Comparison of Support Vector Regression Models of Transcription Factor E2F1 and E2F4's Binding Specificity to DNA Sequences

Sunwoo Yim

September 2015

1 Abstract

Much research on transcription factor binding to DNA sequences has been done, yet there is a surprising lack of information on the differences in binding specificities between individual transcription factors (TFs) in the same family. This study focused on TF E2F1 and E2F4 by using support vector regression (SVR) to train models from genomic-context protein binding microarray (gcPBM) data and then analyzing these models such that the significant features weights can be extracted. Analysis of the most significant features from each model showed that the six-base long flanks on either side of the core significantly contributed to the binding preference of both TFs. The models were also compared by core and by TF, and the features with the greatest deviance between each pair of models tested were studied. It was found that E2F1 had a higher specificity to nucleotides A and T stretches than E2F4 in both flanks and that the preferences of E2F1 could be more easily predicted by sequence features than those of E2F4. Finally, comparison between the sequences' cores demonstrated that the SVR model had higher accuracy when predicting sequences with nucleotide cores consisting of GCGC as compared to those consisting of GCGG.

2 Introduction

Although every cell in the human body contains the exact same DNA, groups of cells express the DNA information differently and thus play different roles in the body. Cells become skin cells rather than blood cells or cancerous cells rather than healthy ones largely due to the interactions of transcription factors (TFs), or proteins that bind to specific DNA binding sites to regulate the process of transcription, thus controlling expression of genetic information (Figure 1). Because TFs play such a major role in deciding how DNA information is decoded in the body, the need to understand TF binding specificities to DNA sequences has become essential in genomics and bioinformatics.

Out of the thousands of TFs found in the human genome, TFs E2F1 and E2F4 were chosen as the focus of the project because they play a vital role in the human body by regulating cell proliferation and apoptosis [6]. As a result, they are crucial during the loss of retinoblastoma (Rb) tumor suppressor function, which can lead to uncontrolled malignant cell growth and thus human cancer [8].

Much research aiming to discover patterns in the sequences bound by specific TFs such as E2F1 and E2F4 base analysis from models such as position weight matrices (PWMs) [5]. However, PWMs fail to account for any interdependencies between nucleotides because they show the relative frequencies of only one nucleotide at each position in the sequences bound by TFs. This can lead to incorrect interpretations because the structures of the TF proteins lead to complex interactions with DNA sequences, often involving multiple nucleotides in the binding process. Other studies support the need to include nucleotide relationships as models that also include 2-mer or 3-mer features perform significantly better when predicting TF binding specificities for DNA sequences [4]. This is further supported when looking at the PWMs for E2F1 and E2F4, which show little difference in their preferred sequences (Figure 2A). However, the data from chromatin immunoprecipitation with DNA sequencing (ChIP-seq) *in vivo* shows a much different story, with the majority of the sequences being bound differently by the two TFs (Figure 2B), thus suggesting that other factors such as

nucleotide interdependency comes into play during the TF binding process.

This research project thus aimed at filling in this gap in literature by exploring the most significant discrepancies in the sequences bound by TFs E2F1 and E2F4 that made these proteins bind so differently to the same DNA sequences despite sharing very similar protein structures.

3 Materials & Methods

Support vector regression (SVR), a form of machine learning, was used to train a model from the experimental data and predict the binding specificity scores for new DNA sequences [7]. So, a library for support vector machines (LIBSVM) was adopted [3] and its Java code altered to analyze the experimental data [2].

The raw experimental data for the binding specificity for TFs E2F1 and E2F4 used in this project contained thousands of 36 base long DNA sequences and their respective log signal intensity scores. A laboratory staff experimentally produced the data for E2F1 and E2F4 using genomic-context protein binding microarrays, or gcPBMs [4], which were also developed in the lab based off of uPBMs, or universal PBMs [1]. The gcPBMs were used to determine relative TF binding affinities to the tested DNA sequences and were preferred over uPBMs because the synthesized DNA sequences contained genetic context to simulate *in vivo* results more closely.

3.1 Process and filter raw gcPBM data

Data for both TFs from the gcPBM was first processed before being filtered. First, all the sequences were made 34-mers by cutting off the "A" or "T" nucleotide that was attached to each end of the sequence to facilitate the primer double-stranding process. Then, the reverse compliment of all the sequences with cores CCGC was taken to turn them into sequences with core GCGG. Each sequence also contained two orientations as the PBM

accounted for both sides of the sequence being attached to the microarray's glass slide. The best orientation score was chosen to represent each unique sequence. Lastly, there contained duplicate sequences with different scores in the raw data, so the the duplicated sequence was assigned the median of the log signal intensity scores. The median was used as the experimental data could have produced extremes, making the mean not practical because it could skew the data.

After this processing had occured, the data was selected such that it satisfied certain conditions. First, the selected sequences had to have a GCGC or GCGG core. This was done to provide better data as E2F1 and E2F4 are known to bind very well to sequences with those cores. In addition, only the sequences in which the absolute value of the difference in orientation was lower than a cutoff score were chosen. The cutoff score was found by testing several values and picking the one that produced the highest correlation coefficient (R^2) values in the model. Then, the sequences that contained "GCGC" or "GCGG" in the farthest 11-mer flanks on either side were taken out to ensure that the TFs would bind at or near the core. This was done so that the TF would not accidentally bind in the farthest flanks and thus misrepresent the sequence with its log signal intensity score.

3.2 Partition and format the data

The processed PBM data for each TF was then randomly shuffled and partitioned into training sets (80% of data) and testing sets (20% of data). This process was repeated to produce 10 different sets of training and testing data. All of the data was then converted into the LIBSVR format by converting each sequence into both 1-mer and 3-mer features and ordering them in the way described by other papers [4]. However, because the data was subselected and tested separately based on the GCGC core and the GCGG core, the features in the core that were same for every sequence, specifically positions 16-19 for 1-mers and 16-17 for 3-mers, were ignored to prevent the weights for those features overwhelming the weights for the other features. This process resulted in features 1-120 for the 1-mer features

and features 121-2040 for 3-mer features.

Because these 10 datasets were later subselected based on cores GCGC and GCGG, this resulted in 10 datasets for TF E2F1 Core GCGC, TF E2F1 Core GCGG, TF E2F4 Core GCGC, and TF E2F4 Core GCGG, or a total of 40 datasets.

