Project Summary

Overview

In this document the PIs describe an approach to identify trends in disease treatment by applying existing and novel machine learning techniques. Data sources will include both direct clinical sources and also pertinent text and metadata acquired from users of social media. High level information, analytical tools, and data resources produced through these analyses will be disseminated back into both clinical and social media venues to improve patient's direct medical outcomes and address social concerns. One specific application will focus on the use of antiretroviral therapy to preventatively treat individuals who are at-risk for HIV infection. This use of antiretroviral therapy is referd to as Pre-Exposure Prophylaxis (PrEP) and uses the small molecule nucleotide reverse transcriptase inhibitor trade named Truvada.

Intellectual Merit

The information produced through data mining and machine learning will be useful to provide clinicians and public health professionals feedback concerning successes and challenges associated with disease treatments. Pertinent tools and summary datasets will also be produced and made available to researchers. In addition novel applications of machine learning and data science methods will be developed as necessary, and shared with the data science community to assist and inspire related applications in other domain areas.

Broader Impacts

The proposed research will contribute to an increased scientific literacy through collaborations and dissemination of our findings through social media. This research will also lead to improvements of the health of all Americans, with in some cases such as PrEP, special emphasis on underrepresented groups such as the LGBT community. This work also has the potential to engage collaborations and partnerships between the social media industry, medical providers and pharmaceutical companies, and academia as we come together to improve public health.

Project Description

1 Mining Pre-Exposure Prophylaxis Trends in Social Media

Pre-Exposure Prophylaxis (PrEP) is a recently developed method for the prevention of Human Immunodeficiency Virus (HIV) via the administration of an oral pharmaceutical trade named Truvada. Truvada contains active ingredients tenofovir and emtricitabine, both Nucleotide Reverse Transcriptase Inhibitors (NRTIs). In the last four years, since Truvada was approved for PrEP in 2012, PrEP has shown demonstrated efficacy at preventing HIV for HIV negative individuals in serodiscordant relationships[?].

Initial studies of PrEP have shown that it is highly effective[?], however because it is still a new treatment, it is facing a number of medical and social obstacles before it reaches full adoption. Incomplete clinical and patient understanding, social stigma, and uncertain insurance status have been identified as challenges preventing continued adoption[?]. Also, since Truvada is an oral NRTI, it must be taken daily. Cases where patients do not adhere to their full prescription have led to loss of viral protection, and some patients and clinicians worry that lack of adherence could lead to increased risk of infection with drug resistant strains[?].

In order to identify factors that lead to the direct mitigation of HIV infection, we would like to be able to access direct medical data, however doing so can be difficult due to privacy regulations. Additionally, medical data often does not contain direct unfiltered opinions and feedback which capture patient and public perception. For this reason we propose the use of social media as one of our principle sources of data. Though we have investigated the use of multiple social media platforms including Facebook, Grindr, Reddit and Twitter, we have focused primarily on the use of Twitter for past projects and intend to use Twitter for future analyses. Twitter is especially useful because in addition to textual data, various useful metadata is also available including datetime, geolocation, username, hastags, and external hyperlinks. We have used many of these metadata attributes in previous analyses and will continue to do so in future analyses. Twitter also features a convenient well documented and free API and has over 300 million monthly active users.

In past analyses we collected over 1 million tweets from the Twitter streaming API filtered on PrEP and HIV related keywords. By using embedding techniques such as Word2Vec and Doc2Vec, we were able to identify new keywords and hashtags that are relevant to the PrEP conversation on Twitter including unexpected political connections ”NancyReagan” and popular hashtags associated with PrEP that might not have been known to researchers in advance such as ”#whereisprep” (see table 1 for examples of hashtags and structurally-related tweets). In addition Doc2Vec can be used to query the top N tweets related to a given hashtag or the top N users related to a given hashtag. This allowed us to identify a small subset of the N tweets most structurally related to ”PrEP” that can be human-read without requiring humans to read the full dataset. Reading from the set of structurally most important tweets, uncovered important blog articles describing some of the fears of drug resistant forms of HIV and concerns as to whether these strains could be caused in part by over use or misuse of PrEP medications. The linked blog articles that we found highlight the usefulness of the hyperlink metadata embedded in tweets. Through these hyperlinks, Twitter acts as an index, providing indirect access to a much larger external social media ecosystem.

