillus_anthracis | 0.677570833 | listeria_monocytogenes | clostridium_perfringens | 0.676415707 |
| listeria_monocytogenes | shigella | 0.653373599 | listeria_monocytogenes | salmonella | 0.4886368013 |
| listeria_monocytogenes | yersinia_pestis | 0.650864802 | listeria_monocytogenes | brucella | 0.484321295 |
| listeria_monocytogenes | campylobacter_jejuni | 0.635935657 | listeria_monocytogenes | shigella | 0.4833730016 |
| listeria_monocytogenes | burkholderia_pseudomallei | 0.61615112 | listeria_monocytogenes | campylobacter_jejuni | 0.4743239831 |
| listeria_monocytogenes | clostridium_perfringens | 0.603655117 | listeria_monocytogenes | yersinia_pestis | 0.4743239831 |
| listeria_monocytogenes | brucella | 0.547157641 | listeria_monocytogenes | burkholderia_mallei | 0.4743239831 |
| listeria_monocytogenes | burkholderia_mallei | 0.538628358 | listeria_monocytogenes | rickettsia_prowazekii | 0.4743239831 |
| listeria_monocytogenes | burkholderia_mallei | 0.538628358 | listeria_monocytogenes | francisella_tularensis | 0.4743239831 |
| listeria_monocytogene | clostridium_botulinum | 0.490690845 | listeria_monocytogenes | burkholderia_pseudo | 0.4743239831 |

| | | | | | |
|---|---|---|---|---|---|
| s | | | | mallei | |
| listeria_monocytogenes | toxoplasma_gondii | 0.489290828 | listeria_monocytogenes | coxiella_burnetii | 0.4743239831 |
| listeria_monocytogenes | coxiella_burnetii | 0.481744976 | listeria_monocytogenes | burkholderia_mallei | 0.4743239831 |
| listeria_monocytogenes | rickettsia_prowazekii | 0.467835382 | listeria_monocytogenes | chlamydia_psittaci | 0.4743239831 |
| listeria_monocytogenes | entamoeba_histolytica | 0.440411903 | listeria_monocytogenes | giardia_lamblia | 0.2703817266 |
| listeria_monocytogenes | chlamydia_psittaci | 0.386162062 | listeria_monocytogenes | entamoeba_histolytica | 0.2703817266 |
| listeria_monocytogenes | giardia_lamblia | 0.374727152 | listeria_monocytogenes | toxoplasma_gondii | 0.2703817266 |
| | | | | | |
| | | | | | |
| rickettsia_prowazekii | yersinia_pestis | 0.641650922 | rickettsia_prowazekii | brucella | 0.7482048658 |
| rickettsia_prowazekii | burkholderia_mallei | 0.603512545 | rickettsia_prowazekii | salmonella | 0.5879741223 |
| rickettsia_prowazekii | francisella_tularensis | 0.593690513 | rickettsia_prowazekii | shigella | 0.5816402194 |
| rickettsia_prowazekii | coxiella_burnetii | 0.58367727 | rickettsia_prowazekii | burkholderia_mallei | 0.5707515826 |
| rickettsia_prowazekii | bacillus_anthracis | 0.543889904 | rickettsia_prowazekii | burkholderia_pseudomallei | 0.5707515826 |
| rickettsia_prowazekii | burkholderia_pseudomallei | 0.525044474 | rickettsia_prowazekii | campylobacter_jejuni | 0.5707515826 |
| rickettsia_prowazekii | toxoplasma_gondii | 0.487718141 | rickettsia_prowazekii | coxiella_burnetii | 0.5707515826 |
| rickettsia_prowazekii | giardia_lamblia | 0.476872575 | rickettsia_prowazekii | francisella_tularensis | 0.5707515826 |
| rickettsia_prowazekii | entamoeba_histolytica | 0.472009314 | rickettsia_prowazekii | yersinia_pestis | 0.5707515826 |
| rickettsia_prowazekii | chlamydia_psittaci | 0.469599243 | rickettsia_prowazekii | chlamydia_psittaci | 0.5409916978 |
| rickettsia_prowazekii | listeria_monocytogenes | 0.467835382 | rickettsia_prowazekii | clostridium_botulinum | 0.4874930494 |
| rickettsia_prowazekii | campylobacter_jejuni | 0.44029191 | rickettsia_prowazekii | bacillus_anthracis | 0.4743239831 |
| rickettsia_prowazekii | shigella | 0.416090153 | rickettsia_prowazekii | clostridium_perfringens | 0.4743239831 |
| rickettsia_prowazekii | clostridium_botulinum | 0.404366669 | rickettsia_prowazekii | listeria_monocytogenes | 0.4743239831 |
| rickettsia_prowazekii | brucella | 0.396640513 | rickettsia_prowazekii | entamoeba_histolytica | 0.2703817266 |
| rickettsia_prowazekii | clostridium_perfringen | 0.359320385 | rickettsia_prowazekii | giardia_lamblia | 0.2703817266 |