3.3 Training, grid search, cross validation, and prediction

The support vector regression training model takes two parameters: the cost variable (c) to penalize deviations from the model and the epsilon variable (p) to set the accepted and unpenalized deviation of each vector from the model. The optimal parameter settings (c, p) was unique for each dataset, and so the best pairing had to be found using a grid search. A coarse grid search was first done by testing every combination of c values consisting of 2^{-9}, 2^{-8}, 2^{-7}, ..., 2^{-2} and p values consisting of 2^{-7}, 2^{-6}, ..., 2^{-1}, 1, and then a fine grid search was done by zooming in on the area with the best R^2 values and repeating the grid search with new c and p values. For each (c, p) pairing, the R^2 value was found by doing a 5-fold cross validation using the training set. After both the coarse and fine grid searches, the parameter pairing that produced the highest R^2 value and thus the closest prediction of the TF binding specificity was used to train the entire training set, resulting in the final prediction model.

This model was then tested for accuracy by predicting the binding specificities of the testing data and comparing the results with the experimental log signal intensity scores of the testing data, thus producing a R^2 value that would represent the entire model's accuracy.

3.4 Finding feature weights

The model could then be interpreted to help find patterns as to why the TF chose to bind to the sequences it did. To do this, the feature weights, showing the relative importance of each of the features in the sequences, of the model were extracted using the matrix equation

$$\begin{bmatrix} y_1 & y_2 & y_3 & \dots & y_s \end{bmatrix} * \begin{bmatrix} v_{1,1} & v_{1,2} & v_{1,3} & \dots & v_{1,f} \\ v_{2,1} & v_{2,2} & v_{2,3} & \dots & v_{2,f} \\ \dots & \dots & \dots & \dots & \dots \\ v_{s,1} & v_{s,2} & v_{s,3} & \dots & v_{s,f} \end{bmatrix} = \begin{bmatrix} w_1 & w_2 & w_3 & \dots & w_f \end{bmatrix}$$

where y_k is the log signal intensity of the kth sequence (maximum of s sequences), $v_{k,l}$ is binary value of 0 or 1 showing whether the lth feature is present in the kth sequence, and w_l is the outputted weight of the lth feature.

Each of the four categories tested for TFs E2F1 and E2F4 and their cores GCGC and GCGG contained 10 unique datasets, each with its own feature weights. So, the feature weights for each category across all 10 iterations were averaged, and the standard deviation of each feature for each category was calculated.

4 Results

4.1 Graph of feature weights

The averaged feature weights were plotted on a clustered column graph and grouped by the 1-mer and 3-mer features for TF E2F1 Core GCGC, TF E2F1 Core GCGG, TF E2F4 Core GCGC, and TF E2F4 Core GCGG (Figures 3 & 4). Standard deviation for each feature weight was shown by the error bars. A cutoff score was then calculated by taking the highest feature weight and dividing it by 2. The features that had a weight above this cutoff were then selected for both 1 and 3-mer graphs (Figure 5).

Analysis of this data demonstrated that the most significant features for all four models seemed to lie in the immediate flanking regions of the core. The data also suggested that TF E2F1 prefers common A and T homotrimers on both sides of the core when binding to sequences. However, the results for TF E2F4 are more ambiguous. While the TF still showed slight preference to sequences with A and T homodimers with core GCGC, it failed to show much specificity to sequences with these features in the GCGG subgroup. Overall,

it seemed that TF E2F4 displayed less specificity to sequences with A and T homodimers.

4.2 Graphs comparing cores in each TF

Before the feature weights for E2F1 and E2F4 were compared, the differences between the two cores for each TF were analyzed. Each set of data was normalized because the different training sequences for each model, created by the different cutoff scores of the orientation differences in the processing of the PBM data, led to disparate feature weight meanings, preventing any accurate method of comparison. So, all the weights were normalized such that each normalized weight would be contained in the interval -1 to 1 using the equation

$$n_k = \frac{w_k}{w_{max}}$$

where n_k represents the normalized feature weight of the unnormalized weight w_k and w_{max} represents the maximum feature weight. This unorthodox way of normalizing the feature weights was used so that an unnormalized weight of 0 would still remain 0 after normalization, avoiding any nonzero weight being given to the corresponding feature and thus preventing false impressions that the feature impacted the model in any way. After the normalization, each feature was graphed using R on a scatter plot with core GCGC on the x-axis and GCGG on the y-axis, and then the line $y = x$ was drawn (Figures 6A & 6B). The line $y = x$ was used as a base for comparing the variations in the two models for both cores. If each feature in both models theoretically weighed the same, then each plotted point would be on the $y = x$ line. As the majority of the points did not lie on the line, this was not the case, and so the features with the greatest distance from the line and thus with the greatest variance in weight between the two cores were found. Depending on which side of the line the points were located, the core that the features were more specific to could be found, and these significant features could be grouped based on core (Figures 7A & 7B).

Using this process, figures could be created for E2F1 Cores GCGC vs. GCGG (Figure 6A & 7A) and for E2F4 Cores GCGC vs. GCGG (Figure 6B & 7B).

The graphs seemed to indicate that many features were not given the same weight in the model for cores GCGC and GCGG because many points did not lie near the $y = x$ line. Of the features that were farthest from this line, the majority contained nucleotides with positions located in the core. In addition, the majority of the features clustered around the point $(0, 0)$, suggesting that their weights were around 0 for both cores in both TFs.

4.3 Graphs comparing TFs in each core

After the feature weights between cores for each TF were compared, the weights between TFs for each core could be analyzed. Using the same process but with E2F1 on the x-axis and E2F4 on the y-axis, graphs were created for Core GCGC E2F1 vs. E2F4 (Figure 6C) and Core GCGG E2F1 vs. E2F4 (Figure 6D). The graphs were interpreted in the same way and the most significant features were grouped based on TF (Figures 7C and 7D).

As with the graphs comparing cores, the graphs comparing TFs showed the majority of feature weights clustered around 0. However, the feature points seemed to lie closer to the $y = x$ line, with fewer outliers. In addition, the farthest outliers proved to be mostly A and T homotrimers.

