Iowa Initiative for Artificial Intelligence

Final Report

Project title:	Predicting effects of Noncoding Variants associated with Breast Cancer Susceptibility using Machine Learning
Principal Investigator:	Robert A. Cornell
Prepared by (IIAI):	Kang P. Lee
Other investigators:	Ronald J Weigel; Lira Pi
Date:	February 23, 2021

Were specific aims fulfilled:	Yes
Readiness for extramural proposal?	No
If yes … Planned submission date	
Funding agency	
Grant mechanism	
If no … Why not? What went wrong?	Good progress has been made, but additional work enabled by this pilot is necessary to have sufficient preliminary data for an external proposal. Please see section entitled, "**Ideas/aims for future extramural project"**

Brief summary of accomplished results:

The project aims to identify the enhancers that are associated with elevated risk for breast cancer. From the processed data set of 1,401 positive examples and 1,401 negative examples, we identified top 30,000 most frequent k-mers from the positive set and the negative set, respectively, where k ranges from 7 to 13, and used the frequency values of the k-mers in the union of the two sets of k-mers as the features for prediction. We applied different classification algorithms including k-Nearest Neighbors (k-NNs), logistic regressions, decision trees, random forest, linear/kernelized Support Vector Machines (SVMs), Neural Networks, and LSTM. The highest prediction accuracy we achieved was 95.1% using neural networks on the 11-mers.

Research report:

Aims (provided by PI):

1. To train state-of-the-art machine-learning-based classifiers on a set of breast epithelium enhancers, multiple classifiers will be deployed, and their performances will be compared.
2. There are four variables: training predictors (X) to train classifiers, binary outcomes (Y) to be predicted by the classifiers, testing predictors (X_positive and X_negative) to compute predicted chromatin effects of variants (SNPs) using the trained classifiers.

Data:

We started with a data set of 1,401 gene sequences of length 500 bp that are labeled with "positive" and a data set of 2,796 gene sequences of the equal length labeled with "negative". The positive sequences are defined using biochemical markers. The positive set were candidate enhancers in MCF7 breast cancer cells defined by three features: 1) chromatin is accessible in MCF7 cells, as defined by Dnase hypersensitive sites-Seq (DHS) data (DHS data from MCF7 cells: SRX1161152 , SRX116115 3), 2) chromatin is marked by histone H3 K27 acetylation, a mark of active enhancers, in MCF7 cells (H3K27Ac ChIP-Seq from MCF7 cells: SRX3083139) and 3) the elements are in physical contact gene promoters (average HiC data, in supplemental data of in Fulco et al 2019 PMID: 31784727)). The negative set are elements randomly selected from the human genome, that do not overlap members of the positive set, but that are matched to the positive set in GC content, length, and repeat regions (Ghandi et al., 2014, PMID: 25033408).

To create a balanced data set for training, we reduced the size of the negative set to 1,401 by random selection, so that the final data set had the equal size of positive and negative examples. To derive as many different types of features for prediction from the sequences as we could, we identified a) the frequencies of the four base pairs A, C, G, and T, respectively, b) the ratios of one base pair to another, c) all 1-mers, all 2-mers, and all 3-mers, and d) the sequence embeddings. We further identified top n most frequent 9-mers from the positive set and the negative set, respectively, where n includes 3,000, 10,000, 30,000, and 100,000, and got the frequency values of the 9-mers in the union of the two sets of k-mers. We finally identified top 30,000 most frequent k-mers from the positive set and the negative set, respectively, where k includes 7, 9, 10, 11, 12, and 13 and got the frequency values of the k-mers in the union of the two sets of k-mers.

Artificial intelligence/Machine Learning Approach:

Using all the features listed for prediction, we built binary classifiers that leveraged a variety of traditional classification algorithms including k-Nearest Neighbors (k-NNs), logistic regressions, decision trees, random forest, linear/kernelized Support Vector Machines (SVMs), Neural Networks, and LSTM (Long Short-Term Memory).

Experimental methods, validation approach:

For each model, we randomly split the whole data set into 75% of a training set and 25% of a test set. We checked the precision, recall, f1-score, and accuracy for each model. Only the accuracy values are listed below in the Results section. For each setting, we used, when applicable, 5-fold cross validation for parameter optimization.

Results:

The highest prediction accuracy we achieved was 95.1% using neural networks applied to strings of 11-mers.

Below is the performance summary. The green cells in the table indicate the highest accuracy for the corresponding set of features. The blue cells show that those simple features such as counts, ratios, and sequence embeddings did not work well, not yielding impressive results. Judging from the results in yellow cells, we learned that selecting the top 30,000 most frequent 9-mers from the positive and negative sets and then taking the union of the two sets of 9-mers yielded the best performance. Based on that finding, we further experimented with 7-mers, 9-mers, 10-mers, 11-mers, 12-mers, and 13-mers – determining that 11-mers resulted in the highest achieved performance of 95.1%.

