JOURNAL OF
EMERGING INVESTIGATORS

Article

DNA-SEnet: A convolutional neural network for classifying DNA-asthma associations

Siva Bubby[1], Brianna Chrisman[2]

[1] BASIS Scottsdale, Scottsdale, Arizona

[2] Stanford University Department of Bioengineering, Stanford, California

SUMMARY

Asthma is a complex disease with a growing global prevalence whose genetic causes remain largely unexplored. The rise of next-generation sequencing has significantly augmented genetic studies in identifying asthma-associated mutations, the most common of which are single nucleotide polymorphisms (SNPs). Population-based and biochemical analyses have been used to identify novel disease-associated loci and their biological consequences; however, SNPs alone do not explain the mechanisms of asthma nor do they offer a context to evaluate candidate SNP-asthma associations. To this end, we developed a model named DNA Sequence Embedding Network (DNA-SEnet) to classify DNA-asthma associations using their genomic patterns. The hypotheses of this study are that DNA-asthma associations can be discerned through high-dimensional vector representations of DNA sequences around SNPs, that these features can be applied to determine novel SNP-asthma associations, and that this model can be generalized to predict SNP-disease associations for other complex traits. On average, this model achieved an Area Under the Curve (AUC) equaling 0.81 when learning and classifying DNA-asthma associations. Additionally, DNA-SEnet corroborated previous studies' SNP-asthma connections and proposed two novel asthma-linked loci based on their surrounding semantic properties. Moreover, DNA-SEnet effectively learned DNA-disease associations when applied to sequence data regarding coronary heart disease, type 2 diabetes mellitus, and rheumatoid arthritis. Therefore, this model can be used to identify novel disease-associated sequences across various disease types.

INTRODUCTION

Asthma is a complex polygenic disease with several subtypes and is influenced by largely unexplored hereditary and environmental factors, often making it difficult to diagnose. By 2025, nearly 400 million people globally will have asthma (1, 2). Advancements in high-throughput sequencing have unveiled several mutations associated with asthma primarily through genome-wide association studies (GWAS), which analyze the distribution of genomic variants across cases and control groups for a specified phenotype (3). These variants are mostly single nucleotide polymorphisms (SNPs), genetic mutations occurring at singular loci across a genome. SNP-asthma associations have been studied via population-based and biochemical analyses, and they offer tremendous opportunity in personalized disease prediction.

Genome-wide prediction studies (GWPS) estimate an individual's susceptibility to a disease based on their SNP composition. Feature selection reduces the high-dimensionality of GWAS data by selecting the most prominent vector-based SNP-disease associations prior to training machine learning models (4-6). While GWAS and GWPS reveal SNP-asthma and SNP-SNP associations, they are often time-consuming since they require individual whole-genome sequencing. Pooling DNA samples across populations into microarrays before sequencing circumvents these issues. Since genetic variation often causes differing minor allele frequencies (MAF) for SNPs across populations, pooled sampling identifies population-based SNP-disease associations and screens for falsely correlated loci within large databases (7-9). Similarly, analyzing MAF distributions, SNP phenotypes, and other population-specific data within case and control groups characterizes the biological impacts of novel SNP loci (10). Moreover, explorations of allelic interactions have significantly increased the number of potential asthma-linked loci. Such studies have developed models to detect epistatic interactions for asthma-associated loci, classify mutation types, and discern functional applications of various loci (11-13).

While SNPs offer insights into the genetic causes of asthma, they alone do not explain its mechanisms nor do they provide a framework to predict the disease risk of novel sequences. As more potentially asthma-associated SNPs are discovered, especially outside coding regions (14), verifying their disease associations and functions becomes increasingly difficult due to high false positive rates, low replicability, and low generalizability (3). To investigate the following three hypotheses, we developed a model named DNA Sequence Embedding Network (DNA-SEnet):

Hypothesis 1. DNA-asthma associations can be revealed through semantic similarities and distributional representations of genomic sequences centered around previously identified SNPs.

Hypothesis 2. DNA-SEnet can identify novel asthma-associated SNPs based on learned semantic features.

Hypothesis 3. DNA-SEnet can be generalized to predict SNP-disease associations for other complex traits including coronary heart disease, type 2 diabetes mellitus, and rheumatoid arthritis.

DNA-SEnet analyzed high-dimensional features of semantic patterns across GWAS loci to discern DNA-asthma associations. On average, the model significantly outperformed classical machine learning methods in both predictive robustness and computation time when learning and clas-

sifying existing SNP-asthma associations. Additionally, the model corroborated the findings of a population-based study whose purpose was to implicate particular SNP-gene pairs linked with asthma. Moreover, DNA-SEnet proposed two novel asthma-associated loci, indicating that the model can effectively apply learned semantic features when determining potential DNA-asthma connections. Finally, DNA-SEnet performed consistently well when applied to SNP data from other complex traits, thereby demonstrating the model's generalizability across various disease types.