4.4 Reliability and statistical errors

To ensure the highest degree of accuracy in the feature weights, the models needed to accurately predict the log signal intensity scores for new testing sequences. When averaging the R^2 values for all 10 models, we found very high R^2 values of 0.895 for TF E2F1 Core GCGC, 0.810 for TF E2F1 Core GCGG, 0.799 for TF E2F4 Core GCGC, and 0.743 for TF E2F4 Core GCGG. So, the models were very reliable as they displayed high prediction accuracy, thus confirming the validity of the feature weights. In addition, the feature weights were very stable as shown by the very small error bars and thus low standard deviations of each weight (Figures 3 & 4), thus reaffirming the small statistical errors in the features. The patterns that were observed from the data also corroborated themselves, as the same observations

were made repeatedly, adding to the validity of the results.

So, by looking at and comparing the feature weights for both TFs by core, the most significant variations in the sequences bound by TF E2F1 and E2F4 can be found. Finding patterns in these feature differences can then lead to critical conclusions as to why TFs sharing very similar structural domains fail to show analogous binding preferences to the same DNA sequences.

5 Illustrations

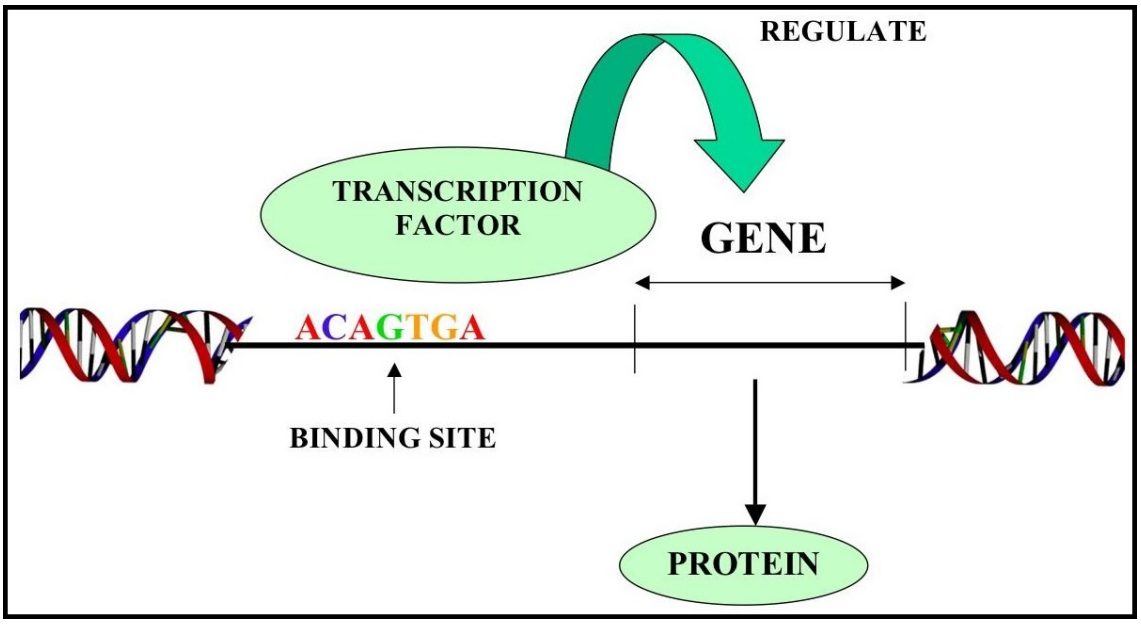


Figure 1: Transcription factors bind to binding sites to regulate adjacent genes.