| | | | | | |
|---|---|---|---|---|---|
| | s | | | | |
| rickettsia_prowazekii | salmonella | 0.296296183 | rickettsia_prowazekii | toxoplasma_gondii | 0.2703817266 |
| | | | | | |
| | | | | | |
| salmonella | shigella | 0.69039736 | salmonella | shigella | 0.8385584166 |
| salmonella | listeria_monocytogenes | 0.690297156 | salmonella | yersinia_pestis | 0.8223864255 |
| salmonella | campylobacter_jejuni | 0.658679635 | salmonella | coxiella_burnetii | 0.6836982238 |
| salmonella | clostridium_perfringens | 0.58739774 | salmonella | francisella_tularensis | 0.6836982238 |
| salmonella | brucella | 0.557399382 | salmonella | brucella | 0.6007489119 |
| salmonella | bacillus_anthracis | 0.536186579 | salmonella | burkholderia_mallei | 0.5879741223 |
| salmonella | francisella_tularensis | 0.520818212 | salmonella | burkholderia_pseudomallei | 0.5879741223 |
| salmonella | yersinia_pestis | 0.482391244 | salmonella | campylobacter_jejuni | 0.5879741223 |
| salmonella | coxiella_burnetii | 0.422091432 | salmonella | rickettsia_prowazekii | 0.5879741223 |
| salmonella | burkholderia_pseudomallei | 0.41936872 | salmonella | chlamydia_psittaci | 0.5573162272 |
| salmonella | burkholderia_mallei | 0.391549924 | salmonella | clostridium_botulinum | 0.5026243353 |
| salmonella | clostridium_botulinum | 0.358533515 | salmonella | bacillus_anthracis | 0.4886368013 |
| salmonella | entamoeba_histolytica | 0.342054334 | salmonella | clostridium_perfringens | 0.4886368013 |
| salmonella | toxoplasma_gondii | 0.331135772 | salmonella | listeria_monocytogenes | 0.4886368013 |
| salmonella | chlamydia_psittaci | 0.306779799 | salmonella | entamoeba_histolytica | 0.2785405476 |
| salmonella | rickettsia_prowazekii | 0.296296183 | salmonella | giardia_lamblia | 0.2785405476 |
| salmonella | giardia_lamblia | 0.284989382 | salmonella | toxoplasma_gondii | 0.2785405476 |
| | | | | | |
| | | | | | |
| shigella | salmonella | 0.69039736 | shigella | salmonella | 0.8385584166 |
| shigella | listeria_monocytogenes | 0.653373599 | shigella | yersinia_pestis | 0.8135273354 |
| shigella | campylobacter_jejuni | 0.628804943 | shigella | francisella_tularensis | 0.6763331409 |
| shigella | bacillus_anthracis | 0.56467806 | shigella | coxiella_burnetii | 0.6763331409 |
| shigella | francisella_tularensis | 0.560840049 | shigella | brucella | 0.5941383356 |
| shigella | yersinia_pestis | 0.556054931 | shigella | campylobacter_jejuni | 0.5816402194 |
| shigella | brucella | 0.550852083 | shigella | rickettsia_prowazekii | 0.5816402194 |