RESULTS

Hyperparameters

This research consists of model-related and data-related hyperparameters. Model-related hyperparameters entail the various combinations of settings in DNA-SEnet and the control model used for performance comparison. We employed random grid search to test five percent of all possible hyperparameter combinations and optimize DNA-SEnet with lower computational cost. The settings which minimized the loss value on the testing set were considered as the optimal hyperparameters. If more than one combination achieved the same minimum loss value, the combination which maximized the Area Under the Curve (AUC) value, which measured predictive robustness and classification accuracy, for the testing set was selected as optimal. All hyperparameters explored in DNA-SEnet were recorded **(Table 1)**.

This study designed a series of Support Vector Machines and selected the one with the highest predictive accuracy as the control model. The most important hyperparameter for SVMs is the kernel, which transforms linearly inseparable input data into linearly separable cases in higher dimensions. Models for each kernel were individually designed and tuned. The Radial Basis Function (RBF) kernel achieved the highest AUC score on the testing dataset. Therefore, the Support Vector Machine with Radial Basis Function kernel (SVM_RBF) was used as a baseline for this experiment. All hyperparameter combinations explored across the SVMs were documented **(Table 2)**.

Word embeddings are growing increasingly popular in natural language processing; however, their effect on predicting DNA-disease associations remains largely unexplored. Specifically, the influence of *k*-mer length *k* (the number of nucleotides in a DNA fragment), stride window *s* (the number of nucleotides between the end of one *k*-mer and the beginning of the next), sequence length *L* (the total number of nucleotides in the sequence), and embedding dimension *b* (the length of the vectorized *k*-mer features) on the performance of DNA-SEnet are unknown. Therefore, these settings were treated as data-related hyperparameters and are discussed below.

DNA-SEnet Hyperparameter	Options
Convolutional Filters	16, 32, **48**, 64, 80, 96, 112
Convolutional Kernel Size	6, 7, **8**
Convolutional Stride Window	1, 2, **3**
Maximum Pooling Pool Size	**2**, 3, 4
Dense Nodes 1	8, **16**
Dense Nodes 2	4, **8**
Dense Nodes 3	2, **4**, 8

Table 1: DNA-SEnet Hyperparameters. All hyperparameters explored in tuning DNA-SEnet using random grid search. Settings that minimized the loss value on the testing set were considered optimal (bolded).

SVM_RBF Hyperparameter	Options
Kernel	**Radial Basis Function (Gaussian)**, Polynomial, Sigmoid
Cost	**0.1**, 0.01, 0.001, 0.0001, 0.00001
Degree (Polynomial kernel only)	1, 2, 3
Epsilon	**0.1**

Table 2: SVM_RBF Hyperparameters. All hyperparameters explored in SVM creation. Hyperparameters for each kernel were tuned individually. Since Radial Basis Function (RBF) kernel maximized the AUC on the testing set, the model using this kernel (named SVM_RBF) was selected as the control model for this experiment. Optimal hyperparameters in SVM_RBF are bolded.

DNA-SEnet Performance Evaluation

All SNPs used in this study had confirmed asthma associations from the GWAS Catalog, which provides diseases associations for experimentally identified SNPs (15). Here, we explored the genomic sequence-based similarities surrounding these variants to better predict the occurrence of novel SNP-asthma associations using DNA-SEnet. We demonstrate the performance of DNA-SEnet and compare it to SVM_RBF as a baseline through a series of hyperparameter experiments. The distribution of AUC scores for each model and average computational time in seconds were documented **(Table 3)**. Significant AUC values typically range from [0.5, 1], where values tending toward 0.5 indicate random association and values closer to 1 indicate greater model performance. On average, DNA-SEnet significantly outperformed SVM_RBF in classifying DNA-asthma associations and required less training time. The incorporation of deep learning in DNA-SEnet allows it to dynamically learn abstract features by adjusting more weights compared to classical machine learning methods while maintaining a fixed architecture. SVMs, however, attempt to separate data by constructing a decision boundary using input features. Increasing the number of features and samples requires longer training times. Moreover, since SVMs attempt to maximize the distance between its data points and decision boundary, they become less generalizable as they require more data points to categorize samples. Thus, this comparison demonstrates the superiority of DNA-SEnet to classical machine learning methods in both robustness of prediction and computational time when discerning sequence-based DNA-asthma associations.

Model	min AUC	mean AUC	max AUC	Training Time (s)
DNA-SEnet	0.75	0.81	0.82	166.72
SVM_RBF	0.54	0.60	0.64	2355.71

Table 3: AUC distributions and training times of DNA-SEnet and SVM_RBF. This table compares the AUC values collected from both models during various hyperparameter experiments. Additionally, the training time of each model was recorded during each experiment using Google Colab.