Our Doc2Vec results also allowed us to identify the top N users most relevant to PrEP. By querying the Twitter REST API for these users’ timelines, and using topic modeling methods like Latent Dirichlet Analysis (LDA), we identified the word-distribution-topics present in the Twitter

Table 1: Cosine similarity to document-vector “#PrEP”

Related hashtag/tweet	Cosine similarity to #PrEP
#lgbtmedia16	0.739128
#hiv	0.727602
#whereisprep	0.707165
#truvada	0.696113
#hivprevention	0.636068
tweet-702179860983189504	0.630055
user-711275699529764864	0.629254
tweet-708519265540907010	0.628778
tweet-712032637024653313	0.628646
#harrogatehour	0.628547

conversation (see figure 1). This analysis went beyond the simple keyword identification analysis from Word2Vec since it clustered keywords into topics and it operated on all tweets that PrEP-related users tweeted, not just PrEP related tweets. The LDA results showed a variety of related concerns such as other sexually transmitted diseases, LGBT related topics, health insurance and political topics. Neither PrEP or Truvada were present in the top 30 keywords related to HIV/AIDS demonstrating that PrEP is still a nascent rare topic in the online discussion. An extension to LDA, Dynamic Topic Modeling (DTM), was able to capture topic and word frequency over time. The DTM results showed that the keyword ”PrEP” is increasing in relative frequency over time, even relative to related words such as ”pill”, ”prevention” and ”drug”. This demonstrates increased interest in PrEP which may correlate with an increase in PrEP adoption over the data acquisition period.

Using an open dataset of tweets which were labeled with binary sentiment labels, we performed a sentiment analysis. This analysis allowed us to identify N PrEP related tweets with the highest sentiment, and N PrEP tweets with the lowest sentiment. After performing the automated sentiment analysis, a human was added into the loop to quickly read the top positive and negative tweets to get a sense of the successes and issues present in public perception. In the positive tweets we found hyperlinks to blogs with positive firsthand accounts from individuals successfully using PrEP to stay HIV negative. In the negative tweets we found concerns of whether Truvada can protect against drug resistant strains of HIV (example negative tweets shown in table 2). Together these approaches and results show that we can take raw text and metadata and extract keywords, hastags, temporal trends, and sentiment information. Doc2Vec and sentiment classification allow the researcher to extract a set of the N most highly relevant tweets from a large corpus that can be easily human-readable.

Though our previous research on data mining social media has uncovered important tweets, keywords, sentiments and hashtags related to the Twitter discussion of PrEP, many important questions remain poorly understood. One important issue is determining why patients stop adhering to their PrEP medication. While our LDA results uncovered ”stigma” and other related keywords, and some of the critical tweets we identified described uncertainty in the efficacy of PrEP, this question still remains to be fully answered. One of the challenges with using Twitter as a data source is that we cant verify ground truth labels on the people authoring the tweets. For example

Table 2: Negative sentiment tweets.

Category	Text
General	“Also, how f***ing vile of Hillary to say. Reagan did f***ing NOTHING during the AIDS epidemic until it was too late. What a stupid old hag.”
General	“I wonder why he beat her a** when she was tryna leave like she wasn’t gone be running back when she found out she had HIV & nobody want her”
General	“Aaannd. Hillary Clinton breathes a sigh of relief that Twitter has left its outrage of her AIDS comments behind to tend to Drumpf debacle.”
PrEP specific	“RT gaston_croupier #Truvada patent’s not expired yet but it is sold online as a generic drug? There’s something rotten in internet #PrEP h”
PrEP specific	“Equality_MI Syph & Hep C have gone up 550% in Gay Men bc many feel tht bc they’re on PrEP, they don’t need condoms. HIV isn’t the only STI.”
PrEP specific	“Xaviom8 in interviews he says he was adherent. strain was highly resistant, and Truvada wouldn’t have blocked it anyways. PrEP didn’t fail.”
Truvada specific	“not surprised at all that someone got HIV on truvada. people get pregnant on birth control. tomato-condoms are still important-tomahto”
Truvada specific	“Now reading that truvada does not protect against certain strains of the HIV virus. Yet people want to take that risk..”
Truvada specific	“I think I have conjunctivitis unless truvada cured it overnight cuz im not feeling as horrible today as last night”

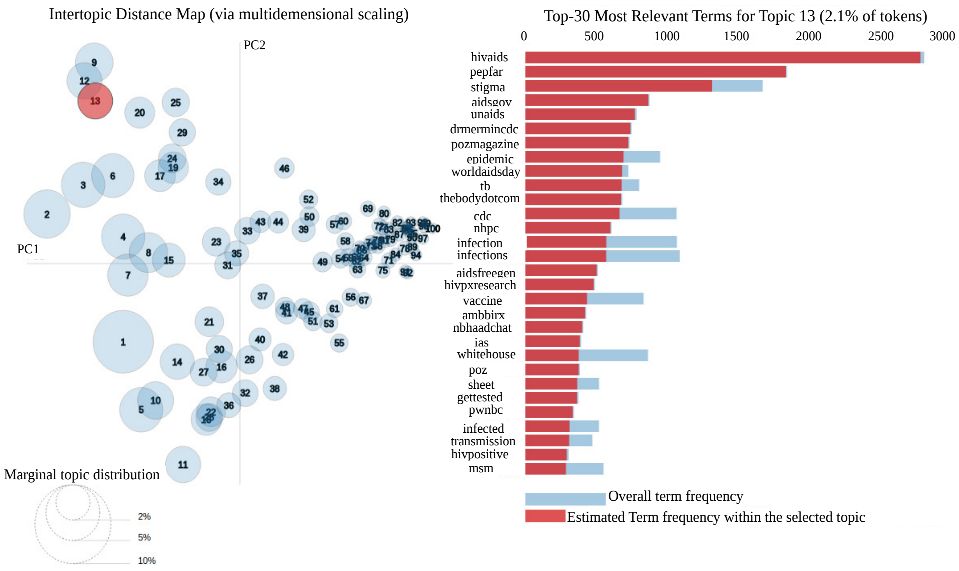

Figure 1: LDA topic modeling for the top 500 users related to PrEP.