Features	k-NNs	Logistic Regression	Decision Trees	Random Forest	Linear SVMs	Kernelized SVMs	Neural Networks	LSTM
counts (4)	0.734	0.769	0.663	0.734	0.767	0.767	0.763	
ratios (16)	0.733	0.774	0.659	0.738	0.772	0.764	0.769	
unigrams (500)	0.711	0.7	0.618	0.755	0.717	0.754	0.694	
bigrams (499)	0.71	0.704	0.604	0.755	0.714	0.757	0.689	
trigrams (498)	0.709	0.703	0.635	0.755	0.716	0.758	0.686	
counts + ratios	0.731	0.769	0.659	0.734	0.772	0.765	0.762	
uni + bi + tri	0.713	0.7	0.645	0.756	0.715	0.757	0.687	
counts + ratios + uni + bi + tri	0.652	0.713	0.677	0.755	0.723	0.757	0.715	
co-counts (7984)	0.769	0.732	0.695	0.784	0.706	0.776	0.757	
counts + ratios + co-counts	0.769	0.737	0.704	0.774	0.707	0.776	0.764	
sequence embedding	0.774	0.789	0.696	0.791	0.797	0.798	0.79	0.74
Embedding + counts + ratios	0.762	0.799	0.707	0.798	0.794	0.802	0.789	
9-mers	0.494	0.506	0.474	0.509	0.478	0.485	0.518	
Counts of top-3,000 most frequent 9-mers in the positive OR negative set	0.586	0.692	0.621	0.659	0.699	0.7	0.715	
Counts of top-10,000 most frequent 9-mers in the positive OR negative set	0.639	0.799	0.629	0.728	0.799	0.802	0.83	
Counts of top-30,000 most frequent 9-mers in the positive OR negative set	0.693	0.832	0.593	0.746	0.833	0.82	0.867	
Counts of top-100,000 most frequent 9-mers in the positive OR negative set	0.719	0.78	0.612	0.779	0.779	0.689	0.822	
Counts of top-30,000 most frequent 7-mers in the positive OR negative set	0.642	0.708	0.601	0.705	0.708	0.696	0.72	
Counts of top-30,000 most frequent 9-mers in the positive OR negative set	0.693	0.832	0.832	0.74	0.833	0.82	0.867	
Counts of top-30,000 most frequent 10-mers in the positive OR negative set	0.649	0.863	0.606	0.699	0.879	0.736	0.934	

Counts of top-30,000 most frequent 11-mers in the positive OR negative set	0.645	0.86	0.633	0.676	0.867	0.643	**0.951**	
Counts of top-30,000 most frequent 12-mers in the positive OR negative set	0.603	0.857	0.645	0.66	0.864	0.618	0.927	
Counts of top-30,000 most frequent 13-mers in the positive OR negative set	0.532	0.857	0.631	0.74	0.892	0.612	0.919	

To identify important features, we proceeded with feature importance from the neutral network setting that yielded the best performance. Neural networks do not provide an explicit way of calculating feature importance, while logistic regression, decision trees, and random forest do. We therefore simply chose logistic regression for feature importance, as the performance of the other two was not as good (0.86 vs. 0.633 vs. 0.676, which are highlighted in red in the table). Below is the feature importance result. The list shows the top 100 features (i.e., 11mers) that have the largest absolute importance values, out of all the 53,044 features. Notice that the importance values can be either positive or negative. The positive scores indicate a feature that predicts class 1 (positive), whereas the negative indicate a feature that predicts class 0 (negative). Again, their absolute values are sorted in descending order.

11-mers	Importance
TTTGTTTGTTT	0.4326
GGGGGGGGGGG	0.3408
ACTTTTTTTTT	0.2956
GCTGTGCGCCA	0.2921
GAGGCTGAGGC	-0.2907
AGACAGAGTCT	0.2899
CACTCCAGCCT	0.2809
AACCTGGGAGG	0.2802
GCGCCTGTAAT	0.2745
CCAGCAGAGGG	0.2709
ACAGAGTGAGA	0.2684
CAGCACTTTGG	0.2652
AAAAAAATAAA	-0.2619
AAATATTTGTT	0.2589
AATATTTGTTG	0.2586
AAAAGAAAAAA	-0.2572
CTTTTTTTTTT	0.2557
GGGCACCCCTC	0.2549
TGGTGGCAGGG	0.2532
TTTTTTTTGA	0.2525
TTTTGTTTGTT	0.2512
CTAAAAATACA	-0.2506
AGCCTCACTCA	0.2482
CTTTTTTTTTG	0.2474
AAAAACAAAGA	0.2463
GAGACAGAGTC	0.2444
TCCACCCGCCT	0.2442
ATGGAGTCTCA	0.2441
ACCACTGCACT	0.2441
GAAAGTGCTAA	0.2429