Sensitivity Analysis

We conducted a sensitivity analysis to study the effect of *k*-mer length *k*, stride window *s*, sequence length *L*, and embedding dimension *b* on model performance. Increasing *k* would tremendously increase the size of the *k*-mer vocabulary, but too large of a *k* value would not capture short-range semantic patterns among DNA sequences. Moreover, too small *k* values would not generate distinguishable sequence embeddings when summed. We reconstructed the *k*-mer corpus for *k* ranging from five to seven and obtained the corresponding embeddings by retraining the word embedding Global Vectors (GloVe) model (16). We found that DNA-SEnet performed slightly better with higher *k* **(Figure 1A)**.

Equation 1 shows the number of possible *k*-mers *N* calculated using genomic sequence length *L*, *k*-mer length *k*, and stride window *s* (16).

$$N = \frac{\lfloor L - k \rfloor}{s} + 1 \quad (1)$$

We found higher *s* was associated with higher AUC from DNA-SEnet **(Figure 1B)**. Equation 1 shows that the size of the *k*-mer corpus is inversely proportional to the stride window. Too large *s* could yield a lack of useful information as potential DNA fragments may be skipped over, which may negatively impact the *k*-mer embeddings. To avoid this, we limited the stride window from two to four, ensuring that $s < k$ so all sequence components were accounted for. We did not explore $s = 1$ because it yields highly similar *k*-mers with a larger corpus (16), which could inflate the quality of the embedding representations.

Next, we examined the effect of varying sequence length *L*. We constrained *L* values to 51, 101, and 151 nucleotides to ensure symmetric, localized sequences around the risk allele. Equation 1 shows that the size of the *k*-mer corpus is directly proportional to *L*. The increase in the number of unique *k*-mers could also improve the quality of the *k*-mer vectors. We observed that DNA-SEnet achieves relatively consistent AUC measurements the aforementioned DNA sequence lengths **(Figure 1C)**. For this model, while larger *L* improve the quality of the *k*-mer vectors, important vector-based DNA-asthma features become lost as more semantic information is included in the overall embeddings.

The final data-related hyperparameter is the embedding dimension *b*. We restricted *b* values to 25, 50, and 75 due to computational constraints. Larger embedding dimensions increase the GloVe model complexity by introducing more weights. DNA-SEnet reflects this improved performance with an upward trend of its AUC value **(Figure 1D)**. Moreover, since GloVe is an unsupervised algorithm, the quality of the emergent embeddings is measured using a loss value. We observed that the GloVe loss value decreased as the embedding dimension increased **(Figure 2)**, indicating better *k*-mer embeddings with higher *b*. If *b* becomes too large, however, both DNA-SEnet and GloVe become prone to overfitting.

Novel DNA-Asthma Associations

To test how DNA-SEnet classifies novel SNP-asthma associations, we applied the model to a small SNP dataset from a population study conducted by Saba *et al.*, whose purpose was to identify SNP-gene links associated with the immunological pathways of asthma (17). We found that DNA-SEnet corroborated three of the study's population-based SNP-asthma correlations and identified two new associations which Saba *et al.* did not.

Regarding the similar findings, DNA-SEnet correctly clas-

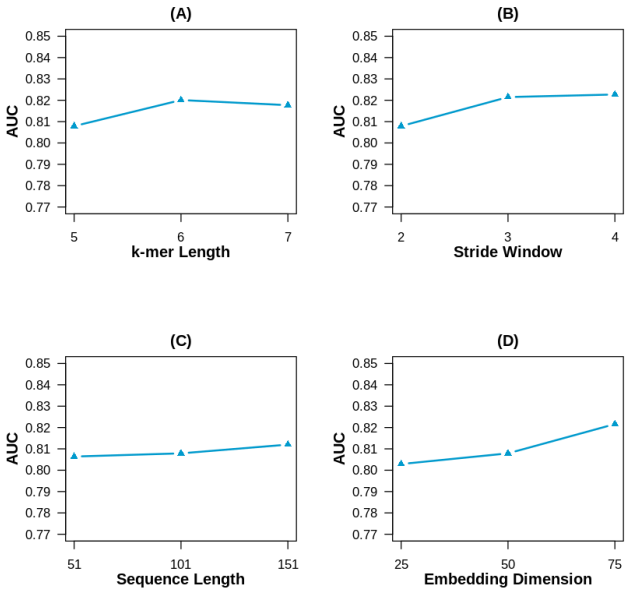

Figure 1: Line graphs displaying impacts of data-related hyperparameters on DNA-SEnet performance. A) Impact of *k*-mer length (*k*) variations on AUC. *k*-mer corpora were generated for *k* values between [5, 7] to retrain GloVe embeddings. B) Impact of stride window (s) variations on AUC. *k*-mer corpora were recreated using *s* values between [2, 4] to retrain GloVe model. C) Impact of sequence length (*L*) variations on AUC. Sequence fragments of lengths 51, 101, and 151 nucleotides were created around each locus. D) Impact of embedding dimension (*b*) variations on AUC. Embedding dimension was set to 25, 50, and 75 to retrain GloVe model. AUC values from DNA-SEnet for each data-related hyperparameter were plotted.