we don't know if they are taking Truvada, or whether they are HIV negative or positive, or other important details of their medical status. By incorporating direct medical data we can identify connections between sentiment and written opinions with more concrete medical outcomes.

Other extensions to our previous analysis could include the investigation of additional diseases or behaviors that could correlate with risk of developing HIV. One of the overall goals of biosurveilance is to predict disease outbreaks before they happen in order to intervene preventatively. An outbreak in Scott county Indiana in early 2016 was thought to have been caused by drug usage. In a town of 4,000 people 135 people were diagnosed with HIV, and about 80% of those diagnosed were codiagnosed with hepatitis C. In theory some online social activity may have been able to indirectly identify this outbreak before it happened[**?**]. In order to identify and predict these hot-spots we can make use of the geolocation metadata, and also do network analyses to identify subgroups, and how information and via inference, social interactions propagate through these subgroups. Some first steps along these lines have already been made by our collaborators [**?**].

The benefit of our results over these other results is that we have started to take advantage of qualitative information in the natural language of the tweet without relying on labels or simple filtering strategies. If we take advantage of some of the natural language processing techniques described in our previous work, combined with geolocation data and hard medical outcomes, we could build a more accurate predictor of HIV outbreaks. We could also use these predictive models to identify the attributes contributing to increased risk and use public health efforts to alleviate those issues. Ultimately we hope to reduce the spread of HIV overall and also predict and prevent acute outbreaks before and as they are happening.

Finally, our previous research on mining qualitative sentiments surrounding disease treatment provides an important contribution to the larger data science community. We have shown a specific application where we mine social sentiment to identify what is working and what the challenges are for PrEP, but a simmilar framework could easily be taken and applied to improve the social barriers surrounding some other disease treatment like chemotherapy. Pharmaceutical companies and hospitals can use our open source code with minimal modification to monitor their disease and treatment of interest to monitor and improve the outcomes and happiness of their patients.

2 Semi supervised learning and operations on arbitrary dimension tensors (working with Arvind Ramanathan, Oak Ridge National Laboratory)

In January 2017 I will start working at Oak Ridge National Laboratory on a data science related fellowship. The goals of the grant supporting me seeks to use novel CPU/GPU hybrid chips to develop novel computational models in the area of semi-supervised deep learning. While I don't understand all of the details of the research proposal, or the specifics of the hardware at the time of writing (early November 2015), I will attempt to give a broad explanation of the research area and some possible directions.

Semi-supervised learning refers to a situation where both labeled and unlabeled data are used to train a discriminative model[?]. In contrast supervised learning uses only labeled data to train a discriminative model and unsupervised learning uses only unlabeled data to train model that captures some structure of the data distribution ie clustering or generative. Supervised learning is often on of the most used and useful types of machine learning since the model is directly predicting a label that has external research value. However, producing that labeled data can often require human annotation or physical experimentation with can be costly. Because labeled data is costly, machine learning models in many domains are error-limited by a lack labeled data to train on[?], which is perhaps ironic in a world of so called ”big data”. By taking advantage of both an unsupervised model of the data-distribution using cheap unlabeled data, and a discriminative model of the labels given the data on some small labeled dataset, semi supervised learning can produce results better than a n approach that uses only supervised learning.

Let me give an example of a semi supervised learning method in action, and for added consistency, I'll use an example from my own work described in the previous section (Mining Pre-Exposure Prophylaxis Trends in Social Media). When we did the sentiment analysis in the previous section we actually did 2 discrete steps. First we used Doc2Vec to transform each tweet into a dense high dimensional doc-vector, then we used a very simple logistic regression model to classify the tweet doc vectors into positive or negative sentiments. As a human reading the tweets in the results, you might be surprised at the relative accuracy of the classification, especially considering that logistic regression is relatively speaking one of the simplest supervised classifiers, and the natural language present in the tweets is very terse and complicated for such a simple supervised logistic regression model to work so well. Doc2Vec, an unsupervised embedding method produced dense document vectors that were able to capture the distribution of the tweet data, mapping tweets that had co-occurring word semantics to be ”close” in a high dimensional space. Thus the combination of unsupervised embedding and supervised classification yielded better results than a simple supervised classification could have done on its own (we didn't try doing classification without using Doc2Vec in the PrEP paper, but see the original Doc2Vec paper for relative quantification of accuracy gained by embedding[?]). Embedding, which is just one example of semi-supervised learning has become very widely used as a single step in a larger deep learning model (see the last section in this proposal below where we use embedding in a bioinformatics deep learning context) also others[?].