| | | | | | |
|---|---|---|---|---|---|
| shigella | burkholderia_pseudomallei | 0.52229287 | shigella | burkholderia_pseudomallei | 0.5816402194 |
| shigella | clostridium_perfringens | 0.518906715 | shigella | burkholderia_mallei | 0.5816402194 |
| shigella | coxiella_burnetii | 0.507879704 | shigella | rickettsia_prowazekii | 0.5816402194 |
| shigella | entamoeba_histolytica | 0.447328837 | shigella | chlamydia_psittaci | 0.551312584 |
| shigella | burkholderia_mallei | 0.423452341 | shigella | clostridium_botulinum | 0.4970565801 |
| shigella | burkholderia_mallei | 0.423452341 | shigella | listeria_monocytogenes | 0.4833730016 |
| shigella | rickettsia_prowazekii | 0.416090153 | shigella | bacillus_anthracis | 0.4833730016 |
| shigella | toxoplasma_gondii | 0.409355717 | shigella | clostridium_perfringens | 0.4833730016 |
| shigella | giardia_lamblia | 0.402907011 | shigella | giardia_lamblia | 0.2755399925 |
| shigella | clostridium_botulinum | 0.365614364 | shigella | entamoeba_histolytica | 0.2755399925 |
| shigella | chlamydia_psittaci | 0.36050759 | shigella | toxoplasma_gondii | 0.2755399925 |
| | | | | | |
| | | | | | |
| toxoplasma_gondii | entamoeba_histolytica | 0.694059817 | toxoplasma_gondii | entamoeba_histolytica | 0.3559424408 |
| toxoplasma_gondii | giardia_lamblia | 0.660509753 | toxoplasma_gondii | giardia_lamblia | 0.3559424408 |
| toxoplasma_gondii | coxiella_burnetii | 0.569772895 | toxoplasma_gondii | salmonella | 0.2785405476 |
| toxoplasma_gondii | listeria_monocytogenes | 0.489290828 | toxoplasma_gondii | clostridium_botulinum | 0.2778885679 |
| toxoplasma_gondii | rickettsia_prowazekii | 0.487718141 | toxoplasma_gondii | brucella | 0.2760805538 |
| toxoplasma_gondii | francisella_tularensis | 0.486882499 | toxoplasma_gondii | shigella | 0.2755399925 |
| toxoplasma_gondii | chlamydia_psittaci | 0.458182392 | toxoplasma_gondii | bacillus_anthracis | 0.2703817266 |
| toxoplasma_gondii | burkholderia_pseudomallei | 0.448829873 | toxoplasma_gondii | burkholderia_mallei | 0.2703817266 |
| toxoplasma_gondii | brucella | 0.445962129 | toxoplasma_gondii | burkholderia_pseudomallei | 0.2703817266 |
| toxoplasma_gondii | yersinia_pestis | 0.415945888 | toxoplasma_gondii | campylobacter_jejuni | 0.2703817266 |
| toxoplasma_gondii | burkholderia_mallei | 0.415286353 | toxoplasma_gondii | chlamydia_psittaci | 0.2703817266 |
| toxoplasma_gondii | burkholderia_mallei | 0.415286353 | toxoplasma_gondii | clostridium_perfringens | 0.2703817266 |
| toxoplasma_gondii | shigella | 0.409355717 | toxoplasma_gondii | coxiella_burnetii | 0.2703817266 |
| toxoplasma_gondii | campylobacter_jejuni | 0.368116237 | toxoplasma_gondii | francisella_tularensis | 0.2703817266 |