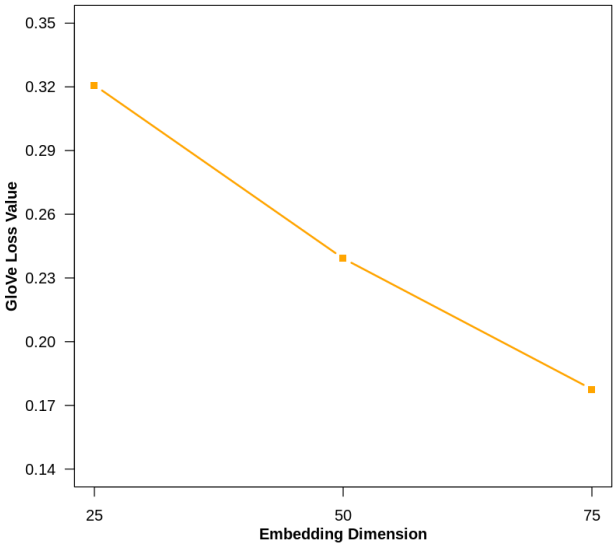

Figure 2: Impact of embedding dimension *b* on GloVe model loss value. Line graph showing respective loss values for *b* of 25, 50, and 75. The GloVe model was retrained for each *b* value and optimized over 50 iterations. The final loss values were recorded and plotted.

sified *rs1131882* on the *TBXA2R* gene as asthma-associated. Moreover, while Saba *et al.* identified *rs2280091* on the *ADAM33* gene within the population as asthma-associated, DNA-SEnet found a higher semantic connection for *rs543749*, a different SNP along the *ADAM33* gene. Saba *et al.* also proposed *rs2583476* on the *FCER1B* gene as predominantly linked with male asthma cases (17). While DNA-SEnet did not discern gender-based associations, it correctly classified this variant as well.

Additionally, DNA-SEnet classified *rs1042713* and *rs1799983* on the *ADRB2* and *NOS3* genes, respectively, as asthma-linked. Previous literature has implicated these two genes as heavily asthma-associated; however, they have primarily been explored through population studies (17-22). In this study, DNA-SEnet implicates the aforementioned SNPs based on their surrounding genomic patterns. The SNPs and genes classified as asthma-associated by DNA-SEnet were documented **(Table 4)**. This experiment indicates that DNA-SEnet can identify semantic properties for known asthma-linked mutations and apply them to predict novel loci.

Applying DNA-SEnet to Other Diseases

Although developed to identify asthma-associated loci, the DNA-SEnet architecture can be trained for a variety of diseases using the genomic sequences around significant loci and an equal number of negative samples (healthy sequences) to avoid class imbalance. We applied DNA-SEnet to fragments containing SNPs associated with coronary heart disease (AUC = 0.8089), type 2 diabetes mellitus (AUC = 0.8081) , and rheumatoid arthritis (AUC = 0.8177) (**Figure 3,** blue bars). SVM_RBF is used as a baseline across all diseases (**Figure 3,** red bars). Our results demonstrate that DNA-SEnet performs consistently well when classifying each SNP-disease association and outperforms SVM_RBF in each case. This consistent performance elucidates that DNA-SEnet can predict SNP-disease associations across myriad genetic diseases.

DISCUSSION

This study proposes DNA-SEnet, a convolutional neural network using *k*-mer-based genomic sequence embeddings to detect asthma-associated loci. First, we extracted *k*-mer embeddings by pre-training the unsupervised GloVe algorithm. We calculated sequence embeddings as the sum of individual *k*-mer embeddings and used these for feature representation to avoid high-dimensional data from one-hot encoding. This method preserved computational resources during model training while helping DNA-SEnet analyze vector similarities across sequences. Second, we applied convolutional neural networks in DNA-SEnet to improve hierarchical feature learning of sequence embeddings. We found that DNA-SEnet significantly outperformed the popular Support Vector Machine when identifying asthma-associated loci. Also, we demonstrated the robustness of the model through data-related hyperparameter experiments.

Furthermore, we showed that DNA-SEnet is capable of classifying novel associations of candidate mutations. The *ADRB2* gene is expressed primarily on smooth muscles of the bronchi and cardiac myocytes and has been extensively correlated with asthma in terms of confirmed and pending SNPs (23). Candidate markers of this gene showed significant blood-based concentrations in patients with nocturnal asthma compared to those with non-nocturnal asthma or no asthma (24). Moreover, the *NOS3* gene exemplifies the complex nature of asthma through an interaction with environmental conditions linked to the disease (25, 26). This experiment demonstrates the ability of DNA-SEnet to apply learned hierarchical features when classifying potential asthma-associated loci based on semantic properties rather than population-based characteristics.