Though we know that we will be using novel hardware at Oak Ridge National Laboratory we aren't currently sure of the hardware details. We know that we will have access to new CPU / GPU hybrid chips that are being produced by Nvidia. GPUs have become increasingly popular in machine learning in the last couple of years, especially in the area of deep learning. This is largely because general purpose programming APIs such as CUDA have allowed researchers to take advantage of cost efficient computing power present in GPU's for computation-bound tasks (such as deep learning). On Nvidia's Blog an article titled "Accelerating AI with GPUs: A New Computing Model" describes how typical deep learning models can be trained 50x faster on a Tesla M40 GPU than on a typical CPU. Though Nvidia typically provides a compiler toolchain supporting a language called CUDA with basic C-like syntax, Nvidia has also provided a series of highly optimized routines for things like linear algebra, matrix multiplication, convolution and manipulation that allow researchers to think in terms of linear algebra primitives on matrices instead of low level array-wise manipulation[**?**]. This makes construction of machine learning models in low-level CUDA code very analogous to using popular higher level machine learning libraries in languages such as Python.

In order to provide support for semi-supervised learning on these novel CPU/GPU hybrid chips, we want to develop computational operations and building blocks that allow for arbitrary dimension tensors. Though most machine learning models use operations on matrices or 3-tensors, some theoretical and practical applications could take advantage of operations on tensors of arbitrary dimension. Think of the embedding example above. We embedded a 1 dimensional sequence of word-tokens (a document) into a dense 1 dimensional real-valued vector representation. In other situations we may want to produce arbitrary dimensional embeddings as a building block to construct semi supervised deep learning models. Several papers show examples at a theoretical level, where operations, projections and decompositions on high rank tensors can be used to extract feature information from the original data[**?**, **?**]. These operations such as Principle Components Analysis, Multi-Dimensional Scaling, Locally Preserving Projection, Neighborhood Preserving projection etc. could be used to provide embeddings to be used in a larger deep learning network either by preprocessing, or by composing the objective function of the embedding kernel with the other parts of the deep learning network in order to train the whole network through back propagation.

Practically we could in theory build small programs that implement these tensor operations in a small CUDA or C-like program. However, recent development of deep learning frameworks gives us another practical implementation option. Frameworks such as Caffe, Theano, Torch and Tensorflow allow machine learning researchers to build and share efficient machine learning models that take advantage of a simple scripting-level API, and take advantage of highly parallel efficient machine implementations. Of these frameworks, the one that I am most familiar with is Tensorflow, an open source project supported by Google [**?**].

Tensorflow is currently under fast paced development, but already it offers support for data-flow parallelism, distributed and multi-device computing, automatic differentiation, and it offers C++ and Python based APIs for specifying models. Creating a deep learning model in Tensorflow is done by composing building blocks, or "layers", into a graph. In the graph operations are nodes and tensors are edges that feed out of one node and into another node. In a typical supervised learning situation data flows in a directed acyclic path from the input node(s) to the objective or classification node(s). Through automatic differentiation, the error or "loss" in the objective function is used to produce parameter updates through gradient decent. These updates are then propagated backwards through the graph in the direction opposite data-flow. This iterative process

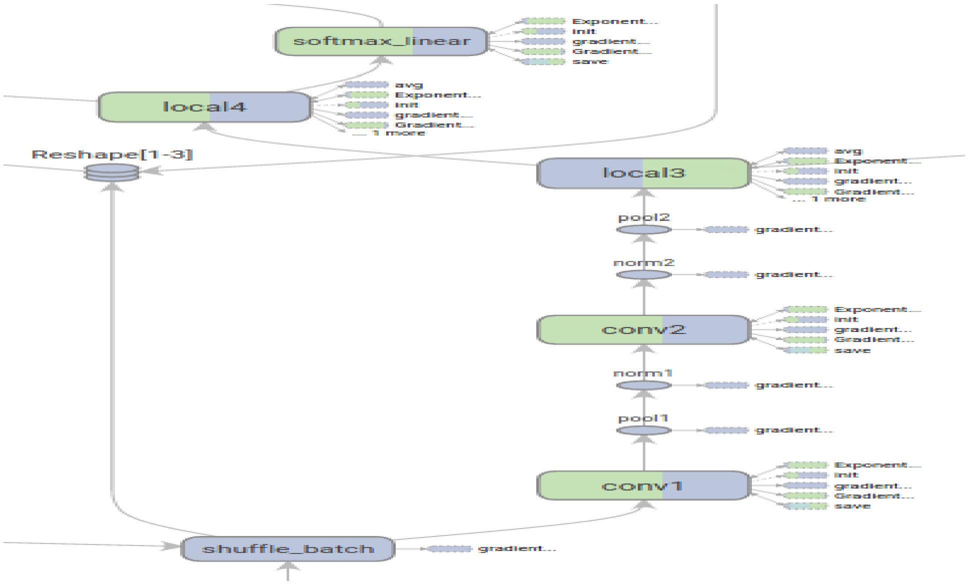

Figure 2: Example Tensorflow graph showing operation-nodes and tensor-edges. In the right panel green operations are tied to CPU and Blue operations are tied to the GPU.

repeats many times as the models parameters are trained.