GGGCAGATCAC	-0.2429
GCAGGCCAGGG	0.2419
GAGGCGGAGGC	0.2416
AAAAAAAAAAT	0.2386
CACAGGCACAT	0.2377
AAAAAAGAAAG	0.2362
CTGAATTCATC	0.2349
GAGTTTTGTTT	0.2333
GGCAGGGCGCG	0.2315
TGGAAGAGCTG	0.2302
AGCCACCGCGC	0.2288
CCTGTTTGTTT	0.2277
TCATGCCTGTA	0.2272
GTTTGTTTGTT	0.2268
AACCAATGCAC	0.2266
ATAAAGTTTAA	0.2258
GAACCTGGGAG	0.2252
TTTTGTTTTTT	0.2243
CCCACCACCAC	-0.2223
ACCCGGGAGGC	0.2221
GGGAAATGTAG	-0.2218
GTGTGTGTGCT	0.221
GAAACTGAGGC	-0.2196
CCCTGCCAGCC	0.2193
GAGCCCGGCTC	0.2191
AATAAATATTT	0.2191
AGGCTGAGGCA	-0.2189
GCCACCGCGCC	0.2185
AGGGGGCGGGC	0.2179
AAGGGGGCGGG	0.2179
AAAGGGGGCGG	0.2179
GTAATCCCAGC	0.2173
TCTGTGTGTGT	0.2172
TGTGTATCTAA	0.217
CCACCGCGCCC	0.2168
TTTTTTTTCT	0.2167
GGGGGCGGGGC	0.2162
GCCTGTTTGTT	0.2159
GTCCCTCCCCA	0.2158
AAAAAAAAGAA	-0.2145
AAAACAAAAAA	0.2143
AGGAGGCAGAG	-0.2126
GAGCCACTGTG	-0.2122
TTTGTTTTTTG	0.2112
AAGGCCAAATT	0.211
TTGGAAGAGCT	0.2101
ATTGGAAGAGC	0.2101
CATTGGAAGAG	0.2101
TTTCTTTCTGT	0.209
CATGCCTGTAA	0.2088
AAAAAAAAATT	0.2085
AACATGGCAAA	-0.2084
GATTTTATCTG	0.207
GGTGGCAGGCG	0.2066
GGAGGCAGAGG	-0.2062
CCTGGTGCTGT	0.206

AATATTTATTT	0.2059
CCGGGAGGCGG	0.2057
GGCGGGGGCGG	0.2055
GCAGAGCCCAG	0.2053
TTGTATTTTTG	0.2051
CTGAGGCGGGA	0.2048
CTCATTTTCCT	0.2045
GAGTGGGGTTC	0.2042
AGGTCACAAAG	0.2041
TCCTCTCTGTG	0.2037
TGAGAAAGCAA	0.2036
GGGACTACAGG	-0.2033
ATTTTCTTTTT	-0.2033
CTCTACTAAAA	-0.2023

Ideas/aims for future extramural project:

Common inherited DNA variants called single nucleotide polymorphisms (SNPs) are associated with elevated risk for cancer, including breast cancer, but the biological mechanisms underlying this risk are unknown. Understanding these mechanisms may be of benefit in diagnosis, prediction of disease risk, and in novel therapy design. The cancer-risk-associated SNPs, identified in genome wide association studies (GWAS), are invariably found in clusters that all travel in meiosis on the same haplotype block. To translate a GWAS result from a statistical observation into biological insight requires distinguishing the SNPs that are functional from those merely in linkage disequilibrium with the functional ones. We hypothesize that functional SNPs alter disrupt enhancers active in breast epithelial cells and in breast cancer cells, while the rider SNPs do not. We predict that functional SNPs will alter the binding of transcription factors essential for enhancer function. The project of this objective was to identify the top-performing machine learning-based algorithm to distinguish such enhancers from other genomic DNA lacking this functionality.

After a systematic comparison of various approaches, and various types of sequence features, we identified a classifier that performed with a remarkable prediction accuracy of 95.1%. These results from this seed project constitute potent preliminary data for a follow up study, in which we will use this classifier to identify the SNPs that demote the element harboring them from belonging in the positive set to belonging in the negative set. The envisioned follow up study will be a marriage of data analytics and wet-bench science. The IIAI (Lee/Sonka) group will identify breast-cancer associated SNPs that pass the bioinformatic filters presented above, the Cornell group will test these SNPs for functional effect in standard luciferase-based reporters assays or alternatively in sequencing-based massively parallel reporter assays.

Publications resulting from project:

No publications yet. There is potential in the near term for a purely bioinformatics paper comparing the performance of the classifier identified here to that of published classifiers, and in the longer term, for one that combines prioritization of SNPs, potentially with multiple classifiers, and tests of those SNPs in reporter assays in vitro.