Additionally, we applied DNA-SEnet to sequences surrounding SNPs linked with other common complex diseases including coronary heart disease, type 2 diabetes mellitus, and rheumatoid arthritis. This experiment corroborated the generalizability of DNA-SEnet to predict other complex disease associations. This extension emphasizes the role of natural language processing and deep learning in genomic

Mutation	Gene
rs1042713	*ADRB2*
rs1799983	*NOS3*
rs2583476	*FCER1B*
rs1131882	*TBXA2R*
rs543749	*ADAM33*

Table 4: rsID and genes for all novel asthma-associated SNP candidates identified by DNA-SEnet. DNA sequences around each SNP Saba *et al.* classified as asthma-associated were extracted from BEDTools and ran through DNA-SEnet. The model classified the above mutation as asthma-associated based on their semantic properties.

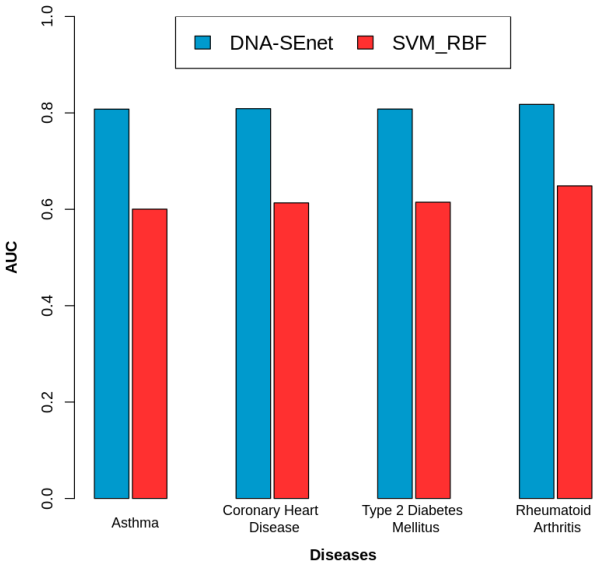

Figure 3: DNA-SEnet performance on classifying SNP-disease associations of other complex traits. Bar graph of AUC from DNA-SEnet (blue bars) on asthma, coronary heart disease, type 2 diabetes mellitus, and rheumatoid arthritis. SVM_RBF (red bars) was used for comparison. Sequence fragments for SNPs of each disease were created to retrain GloVe and DNA-SEnet.

sequence analysis and disease prediction.

DNA-SEnet can also be applied in many areas of medical research. First, it can supplement GWAS to unveil potential mutation-phenotype associations across myriad disease types. Second, it can explore disease-disease interactions via unsupervised learning and sequence embedding cluster analyses. Similarly, if the model can analyze disease-specific SNPs without being trained on them, it would detect common genomic features between its current disease classifications and the additional disease. Third, it can subtype diseases using locus heterogeneity. Understanding the relationship between a disease's diverse subphenotypes and genetics would augment this application (27). Ultimately, analyzing the effects of disease-specific genomic patterns on protein production, transcription factor binding, gene regulation, chromatin accessibility, and other biological functions can yield greater insights into the causes and progression of complex traits.

However, there are limitations to this model. Namely, DNA-SEnet includes SNPs from the GWAS Catalog, but does not account for other mutations (including insertion, deletion, or genetic amplification) or sex-based associations (28). Additionally, DNA-SEnet focuses on localized genomic patterns to generate global statistics for a *k*-mer corpus. This accounts for semantic similarities across epistatic sites, but not their biological interactions. Incorporating linkage disequilibrium data would help overcome this limitation (29). Nonetheless, DNA-SEnet shows promise in further understanding the relationships between genomic sequences and genetic diseases, which can yield greater insights into the biological mechanisms of complex diseases including asthma.

MATERIALS AND METHODS

Datasets and Data Augmentation

GWAS Catalog is a publicly available database containing collections of SNP-phenotype associations across several disease types analyzed in population studies and genomic sequencing (15). For this study, the asthma dataset was downloaded. SNPs were filtered to remove all entries missing mutated alleles and chromosomal loci. For all entries missing either the SNP or locus, the missing information was manually extracted from SNPedia (30) or dbSNP (31). Each genomic locus was expanded to symmetric sequence lengths—the *L* values—about the risk allele and inputted into BEDTools (32), a software used for genomic arithmetic, alongside its corresponding chromosome to extract the appropriate DNA sequences from the *GRCh38* reference genome (33). Sequences directly extracted from the reference genome had no disease associations because they lacked their corresponding risk alleles, and were thus classified as "healthy." Risk alleles were substituted into sequences at their corresponding loci to generate the disease-associated sequences.

We used data augmentation to simulate minor nucleotide variations around risk alleles so DNA-SEnet could identify short- and long-range patterns when classifying sequence-based disease associations. The primary difference between healthy and disease-associated sequences was the central allele. Augmented sequences preserved their corresponding central allele and classification. Appending slightly modified versions of all sequences increased the total number of sequences. Additionally, the reverse complements of all possible sequences were fed into the model alongside the original to account for double stranded DNA variations, thereby doubling the number of sample sequences (34, 35).