For high throughput and flexibility the compute graph's operations can be assigned to various ”devices” such as CPU or GPU. These devices need not be co-located on the same machine, and during the evaluation of the graph tensors that are output from one operation that are needed on another machine or device will be transported over a bus or over a network connection. The same operation can be instantiated on different devices by instantiating a ”kernel” for the given device. The kernel is usually specified in C, C++ or CUDA-C for performance and for compatibility with the device. Once these operations and kernels are instantiated, they can be made available along with the other kernels present in Tensorflow to provide yet another building block for other researchers to use to build their own compute-graphs. Thus once we implement our high rank tensor operation(s), we can relatively easily use our operation to extend Tensorflow which makes our operations available for many other researchers in the community.

Ultimately in addition to building embedding and arbitrary tensor operations to extend computational capabilities for other machine learning researchers, we also want to produce some specific applications to help solve our surveillance problems. Since we don't know the exact details of the hardware at this point, it is unclear exactly what our capabilities will be. However I think that a basic re-implementation of existing methods like Word2Vec or other deep learning components, perhaps in a way that uses high-rank tensors would provide a familiar embedding-like method that could be used on social media data such as described in the first section of this proposal. By using our newly developed algorithm to do embedding or sentiment classification on twitter data, we can perhaps uncover new trends and mechanisms underlying the spread of disease. This would help facilitate the growing need for data analysis mechanisms in the biological and medical sciences[**?**]. Ultimately we will use these new methods to predict disease outbreaks, prevent disease, and identify social issues and concerns surrounding existing diseases and treatments with the goal of improving the social outcomes in addition to medical outcomes for patients.

3 Using machine learning to determine user personality associated with disease and treatment responses

For the third section of this proposal, I would like to extend biosurveillance measures to other disease/treatment combinations that were not covered in the PrEP section, and also take into account social network connectivity (number of followers and followees), and other forms of media related to social profiles such as images. Connectivity is important in part due to its association with certain personality traits. Previous research in psychology has determined that certain narcissistic personality traits correlate with connectivity in social networks[**?**]. In this section we propose an exploration of whether personality traits correlate with disease/treatment discussions in social media.

Because personality traits have been shown previously to correlate with profile pictures and other posted pictures[**?**], for our investigation of personality traits we wanted to use a social media platform in which users would post images of themselves. While Instagram, a social media platform that specializes in images and short videos, seemed like a great choice for this analysis, unfortunately their public API feed was deactivated indefinitely in June 2016. Facebook also does not have a free API that can be used by data mining researchers, in part because its privacy policy limits many comments to only be viewable by friends. These public API limitations left us with Twitter as the only remaining platform with millions of monthly active users who have profile pictures available for datamining. Twitter is also uniquely useful because it offers a variety of meta data including datetime, hashtag and geolocation data. Twitter also offers a public streaming API and various REST APIs that can be used to query all public tweets and unlike social media platforms like facebook, tweets are public by default. Twitter users each have a profile picture available through the REST API, and images posted in tweets can be easily filtered out based on the image format (i.e. *.jpeg or *.png).

A recent psychology study has annotated over 1 million tweets by Myers-Briggs personality type and gender[**?**]. We propose the use of this labeled tweet corpus to perform supervised label inference on our public health dataset of interest in a similar way to how we did the sentiment analysis in our PrEP-related research (see section one of this document). In this way we can transpose personality labels onto our disease-related tweets. We can also see if there are any aspects of user profile pictures that correlate with either personality or disease topics, or finally whether follower or followee counts associate with disease or personality traits. In addition using network-based semi supervised learning methods, such as graph-laplacian[**?**], we can do unsupervised learning wherein we use unlabeled twitter users's graph connections to help us better classify labeled user's personalities.

For this project I am proposing the use of several tweet corpuses. Firstly we have the tweets with labeled personality score and gender provided as an open dataset[**?**]. Next we have a corpus of general tweets that mention an infectious disease, a corpus that mentions at least one of the top 100 prescription pharmaceuticals, and a corpus that mentions at least one of the top 100 hospitals in the US. In addition to these disease related corpuses, I also want to repeat any analyses on a random "background" corpus of general tweets from Twitter to make sure that trends and patterns identified in disease related corpuses are disease-specific and not trends that hold over all tweets.