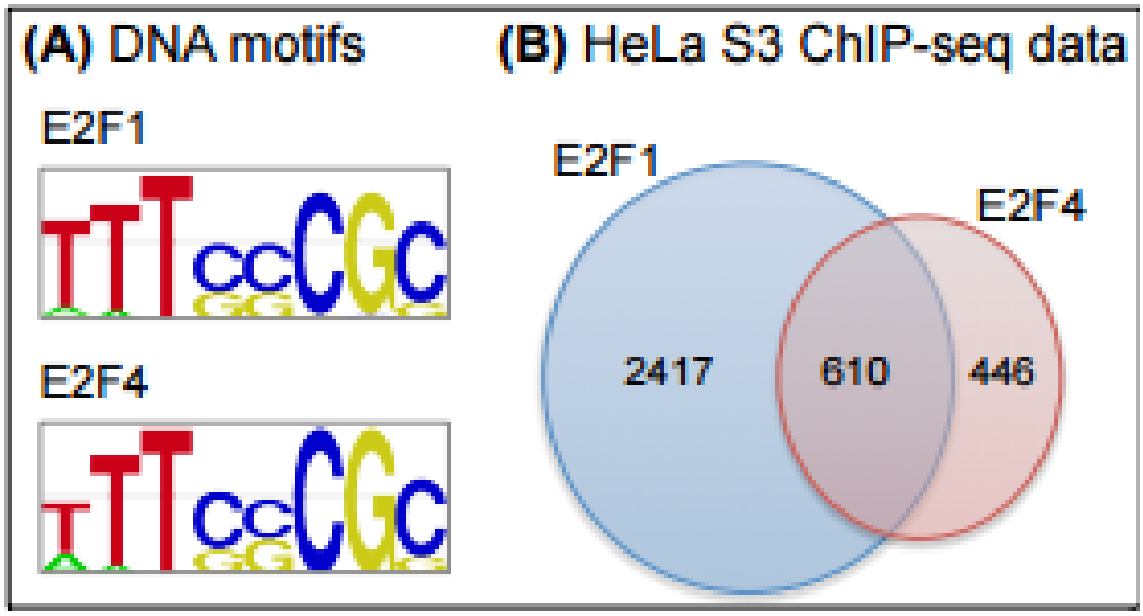
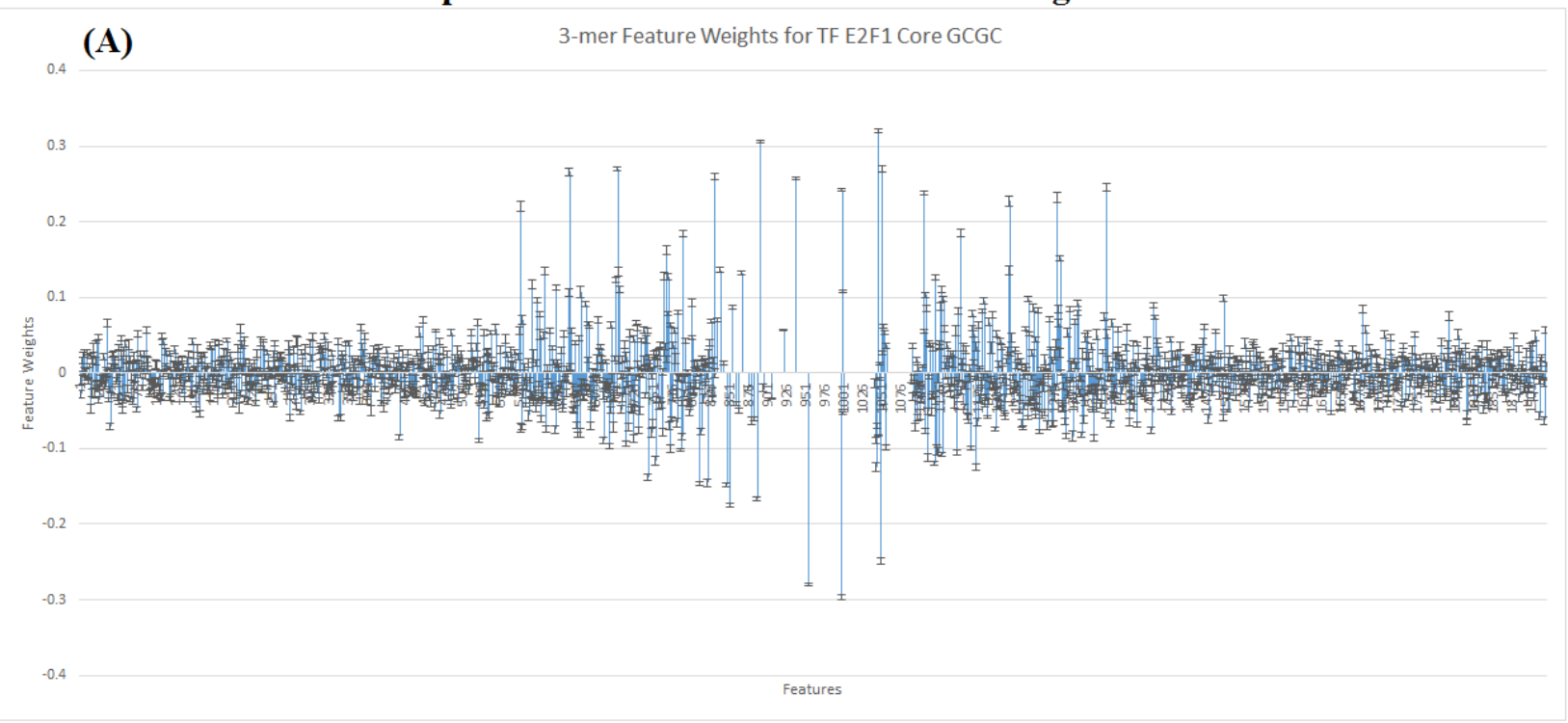


Figure 2: (A) DNA binding motifs (from Transfac) represented with PWMs
(B) *in vivo* DNA binding data (from ENCODE) for E2F1 and E2F4 showing little shared sequences

Graph of E2F1 Models' 3-mer Feature Weights

(A)



(B)