Embedding Representations

The term *k*-mer refers to pieces of genomic sequences obtained using a fragment length *k*, a stride window *s*, and a genomic sequence length *L*, as shown in Equation 1. For this study, all *k*-mers produced from a genomic sequence were strung together as a *k*-mer sequence indexed by positive integers $j \in [1, N]$, where *N* represents the number of *k*-mers obtained from a sequence. To generate embedding vectors, we used the R implementation of the unsupervised Global Vectors (GloVe) algorithm in the text2vec package to analyze global and local statistics of individual *k*-mers (36). By treating *k*-mers within a *k*-mer sequence as words within a sentence, we trained an embedding model which converted each *k*-mer k_j into a *b*-dimensional vector kv_j. All embedding vectors culminated into a matrix of dimension *n* x *b*, where *n* is the number of unique *k*-mers and *b* is the specified embedding dimension. All sequence embedding vectors *SEv* were calculated as the sum of their *k*-mer vectors, as shown in Equation 2. These sequence embeddings culminated into a matrix of dimension n_s x *b*, where n_s is the number of sample sequences and *b* is the embedding dimension.

$$SEv = \sum_{j=1}^{N} kv_j \quad (2)$$

Classifying DNA-Asthma Associations

To investigate the semantic properties of DNA-asthma associations, we used Google Colab and the Keras platform in R to analyze the aforementioned sequence embeddings. We designed DNA-SEnet to accomplish this goal. DNA-SEnet employed convolutional neural networks (CNNs) to adaptively learn and generalize hierarchical spatial features (37). CNNs helped DNA-SEnet learn short-range embedding associations during convolution and long-range associations during classification. Hyperparameters were tuned using random grid search.

Additionally, we created SVM_RBF as a baseline when evaluating DNA-SEnet. SVMs apply statistical learning theory to classification problems by constructing hyperplanes to separate data in *f*-dimensional space, where *f* is the number of features. For linearly inseparable cases in *f* dimensions, SVMs convert the data to linearly separable form in *m*-dimensional space, where $m > f$ (38). This model serves as an effective comparison for DNA-SEnet because both models are capable of high-dimensional feature extraction and reduction for classification problems. Moreover, both models train to converge their cost functions, meaning that optimization can be controlled using hyperparameters and will improve with training. We created SVMs for Linear, Polynomial, Sigmoid, and Radial Basis Function kernels and tuned them individually using the publicly available e1071 package within R.

We measured each model's ability to rank patterns distinguishing the binary classifications using AUC. Receiver Operating Characteristics (ROC) Curves plot the true positive rate against the false positive rate for a given decision threshold. We extracted the threshold yielding the highest accuracy before evaluating each model.

Identifying Novel Loci

We downloaded the dataset of candidate SNP-asthma connections from Saba *et al.* (17) to measure the ability of DNA-SEnet to identify novel asthma-linked loci. Corresponding healthy and risk sequences were generated for binary classification. DNA-SEnet predicted the probability of each association using its vector-based semantic properties with a probability threshold of 0.5 to discern asthma associations.

Applying DNA-SEnet to Other Diseases

We downloaded the coronary heart disease, type 2 diabetes mellitus, and rheumatoid arthritis datasets from the GWAS Catalog to apply DNA-SEnet to additional disease types including cardiovascular, metabolic, and immunological, respectively. We retrained GloVe with reconstructed *k*-mer corpora and used SVM_RBF as a baseline for each disease type.

Code and data for this study can be found here: https://github.com/sivab468/DNA-SEnet.

ACKNOWLEDGEMENTS

I am grateful to Polygence for this wonderful research opportunity.