I want to use a variety of simplistic models such as linear regression and decision trees to determine the features that contribute to whether a tweet is associated with a certain personality or mentions a certain disease negatively or positively. These simple models are useful because they can more easily be interpreted. Next I want to do a variety of hypothesis tests to determine if certain diseases or personality traits correlate significantly with each other. This correlation will have to take into account whether or not the disease or personality scores are continuous or discrete, though this will be decided on at the time. Significance tests will use the background tweet corpus described above as an empirical ”null distribution”.

Taking advantage of metadata available in tweets, we can determine if there are any correlations or relevant summary statistics that correlate between disease or personality variables. Given two personality types or two sets of positive or negative disease-related tweets, we can measure the most frequent domain name prefixes in the hyper links. We can also measure if there are non-random temporal or spacial (using geolocation metadata) distribution of these tweets relative to some background distribution (hence the need for a set of background tweets).

If successful, this research would help public health officials determine the interactions between social media networks, infectious disease, and personality traits. This would continue to inform our understanding developed from research described in section 1 that attempts to determine the social aspects of disease treatment outcomes.

4 Generating novel peptides based on unsupervised latent distribution of peptides

Earlier this semester I decided that biosurvielince isn’t going to be relevant to a lot of the sequence-related and biologically related things that the other students in my graduate program are doing. With that in mind I set out to prepare a separate project for the purposes of my student seminar in Fall 2016, and I think that it may be a significant enough project to publish, or at least post in in Arxiv and include it as part of my dissertation. The concept, though developed independantly back in September 2016 is similar to a paper by Gomez-Bombarelli et al published in October 2016 titled ”Automatic chemical design using data-driven continuous representation of molecules”[**?**]. The cited paper attempts to create a generative continuous representation of small molecules for drug-related machine learning purposes. My work will propose a generative continous representation of peptides. A Variational Autoencoder (VAE) combined with a Recurrent eural Network (RNN) Sequence to Sequence (seq2seq) model is the specific model we used in our project. The model used by Gomez-Bombarelli et al also used a VAE, but used a more ”vanilla” deep neural network instead of a seq2seq network. In addition, the seq2seq network that we used took advantage of some advanced features such as an ”attention mechanism” and ”bucketing”. Our model was implemented in tensorflow.

In this project we take have taken the open uniref dataset, filtered for sequences between 10 and 100 amino acids in length, and subsampled one million peptide sequences (enough to fit into memory). One of the most important hyper parameters in our model is n_z, which denotes the size of the latent distribution that ”encodes” the peptide. By choosing n_z to be large (32) our model was easily able to encode peptides with nearly 100% accuracy. When n_z was small (2), our model was only able to encode patterns in peptides that are extremely ”highly conserved” such as the peptide beginning with a Methionine.

We hope to in the next few weeks demonstrate the utility of our seq2seq VAE peptide model. We will show a variety of figures including the relative encoding distances between the input tokens (tokens are biamino acids and encoded prior to being input into the RNN). Inspection of the embeddings will show us whether the token XY is similar to the token YX for various amino acids. Embeddings will also let us see relative substitution rates, which are not explicitly fed into the model from something like a Blossum matrix, but rather are "learned organically".

We also want to look at error (both reconstruction error and a latent error which captures the entropy of the latent distribution) and example reconstructions for a range of hyper parameter choices of n_z. We want to see if there is any accuracy performance correlation with sequence length.

Finally, we want to examine the latent distribution to see if we can see structures, or clusters that can be interpreted as "biologically meaningful". We can do this by visualizing the latent distribution in 2 dimensions using dimensionality reduction tools like tSNE or PCA. We can use a variety of clustering algorithms such as kmeans or spectral clustering to determine clusters, and or perhaps color data points in latent space by biological function. Having biological function labels would also allow us to do supervised classification, and to use latent space representation to extend this with semisupervised classification. Thus our model would provide a representation that could be shown to be useful for a variety of machine learning problems concerning peptides.

Our model provides more than just a continuous representation of peptides. By controlling the regularization paramettter (n_z) we can generate a distribution for a given input peptide of peptides that are related to the input peptide. This allows us to do say "guided evolution" where we can draw a mutated peptide from a distribution that models relative substitution rates in the uniref data set. This generated distribution allows us to perform a variety of subsequent machine learning applications on peptides. Say we have an objective function for how well a given peptide performs at binding or for drug treatment. This objective function would be filled with local optima, not differentiable, and expensive to compute (requiring experiment or molecular simulation). We could construct an an optimization strategy of this function more effective than monto carlo or exhastive search, if we use our proposed model as the distribution function in a Metropolis Hastings optimization. Such a proposed application would be similar in theory to previous work, but with a different distribution function[**?**].