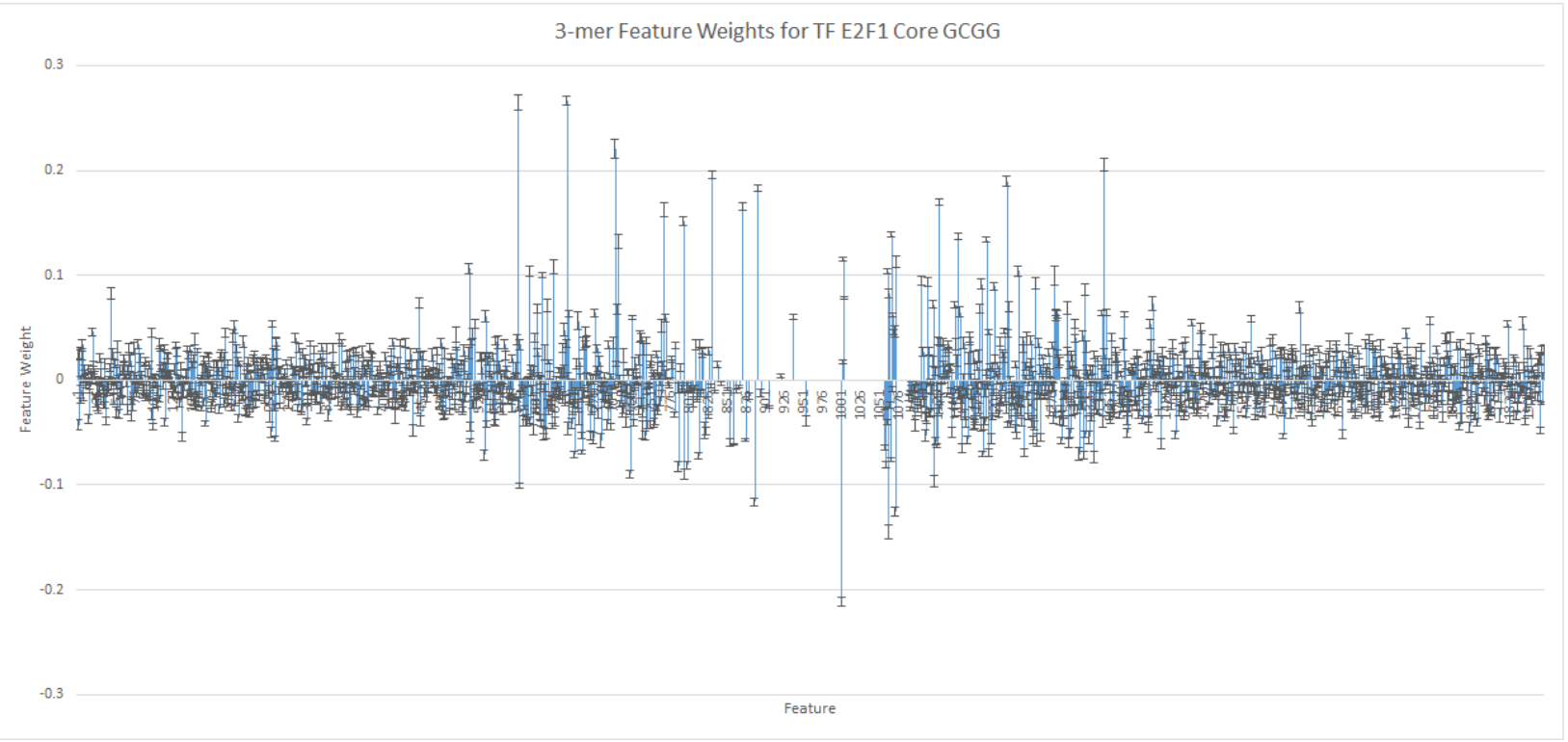


Figure 3

Graph of E2F4 Models' 3-mer Feature Weights

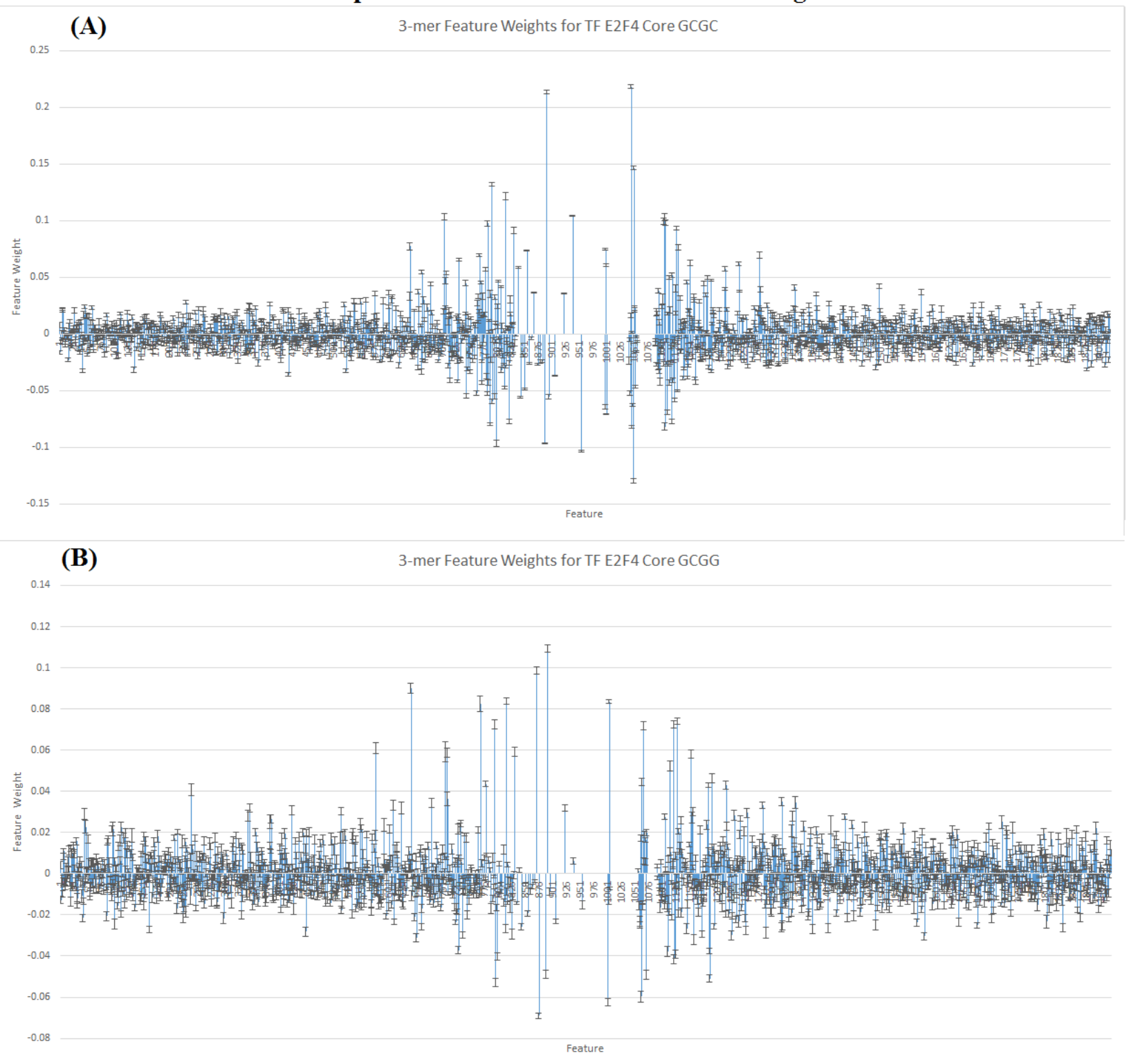
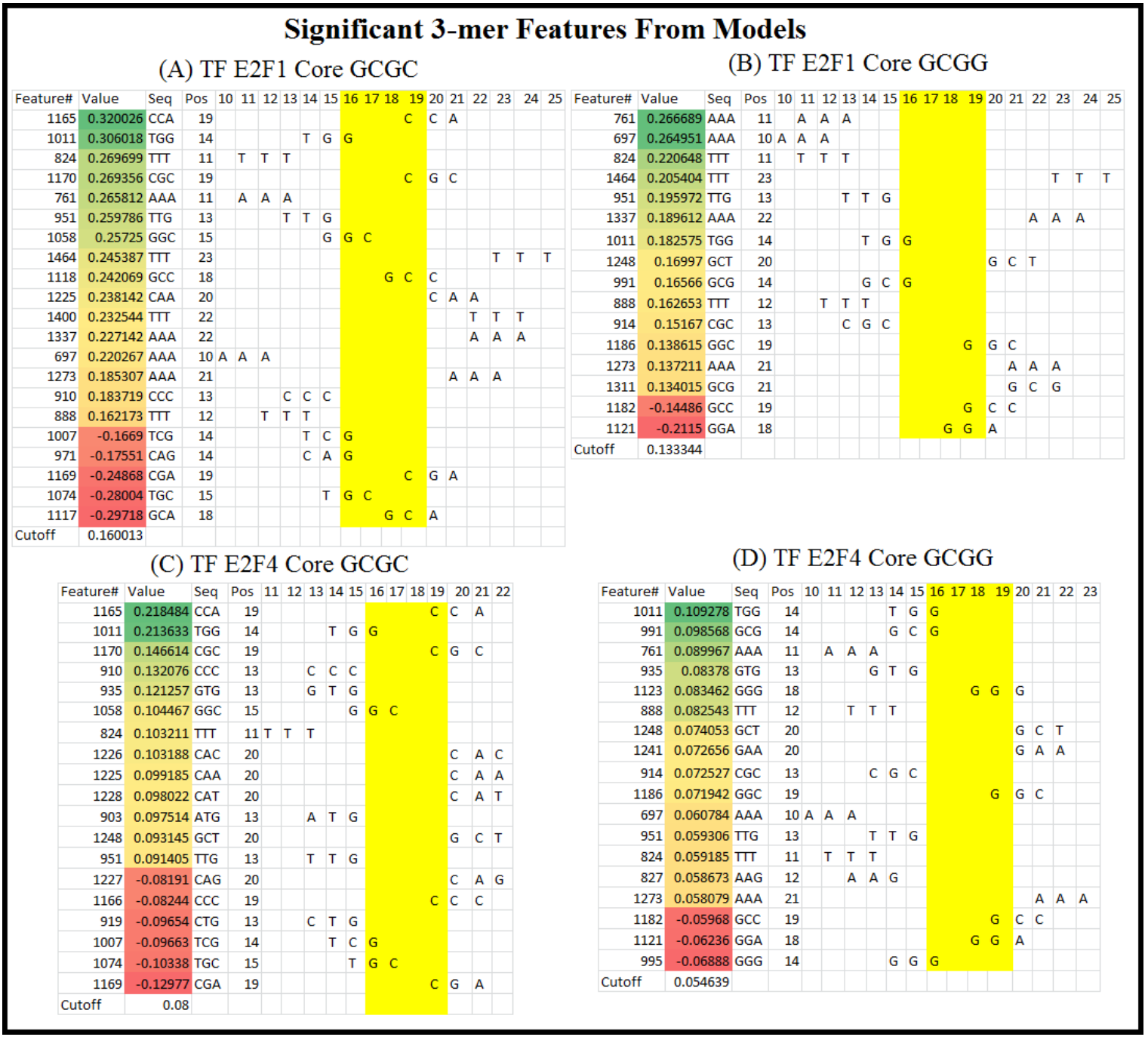