Received: May 21, 2021
Accepted: August 13, 2021
Published: November 16, 2021

REFERENCES

1. Ntontsi, Polyxeni, *et al*. "Genetics and Epigenetics in Asthma." *International Journal of Molecular Sciences*, vol. 22, no. 5, 2021, p. 2412., doi:10.3390/ijms22052412.
2. Masoli, Matthew, *et al*. "The Global Burden of Asthma: Executive Summary of the GINA Dissemination Committee Report." *Allergy*, vol. 59, no. 5, 6 Apr. 2004, pp. 469–478., doi:10.1111/j.1398-9995.2004.00526.x.
3. Quezada, Héctor, *et al*. "Omics-Based Biomarkers: Current Status and Potential Use in the Clinic." *Boletín Médico Del Hospital Infantil De México (English Edition)*, vol. 74, no. 3, 2017, pp. 219–226., doi:10.1016/j.bmhime.2017.11.030.
4. Gaudillo, Joverlyn et al. "Machine learning approach to single nucleotide polymorphism-based asthma prediction." *PloS one*, vol. 14,12 e0225574. 4 Dec. 2019, doi:10.1371/journal.pone.0225574
5. Xu, Mousheng, *et al*. "Genome Wide Association Study to Predict Severe Asthma Exacerbations in Children Using Random Forests Classifiers." *BMC Medical Genetics*, vol. 12, no. 1, 30 June 2011, doi:10.1186/1471-2350-12-90.
6. Tomita, Yasuyuki, *et al*. "Artificial Neural Network Approach for Selection of Susceptible Single Nucleotide Polymorphisms and Construction of Prediction Model on Childhood Allergic Asthma." *BMC Bioinformatics*, vol. 5, no. 1, 1 Sept. 2004, p. 120., doi:10.1186/1471-2105-5-120.
7. Craig, David W, et al. "Identification of Disease Causing Loci Using an Array-Based Genotyping Approach on Pooled DNA." *BMC Genomics*, vol. 6, no. 1, 30 Sept. 2005, doi:10.1186/1471-2164-6-138.
8. Ricci, Giampaolo, *et al*. "Pooled Genome-Wide Analysis to Identify Novel Risk Loci for Pediatric Allergic Asthma." *PLoS one*, vol. 6, no. 2, 16 Feb. 2011, doi:10.1371/journal.pone.0016912.
9. Castro-Giner, Francesc, *et al*. "A Pooling-Based Genome-Wide Analysis Identifies New Potential Candidate Genes for Atopy in the European Community Respiratory Health Survey (ECRHS)." *BMC Medical Genetics*, vol. 10, no. 1, 6 Dec. 2009, doi:10.1186/1471-2350-10-128.
10. Xu, Yan, *et al*. "Model-Based Clustering for Identifying Disease-Associated SNPs in Case-Control Genome-Wide Association Studies." *Scientific Reports*, vol. 9, no. 1, 23 Sept. 2019, doi:10.1038/s41598-019-50229-6.
11. Han, Bing, *et al*. "FEPI-MB: Identifying SNPs-Disease Association Using a Markov Blanket-Based Approach." *BMC Bioinformatics*, vol. 12, no. Suppl 12, 24 Nov. 2011, doi:10.1186/1471-2105-12-s12-s3.
12. Han, Bing, *et al*. "Genetic Studies of Complex Human Diseases: Characterizing SNP-Disease Associations Using Bayesian Networks." *BMC Systems Biology*, vol. 6, no. Suppl 3, 7 Dec. 2012, doi:10.1186/1752-0509-6-s3-s14.
13. Neville, Matthew D., *et al*. "Identification of Deleterious and Regulatory Genomic Variations in Known Asthma Loci." *Respiratory Research*, vol. 19, no. 1, 12 Dec. 2018, doi:10.1186/s12931-018-0953-2.
14. Höglund, Julia, *et al*. "Improved Power and Precision with Whole Genome Sequencing Data in Genome-Wide Association Studies of Inflammatory Biomarkers." *Scientific Reports*, vol. 9, no. 1, 14 Nov. 2019, doi:10.1038/s41598-019-53111-7.
15. Buniello A, *et al*. "The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019." *Nucleic Acids Research*, vol. 47. *GWAS Catalog*. https://www.ebi.ac.uk/gwas/
16. Shen, Zhen, *et al*. "Recurrent Neural Network for Predicting Transcription Factor Binding Sites." *Scientific Reports*, vol. 8, no. 1, 15 Oct. 2018, doi:10.1038/s41598-018-33321-1.
17. Saba, Nusrat, *et al*. "Single Nucleotide Polymorphisms in Asthma Candidate Genes TBXA2R, ADAM33 FCER1B and ORMDL3 in Pakistani Asthmatics a Case Control Study." *Asthma Research and Practice*, vol. 4, no. 1, 22 Mar. 2018, doi:10.1186/s40733-018-0039-4.
18. Basu, Kaninika, *et al*. "Adrenergic β2-Receptor Genotype Predisposes to Exacerbations in Steroid-Treated Asthmatic Patients Taking Frequent Albuterol or Salmeterol." *Journal of Allergy and Clinical Immunology*, vol. 124, no. 6, 5 Oct. 2009, doi:10.1016/j.jaci.2009.07.043.
19. Coto-Segura, Pablo, *et al*. "Influence of Endothelial Nitric Oxide Synthase Polymorphisms in Psoriasis Risk." *Archives of Dermatological Research*, vol. 303, no. 6, 4 Jan. 2011, pp. 445–449., doi:10.1007/s00403-011-1129-9.
20. Hizawa, N., *et al*. "Genetic Polymorphisms at FCER1B and PAI-1 and Asthma Susceptibility." *Clinical Experimental Allergy*, vol. 36, no. 7, 31 Jan. 2006, pp. 872–876., doi:10.1111/j.1365-2222.2006.02413.x.
21. Kim, J-H., *et al*. "TBXA2R Gene Polymorphism and Responsiveness to Leukotriene Receptor Antagonist in Children with Asthma." *Clinical & Experimental Allergy*, 21 Nov. 2007, doi:10.1111/j.1365-2222.2007.02874.x.
22. Liang, Siqiao, *et al*. "A Disintegrin and Metalloprotease 33 (ADAM33) Gene Polymorphisms and the Risk of Asthma: A Meta-Analysis." *Human Immunology*, vol. 74, no. 5, 2013, pp. 648–657., doi:10.1016/j.humimm.2013.01.025.
23. Danielewicz, Hanna. "What the Genetic Background