One large assumption used in this project is the seq2seq error function which serves as the objective function for the whole model. Currently we used a very simple dense sequence comparison (I don't know the details, need to look this up.) but in the future we might want to extend this model by using a seq2seq error function that explicitly accounts for insertions and deletions present in biological sequences. For now, we consider this problem to be mitigated in part by focusing on relatively small peptides (less than 100 amino acids) and we assume that small conserved motifs of amino acids (which is what our model is capturing) are unlikely to contain insertions or deletions. Insertions and deletions are more likely to occur between biologically concerced motifs. Our model currently does allow for insertions and deletions in a general seq2seq way, but it does not handle them explicitly.

Broader Impacts

@articleliu2014early, title=Early experiences implementing pre-exposure prophylaxis (PrEP) for HIV prevention in San Francisco, author=Liu, Albert and Cohen, Stephanie and Follansbee, Stephen and Cohan, Deborah and Weber, Shannon and Sachdev, Darpun and Buchbinder, Susan, journal=PLoS Med, volume=11, number=3, pages=e1001613, year=2014, publisher=Public Library of Science

@articlecalabrese2015stigma, title=How stigma surrounding the use of HIV preexposure prophylaxis undermines prevention and pleasure: a call to destigmatize Truvada Whores, author=Calabrese, Sarah K and Underhill, Kristen, journal=American journal of public health, volume=105, number=10, pages=1960–1964, year=2015, publisher=American Public Health Association

@articlearnold2012qualitative, title=A qualitative study of provider thoughts on implementing pre-exposure prophylaxis (PrEP) in clinical settings to prevent HIV infection, author=Arnold, Emily A and Hazelton, Patrick and Lane, Tim and Christopoulos, Katerina A and Galindo, Gabriel R and Steward, Wayne T and Morin, Stephen F, journal=PloS one, volume=7, number=7, pages=e40603, year=2012, publisher=Public Library of Science

@articlegolub2013efficacy, title=From efficacy to effectiveness: facilitators and barriers to PrEP acceptability and motivations for adherence among MSM and transgender women in New York City, author=Golub, Sarit A and Gamarel, Kristi E and Rendina, H Jonathon and Surace, Anthony and Lelutiu-Weinberger, Corina L, journal=AIDS patient care and STDs, volume=27, number=4, pages=248–254, year=2013, publisher=Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA

@articleconrad2015community, title=Community outbreak of HIV infection linked to injection drug use of oxymorphoneIndiana, 2015, author=Conrad, Caitlin and Bradley, Heather M and Broz, Dita and Buddha, Swamy and Chapman, Erika L and Galang, Romeo R and Hillman, Daniel and Hon, John and Hoover, Karen W and Patel, Monita R and others, journal=MMWR Morb Mortal Wkly Rep, volume=64, number=16, pages=443–444, year=2015

@articleyoung2014methods, title=Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes, author=Young, Sean D and Rivers, Caitlin and Lewis, Bryan, journal=Preventive medicine, volume=63, pages=112–115, year=2014, publisher=Elsevier

@incollectionzhu2011semi, title=Semi-supervised learning, author=Zhu, Xiaojin, booktitle=Encyclopedia of machine learning, pages=892–897, year=2011, publisher=Springer

@inproceedingsguillaumin2010multimodal, title=Multimodal semi-supervised learning for image classification, author=Guillaumin, Matthieu and Verbeek, Jakob and Schmid, Cordelia, booktitle=CVPR 2010-23rd IEEE Conference on Computer Vision & Pattern Recognition, pages=902–909, year=2010, organization=IEEE Computer Society

@inproceedingsle2014distributed, title=Distributed Representations of Sentences and Documents., author=Le, Quoc V and Mikolov, Tomas, booktitle=ICML, volume=14, pages=1188–1196, year=2014

@incollectionweston2012deep, title=Deep learning via semi-supervised embedding, author=Weston, Jason and Ratle, Frédéric and Mobahi, Hossein and Collobert, Ronan, booktitle=Neural Networks: Tricks of the Trade, pages=639–655, year=2012, publisher=Springer

@articlechetlur2014cudnn, title=cudnn: Efficient primitives for deep learning, author=Chetlur, Sharan and Woolley, Cliff and Vandermersch, Philippe and Cohen, Jonathan and Tran, John and Catanzaro, Bryan and Shelhamer, Evan, journal=arXiv preprint arXiv:1410.0759, year=2014

@articlecichocki2014tensor, title=Tensor networks for big data analytics and large-scale optimization problems, author=Cichocki, Andrzej, journal=arXiv preprint arXiv:1407.3124, year=2014

@articlecichocki2014era, title=Era of big data processing: A new approach via tensor networks and tensor decompositions, author=Cichocki, Andrzej, journal=arXiv preprint arXiv:1403.2048, year=2014

@articlecichocki2016low, title=Low-Rank Tensor Networks for Dimensionality Reduction and Large-Scale Optimization Problems: Perspectives and Challenges PART 1, author=Cichocki, A and Lee, N and Oseledets, IV and Phan, AH and Zhao, Q and Mandic, D, journal=arXiv preprint arXiv:1609.00893, year=2016