Figure 4



Feature#	Value	Seq	Pos	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1165	0.320026	CCA	19										C	C	A				
1011	0.306018	TGG	14					T	G	G									
824	0.269699	TTT	11		T	T	T												
1170	0.269356	CGC	19										C	G	C				
761	0.265812	AAA	11		A	A	A												
951	0.259786	TTG	13				T	T	G										
1058	0.25725	GGC	15						G	G	C								
1464	0.245387	TTT	23														T	T	T
1118	0.242069	GCC	18									G	C	C					
1225	0.238142	CAA	20											C	A	A			
1400	0.232544	TTT	22													T	T	T	
1337	0.227142	AAA	22													A	A	A	
697	0.220267	AAA	10	A	A	A													
1273	0.185307	AAA	21												A	A	A		
910	0.183719	CCC	13				C	C	C										
888	0.162173	TTT	12			T	T	T											
1007	-0.1669	TCG	14					T	C	G									
971	-0.17551	CAG	14					C	A	G									
1169	-0.24868	CGA	19										C	G	A				
1074	-0.28004	TGC	15						T	G	C								
1117	-0.29718	GCA	18									G	C	A					
Cutoff	0.160013																		

Feature#	Value	Seq	Pos	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
761	0.266689	AAA	11		A	A	A												
697	0.264951	AAA	10	A	A	A													
824	0.220648	TTT	11		T	T	T												
1464	0.205404	TTT	23														T	T	T
951	0.195972	TTG	13				T	T	G										
1337	0.189612	AAA	22													A	A	A	
1011	0.182575	TGG	14					T	G	G									
1248	0.16997	GCT	20											G	C	T			
991	0.16566	GCG	14					G	C	G									
888	0.162653	TTT	12			T	T	T											
914	0.15167	CGC	13				C	G	C										
1186	0.138615	GGC	19										G	G	C				
1273	0.137211	AAA	21												A	A	A		
1311	0.134015	GCG	21												G	C	G		
1182	-0.14486	GCC	19										G	C	C				
1121	-0.2115	GGA	18									G	G	A					
Cutoff	0.133344																		

Feature#	Value	Seq	Pos	11	12	13	14	15	16	17	18	19	20	21	22
1165	0.218484	CCA	19									C	C	A	
1011	0.213633	TGG	14				T	G	G						
1170	0.146614	CGC	19									C	G	C	
910	0.132076	CCC	13			C	C	C							
935	0.121257	GTG	13			G	T	G							
1058	0.104467	GGC	15					G	G	C					
824	0.103211	TTT	11	T	T	T									
1226	0.103188	CAC	20										C	A	C
1225	0.099185	CAA	20										C	A	A
1228	0.098022	CAT	20										C	A	T
903	0.097514	ATG	13			A	T	G							
1248	0.093145	GCT	20										G	C	T
951	0.091405	TTG	13			T	T	G							
1227	-0.08191	CAG	20										C	A	G
1166	-0.08244	CCC	19									C	C	C	
919	-0.09654	CTG	13			C	T	G							
1007	-0.09663	TCG	14				T	C	G						
1074	-0.10338	TGC	15					T	G	C					
1169	-0.12977	CGA	19									C	G	A	
Cutoff	0.08														

Feature#	Value	Seq	Pos	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1011	0.109278	TGG	14					T	G	G							
991	0.098568	GCG	14					G	C	G							
761	0.089967	AAA	11		A	A	A										
935	0.08378	GTG	13				G	T	G								
1123	0.083462	GGG	18									G	G	G			
888	0.082543	TTT	12			T	T	T									
1248	0.074053	GCT	20											G	C	T	
1241	0.072656	GAA	20											G	A	A	
914	0.072527	CGC	13				C	G	C								
1186	0.071942	GGC	19										G	G	C		
697	0.060784	AAA	10	A	A	A											
951	0.059306	TTG	13				T	T	G								
824	0.059185	TTT	11		T	T	T										
827	0.058673	AAG	12			A	A	G									
1273	0.058079	AAA	21												A	A	A
1182	-0.05968	GCC	19										G	C	C		
1121	-0.06236	GGA	18									G	G	A			
995	-0.06888	GGG	14					G	G	G							
Cutoff	0.054639																