of Individuals with Asthma and Obesity Can Reveal: Is β2-Adrenergic Receptor Gene Polymorphism Important?" *Pediatric Allergy, Immunology, and Pulmonology*, vol. 27, no. 3, 16 Sept. 2014, pp. 104–110., doi:10.1089/ped.2014.0360.
24. Online Mendelian Inheritance in Man, OMIM. Johns Hopkins University, Baltimore, MD. MIM Number: 600807: {08 Mar 2020}: . World Wide Web URL: https://omim.org/
25. Spanier, Adam J., *et al*. "Environmental Exposures, Nitric Oxide Synthase Genes, and Exhaled Nitric Oxide in Asthmatic Children." *Pediatric Pulmonology*, vol. 44, no. 8, 14 July 2009, pp. 812–819., doi:10.1002/ppul.21071.
26. Van's Gravesande, Karin Storm, *et al*. "Association of a Missense Mutation in theNOS3Gene with Exhaled Nitric Oxide Levels." *American Journal of Respiratory and Critical Care Medicine*, vol. 168, no. 2, 1 May 2003, pp. 228–231., doi:10.1164/rccm.200212-1491bc.
27. Dahl, Andy, and Noah Zaitlen. "Genetic Influences on Disease Subtypes." *Annual Review of Genomics and Human Genetics*, vol. 21, no. 1, 31 Aug. 2020, pp. 413–435., doi:10.1146/annurev-genom-120319-095026.
28. Hendrickson, H., *et al*. "Amplification-Mutagenesis: Evidence That 'Directed' Adaptive Mutation and General Hypermutability Result from Growth with a Selected Gene Amplification." *Proceedings of the National Academy of Sciences*, vol. 99, no. 4, 19 Feb. 2002, pp. 2164–2169., doi:10.1073/pnas.032680899.
29. Cordell, H. J. "Epistasis: What It Means, What It Doesn't Mean, and Statistical Methods to Detect It in Humans." *Human Molecular Genetics*, vol. 11, no. 20, 1 Oct. 2002, pp. 2463–2468., doi:10.1093/hmg/11.20.2463.
30. Cariaso, Michael, and Greg Lennon. "SNPedia: a Wiki Supporting Personal Genome Annotation, Interpretation and Analysis." *Nucleic Acids Research*, vol. 40, no. D1, 2 Dec. 2011, doi:10.1093/nar/gkr798.
31. Sherry, S. T. *et al*. "DbSNP: the NCBI Database of Genetic Variation." *Nucleic Acids Research*, vol. 29, no. 1, 1 Jan. 2001, pp. 308–311., doi:10.1093/nar/29.1.308.
32. Quinlan, Aaron R., and Ira M. Hall. "BEDTools: a Flexible Suite of Utilities for Comparing Genomic Features." *Bioinformatics*, vol. 26, no. 6, 15 Mar. 2010, pp. 841–842., doi:10.1093/bioinformatics/btq033.
33. Schneider, Valerie A. *et al*. "Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly." *Genome Research* vol. 27,5 (2017): 849-864. doi:10.1101/gr.213611.116.
34. Park, Sungjoon, *et al*. "Enhancing the Interpretability of Transcription Factor Binding Site Prediction Using Attention Mechanism." *Scientific Reports*, vol. 10, no. 1, 7 Aug. 2020, doi:10.1038/s41598-020-70218-4.
35. Cao, Zhen, and Shihua Zhang. "Simple Tricks of Convolutional Neural Network Architectures Improve DNA–Protein Binding Prediction." *Bioinformatics*, vol. 35, no. 11, 1 June 2019, pp. 1837–1843., doi:10.1093/bioinformatics/bty893.
36. Pennington, Jeffrey, *et al*. "Glove: Global Vectors for Word Representation." *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Jan. 2014, pp. 1532–1543., doi:10.3115/v1/d14-1162.
37. Indolia, Sakshi, *et al*. "Conceptual Understanding of Convolutional Neural Network- A Deep Learning Approach." *Procedia Computer Science*, vol. 132, 2018, pp. 679–688., doi:10.1016/j.procs.2018.05.069.
38. Cai, C.Z. "SVM-Prot: Web-Based Support Vector Machine Software for Functional Classification of a Protein from Its Primary Sequence." *Nucleic Acids Research*, vol. 31, no. 13, 1 July 2003, pp. 3692–3697., doi:10.1093/nar/gkg600.