@articleabadi2016tensorflow, title=Tensorflow: Large-scale machine learning on heterogeneous distributed systems, author=Abadi, Martın and Agarwal, Ashish and Barham, Paul and Brevdo, Eugene and Chen, Zhifeng and Citro, Craig and Corrado, Greg S and Davis, Andy and Dean, Jeffrey and Devin, Matthieu and others, journal=arXiv preprint arXiv:1603.04467, year=2016

@articlebuffardi2008narcissism, title=Narcissism and social networking web sites, author=Buffardi, Laura E and Campbell, W Keith, journal=Personality and social psychology bulletin, volume=34, number=10, pages=1303–1314, year=2008, publisher=Sage Publications

@articleong2011narcissism, title=Narcissism, extraversion and adolescents self-presentation on Facebook, author=Ong, Eileen YL and Ang, Rebecca P and Ho, Jim CM and Lim, Joylynn CY and Goh, Dion H and Lee, Chei Sian and Chua, Alton YK, journal=Personality and individual differences, volume=50, number=2, pages=180–185, year=2011, publisher=Elsevier

@inproceedingsplank2015personality, title=Personality Traits on TwitterorHow to Get 1,500 Personality Tests in a Week, author=Plank, Barbara and Hovy, Dirk, booktitle=Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages=92–98, year=2015

@inproceedingssindhwani2005beyond, title=Beyond the point cloud: from transductive to semi-supervised learning, author=Sindhwani, Vikas and Niyogi, Partha and Belkin, Mikhail, booktitle=Proceedings of the 22nd international conference on Machine learning, pages=824–831, year=2005, organization=ACM

@articlegomez2016automatic, title=Automatic chemical design using a data-driven continuous representation of molecules, author=Gómez-Bombarelli, Rafael and Duvenaud, David and Hernández-Lobato, José Miguel and Aguilera-Iparraguirre, Jorge and Hirzel, Timothy D and Adams, Ryan P and Aspuru-Guzik, Alán, journal=arXiv preprint arXiv:1610.02415, year=2016

@articlegiguere2013improved, title=Improved design and screening of high bioactivity peptides for drug discovery, author=Giguere, Sébastien and Laviolette, François and Marchand, Mario and Tremblay, Denise and Moineau, Sylvain and Biron, Éric and Corbeil, Jacques, journal=arXiv preprint arXiv:1311.3573, year=2013

Biographical Sketch: Patrick Breen

(a) Professional Preparation

Bowdoin College BA Biochemistry and Mathematics, Brunswick Maine, 2013
University of Georgia PhD Bioinformatics, Athens Georgia, (attending)

(b) Appointments

Oak Ridge National Laboratory Fellowship (2016)
President of Bioinformatics Graduate Student Association (2015)
University of Georgia Presidential Graduate Fellowship (2013)

(c) Products

Mining Pre-Exposure Prophylaxis Trends in Social Media. Patrick Breen, Jane Kelly, Timothy Heckman, Shannon Quinn. DSAA2016.
P2Y6 receptor antagonist, MRS2578, inhibits neutrophil activation and aggregated NET formation induced by gout-associated monosodium urate crystals. Payel Sil ... Patrick Breen ...

(d) Synergistic Activities

Interacted with the local scientific community by judging at high school science fair (2014)

Data Management Plan

Multiple filtered streams of twitter data relevant to diseases studied will be acquired using Twitter's public streaming API. This data includes tweets related to Pre Exposure Prophylaxis, Truvada, HIV, and AIDS as well as other infectious diseases, and commonly used prescription medications. In addition to twitter data we will also acquire other social media data such as Reddit or Facebook, either from an open data repository, or in the case of Reddit through the public API. Direct medical data will be acquired from Oak Ridge National Laboratory and other sources and will be used in accordance with institutional and national laws and regulations (including HIPAA) governing appropriate use of medical data.

Data will be analyzed on local researcher's desktop machines and on research server clusters including the Quinn research group cluster, the Georgia Advanced Computing Resource Center, and Oak Ridge's institutional computing resources. Code developed for analyses will be made openly available on github.com under open source licenses.

Most of the primary information acquired through Twitter's API or direct medical information cannot be shared directly due to Twitter's terms of service and federal regulations such as HIPAA. However anonymized summary information can and will be disseminated in the form of research article publications, and perhaps also through blog articles and other non-technical mediums.

Collaborators and Other Affiliations Information

Collaborators and Co-Editors

Shannon Quinn (University of Georgia)
Jane Kelly (Georgia Department of Public Health)
Timothy Heckman (University of Georgia)
Arvind Ramanathan (Oak Ridge National Laboratory)

Graduate Advisors and Postdoctoral Sponsors

Shannon Quinn (University of Georgia)

Committee Members

Shannon Quinn (University of Georgia)
Jan Mrazek (University of Georgia)
Timothy Heckman (University of Georgia)
Keith Campbell (University of Georgia)