| | | | | | |
|---|---|---|---|---|---|
| toxoplasma_gondii | bacillus_anthracis | 0.357918893 | toxoplasma_gondii | listeria_monocytogenes | 0.2703817266 |
| toxoplasma_gondii | salmonella | 0.331135772 | toxoplasma_gondii | rickettsia_prowazekii | 0.2703817266 |
| toxoplasma_gondii | clostridium_perfringens | 0.262706126 | toxoplasma_gondii | rickettsia_prowazekii | 0.2703817266 |
| toxoplasma_gondii | clostridium_botulinum | 0.203076924 | toxoplasma_gondii | yersinia_pestis | 0.2703817266 |
| | | | | | |
| | | | | | |
| yersinia_pestis | bacillus_anthracis | 0.758683807 | yersinia_pestis | salmonella | 0.8223864255 |
| yersinia_pestis | francisella_tularensis | 0.756898156 | yersinia_pestis | shigella | 0.8135273354 |
| yersinia_pestis | burkholderia_pseudomallei | 0.719093848 | yersinia_pestis | coxiella_burnetii | 0.6636717985 |
| yersinia_pestis | burkholderia_mallei | 0.658958001 | yersinia_pestis | francisella_tularensis | 0.6636717985 |
| yersinia_pestis | listeria_monocytogenes | 0.650864802 | yersinia_pestis | brucella | 0.5827812959 |
| yersinia_pestis | rickettsia_prowazekii | 0.641650922 | yersinia_pestis | burkholderia_mallei | 0.5707515826 |
| yersinia_pestis | rickettsia_prowazekii | 0.641650922 | yersinia_pestis | burkholderia_pseudomallei | 0.5707515826 |
| yersinia_pestis | campylobacter_jejuni | 0.625828469 | yersinia_pestis | campylobacter_jejuni | 0.5707515826 |
| yersinia_pestis | coxiella_burnetii | 0.605156222 | yersinia_pestis | rickettsia_prowazekii | 0.5707515826 |
| yersinia_pestis | shigella | 0.556054931 | yersinia_pestis | rickettsia_prowazekii | 0.5707515826 |
| yersinia_pestis | clostridium_perfringens | 0.521380081 | yersinia_pestis | chlamydia_psittaci | 0.5409916978 |
| yersinia_pestis | clostridium_botulinum | 0.493559377 | yersinia_pestis | clostridium_botulinum | 0.4874930494 |
| yersinia_pestis | salmonella | 0.482391244 | yersinia_pestis | bacillus_anthracis | 0.4743239831 |
| yersinia_pestis | brucella | 0.474110716 | yersinia_pestis | clostridium_perfringens | 0.4743239831 |
| yersinia_pestis | chlamydia_psittaci | 0.459664546 | yersinia_pestis | listeria_monocytogenes | 0.4743239831 |
| yersinia_pestis | entamoeba_histolytica | 0.421061084 | yersinia_pestis | entamoeba_histolytica | 0.2703817266 |
| yersinia_pestis | toxoplasma_gondii | 0.415945888 | yersinia_pestis | giardia_lamblia | 0.2703817266 |
| yersinia_pestis | giardia_lamblia | 0.375460944 | yersinia_pestis | toxoplasma_gondii | 0.2703817266 |