Figure 5

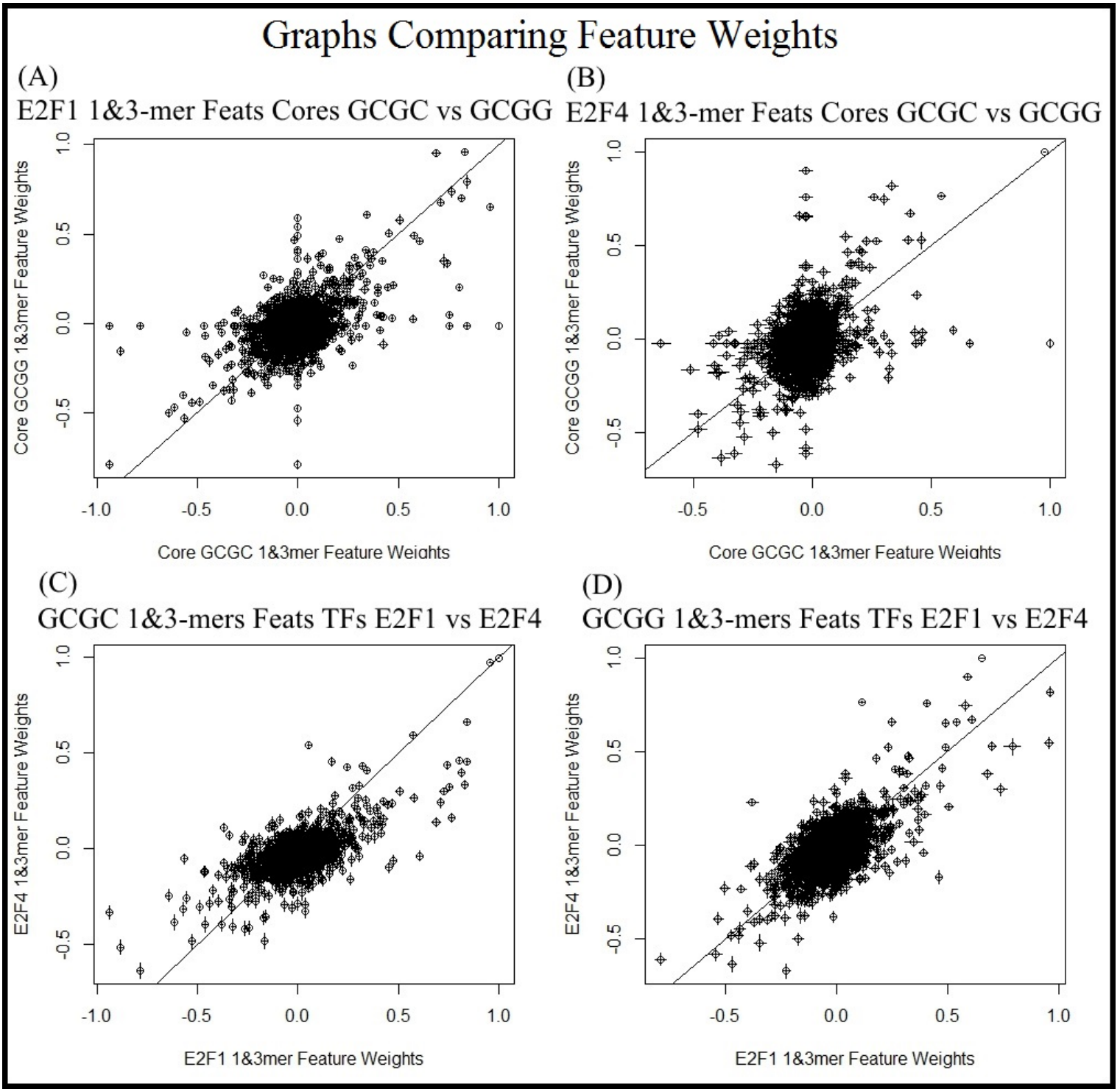

Graphs Comparing Feature Weights
(A)
E2F1 1&3-mer Feats Cores GCGC vs GCGG
Core GCGG 1&3mer Feature Weights
Core GCGC 1&3mer Feature Weights
(B)
E2F4 1&3-mer Feats Cores GCGC vs GCGG
Core GCGG 1&3mer Feature Weights
Core GCGC 1&3mer Feature Weights
(C)
GCGC 1&3-mers Feats TFs E2F1 vs E2F4
E2F4 1&3mer Feature Weights
E2F1 1&3mer Feature Weights
(D)
GCGG 1&3-mers Feats TFs E2F1 vs E2F4
E2F4 1&3mer Feature Weights
E2F1 1&3mer Feature Weights

Figure 6

Features with Greatest Deviance from y=x in Comparison Graphs

Significant Features E2F1 GCGC vs. GCGG

Feat#	Dist(y=x)	Name	Seq	Pos	12	13	14	15	16	17	18	19	20	21
1165	0.71786	GCGC	CCA	19								C	C	A
1170	0.605593	GCGC	CGC	19								C	G	C
1121	0.55651	GCGC	GGA	18							G	G	A	
1118	0.545259	GCGC	GCC	18							G	C	C	
62	0.501359	GCGC	C	20									C	
59	0.423562	GCGC	G	15				G						
1058	0.423562	GCGC	GGC	15				G	G	C				
910	0.386398	GCGC	CCC	13		C	C	C						
1117	0.650226	GCGG	GCA	18							G	C	A	
1169	0.542613	GCGG	CGA	19								C	G	A
60	0.51293	GCGG	T	15				T						
1074	0.51293	GCGG	TGC	15				T	G	C				
991	0.420161	GCGG	GCG	14			G	C	G					
914	0.384076	GCGG	CGC	13		C	G	C						
971	0.357761	GCGG	CAG	14			C	A	G					