# Appendix B: Chapter 3 Figures

**Figure 3.1 "Pathogen" cluster of related terms.**

**Figure 3.2 Carbohydrate cluster of related terms.**

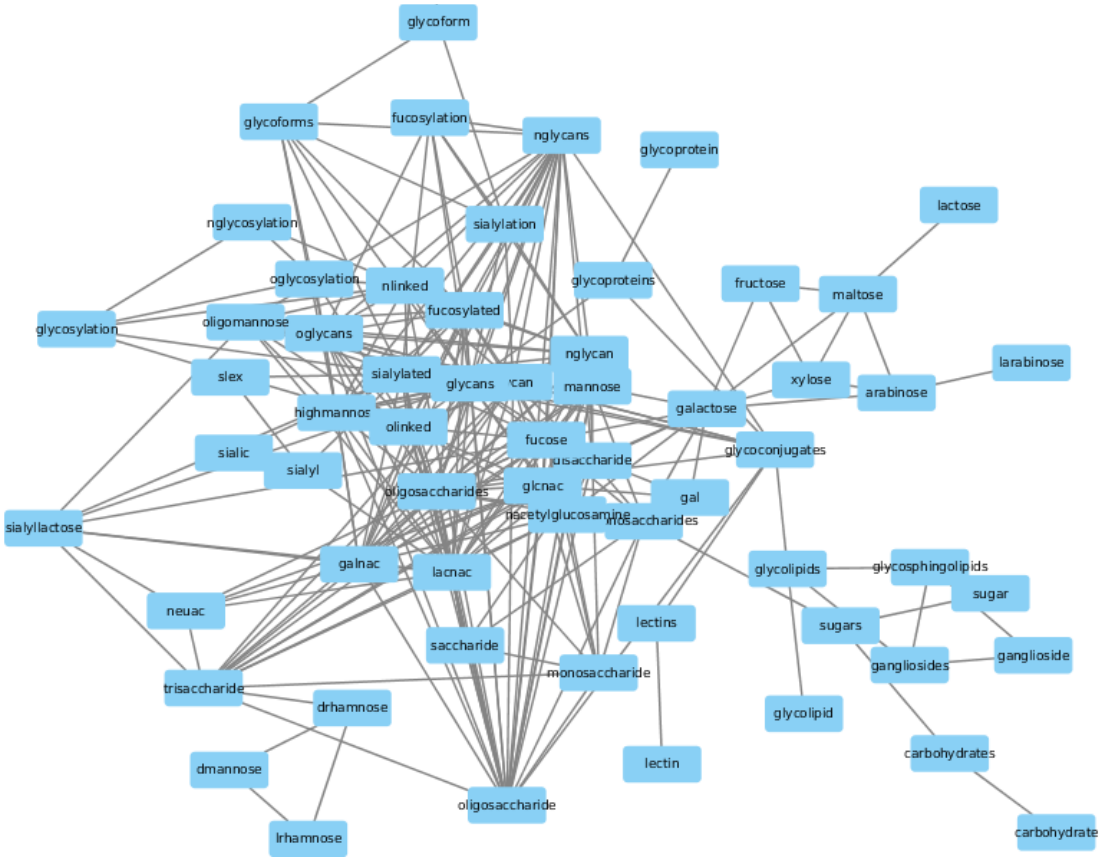

**Figure 3.3 Lyme disease cluster of related terms**

**Figure 3.4 Virus cluster of related terms**

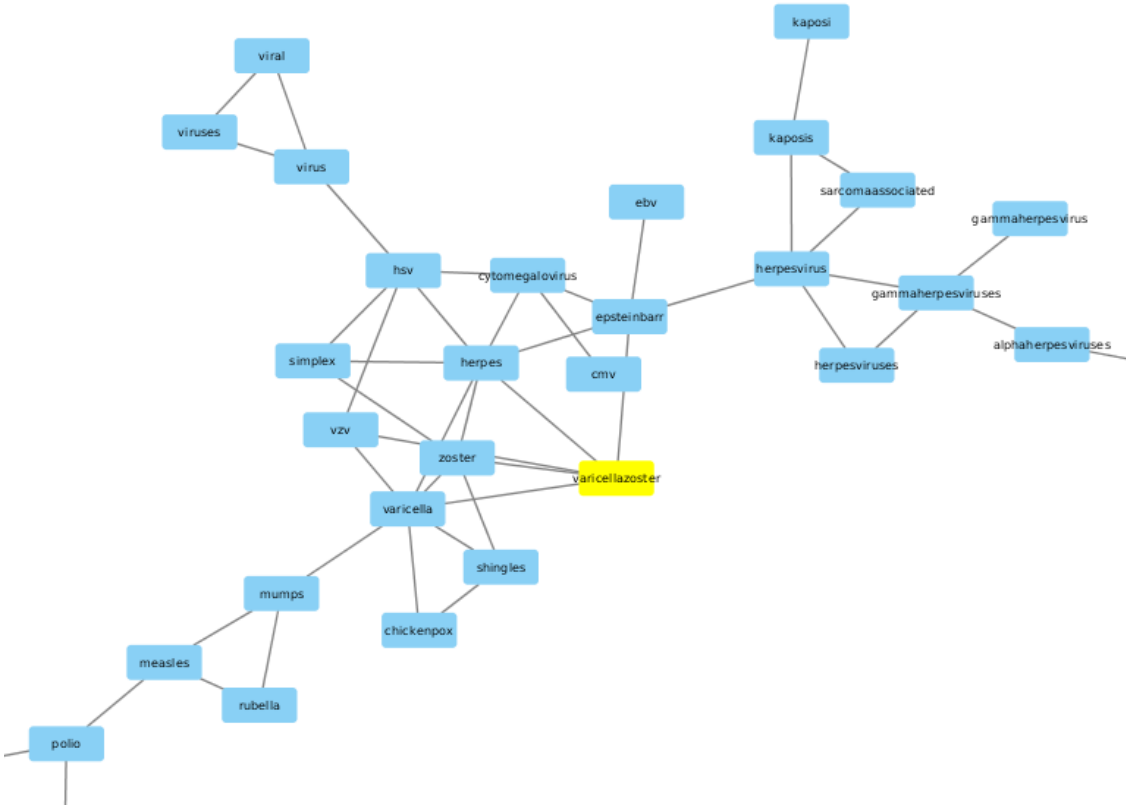

**Figure 3.5 Adhesion related virulence factors cluster of related terms.**

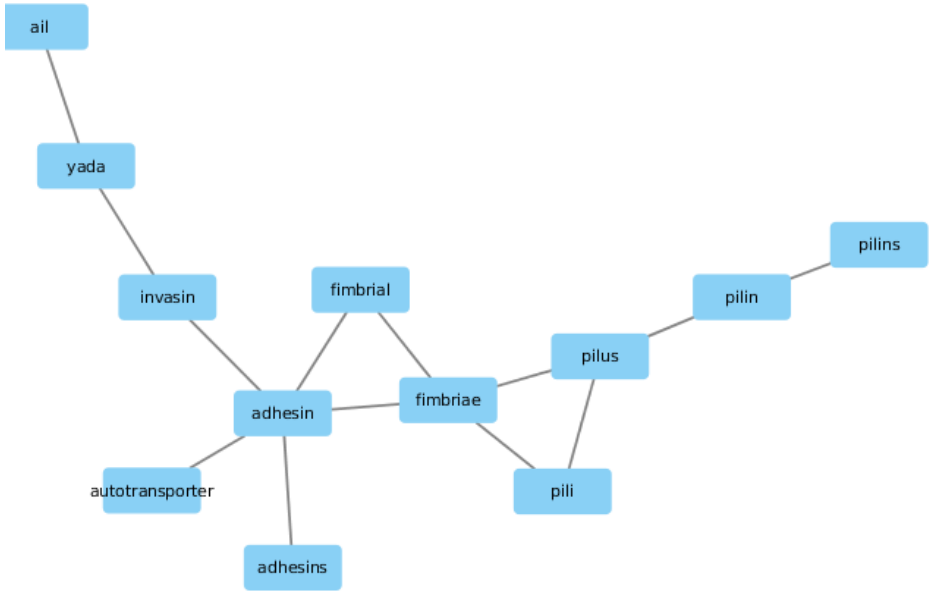

**Figure 3.6 Genetic mutation cluster of related terms**

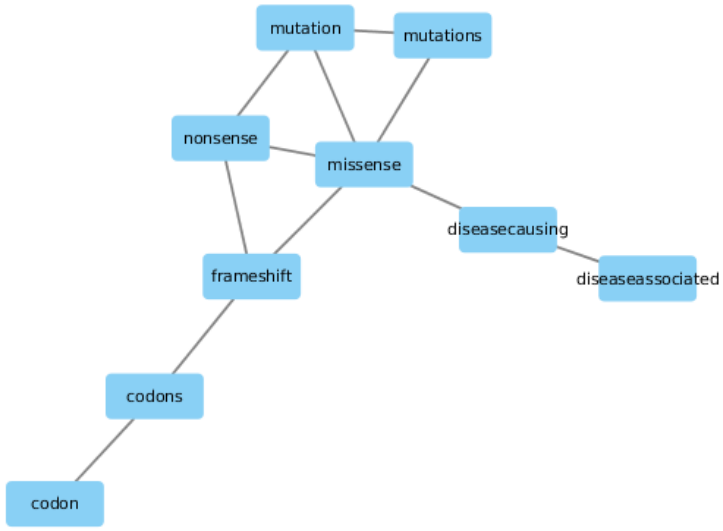

**Figure 3.7 Antiseptic cluster of related terms.**

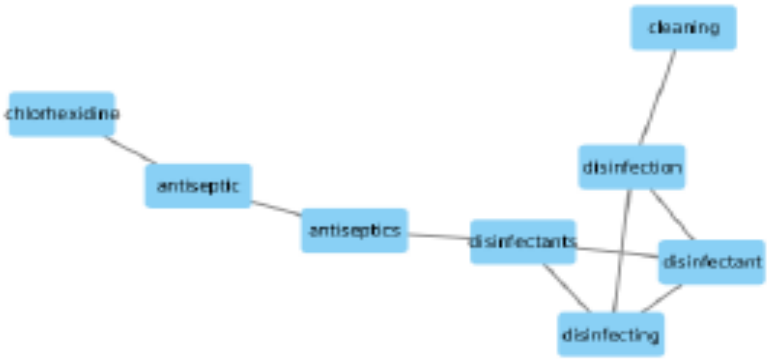

**Figure 3.8 Organism taxonomy related cluster of related terms.**

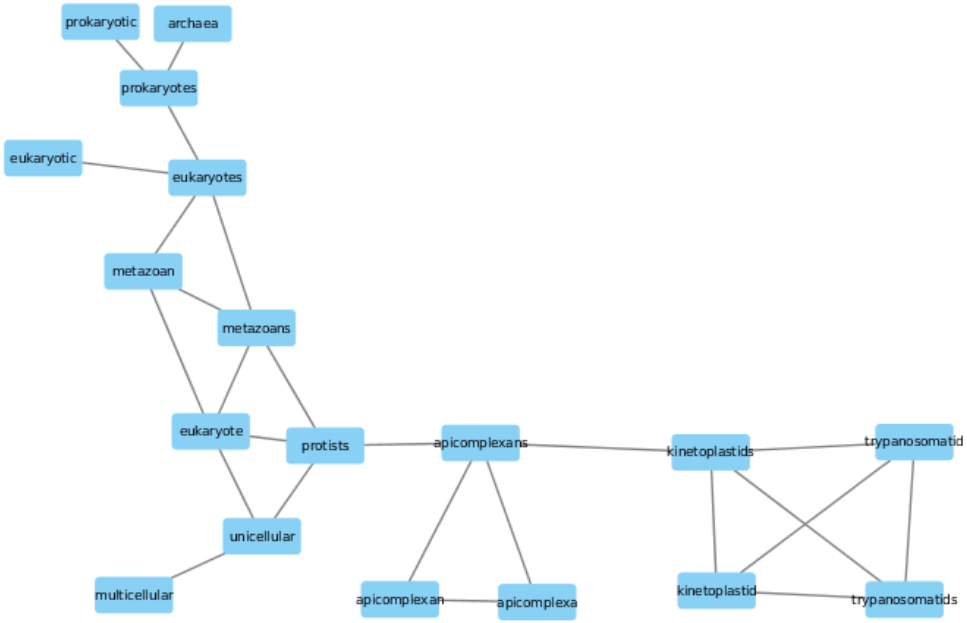

**Figure 3.9 Immunology related cluster of related terms.**