Significant Features E2F4 GCGC vs. GCGG

Feat#	Dist(y=x)	Name	Seq	Pos	11	12	13	14	15	16	17	18	19	20	21	22
1165	0.723345	GCGC	CCA	19									C	C	A	
1170	0.484272	GCGC	CGC	19									C	G	C	
1121	0.410698	GCGC	GGA	18								G	G	A		
1182	0.39263	GCGC	GCC	19									G	C	C	
910	0.387087	GCGC	CCC	13			C	C	C							
975	0.371886	GCGC	CCG	14				C	C	G						
995	0.364946	GCGC	GGG	14				G	G	G						
62	0.339523	GCGC	C	20										C		
991	0.655812	GCGG	GCG	14				G	C	G						
1123	0.55578	GCGG	GGG	18								G	G	G		
1241	0.504209	GCGG	GAA	20										G	A	A
914	0.482719	GCGG	CGC	13			C	G	C							
1186	0.480435	GCGG	GGC	19									G	G	C	
1169	0.432901	GCGG	CGA	19									C	G	A	
63	0.354267	GCGG	G	20										G		

Significant Features GCGC E2F1 vs. E2F4

Feat#	Dist(y=x)	Name	Seq	Pos	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
52	0.455889	E2F1	T	13				T												
1464	0.424803	E2F1	TTT	23														T	T	T
69	0.388282	E2F1	A	22												A				
697	0.386257	E2F1	AAA	10	A	A	A													
1404	0.374875	E2F1	AAT	23														A	A	T
761	0.349865	E2F1	AAA	11		A	A	A												
1337	0.32985	E2F1	AAA	22												A	A	A		
61	0.428402	E2F4	A	20											A					
1117	0.428402	E2F4	GCA	18								G	C		A					
54	0.36279	E2F4	C	14					C											
935	0.345728	E2F4	GTG	13				G	T	G										
51	0.338634	E2F4	G	13				G												
70	0.292656	E2F4	C	22												C				
67	0.282909	E2F4	G	21											G					

Significant Features GCGG E2F1 vs. E2F4

Feat#	Dist(y=x)	Name	Seq	Pos	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
52	0.447626	E2F1	T	13										T												
995	0.314148	E2F1	GGG	14											G	G	G									
1464	0.310794	E2F1	TTT	23																				T	T	T
68	0.303846	E2F1	T	21																		T				
697	0.288522	E2F1	AAA	10							A	A	A													
1321	0.278374	E2F1	TAA	21																		T	A	A		
1416	0.274689	E2F1	ATT	23																				A	T	T
51	0.430166	E2F4	G	13										G												
1241	0.287615	E2F4	GAA	20																	G	A	A			
63	0.247528	E2F4	G	20																	G					
1123	0.247528	E2F4	GGG	18															G	G	G					
1011	0.245636	E2F4	TGG	14											T	G	G									
898	0.242753	E2F4	AGC	13										A	G	C										
1458	0.239196	E2F4	TGC	23																				T	G	C

Figure 7

6 Discussion

During the protein-DNA interaction during the binding process, the area of focus is primarily the core of the sequence and its closest surrounding flanks, and so the distant flanks have less of an influence on the TF binding specificity. This is indicated as the most significant features for each model represent the six-base-long flanks on either side of the core, suggesting that this area has the greatest influence on the binding preferences of TFs. Because the sequences bound by TFs E2F1 and E2F4 intersect to some extent, the important features for these TFs are expected to overlap. The preference for sequences with homotrimers of As and Ts thus is shared by E2F1 and E2F4, but the larger specificity for these 3-mers shown in the E2F1 features suggests that sequences with these features are expected to show more binding potential to E2F1. As only 1-mer and 3-mer features were used in the model, the information gained from the sequences was slightly limited. Although homotrimers were observed, it is hard to say whether they may be part of larger homogenous motifs. Indeed, the significant features from the models has shown that identical homotrimers have been found slightly overlapping in the flanks, and this could point to larger sequences of identical nucleotides. However, including k-mer features larger than the 3-mers would require a much larger training set to accomodate the exponentially increased range of features, and as the processing and filtering of the raw PBM data limited the accepted data size, this was not tested.

The deviance of the feature points from the $y = x$ line in the graphs comparing cores can be explained by the nature of the nucleotide interactions. The separation of the sequences by core would impact the bonding structures of the flanks, and so the important features would be different for each core. If models had been constructed without this distinction of cores, the features containing only the core could not have been ignored, resulting in overrepresented and disproportionately high weights for those features. Thus, the feature points for the graphs comparing cores was expected to correlate less with the $y = x$ line than the graphs comparing TFs. In addition, the majority of the features used in the model

had a weight of 0. Because 3-mer features accounted for $4^3 = 64$ different combinations of nucleotides for every position, many features were expected to be represented only a few times in the training sequences. So, the SVR model would not have had enough information to accurately place a weight for these features, resulting in a very low weight close to 0. The lower R^2 scores found for E2F4 models compared to E2F1 models indicates that TF E2F1's binding specificities can be better predicted by sequence information alone than those of E2F4. The reason for this is still unclear but may be related to the presence of cofactors in the E2F4 binding process. Thus, without these additional factors included in the models, the ability to predict E2F4's bound sequences is diminished. The project's results also enhance the conclusions of studies on human TFs that demonstrate that TFs show greater preference to sequences with A- and T- stretches [5].

7 Conclusion and Future Work

Although many studies aim to find binding preferences of whole families of TFs, there is a lack of data on individual TFs within a particular family. This project found that TF E2F1 preferred sequences with A and T stretches in the flanking regions more than TF E2F4 did. The flanks that were six bases on either side of the core especially influenced the outcome of binding for these two TFs. In addition, TF E2F4 was shown to be less predictable than E2F1 due to possible factors such as cofactors. As the significant features for each model as well as that for the comparison graphs were analyzed in conjunction, the conclusions made are sound and supported by multiple sources. The standard deviations were low and the reliability high, thus supporting the validity of the data. Many of the conclusions made for this project applied to TFs in different families from different studies as well.

Further studies should look at additional methods of comparison for the data to either find new observations or corroborate established ones. In addition, other TFs in the E2F family would be added to the study such as TF E2F2 or E2F3 in order to answer the

question: Could the observations made could be extended to other proteins in the same family? Additional areas for research include why these TFs prefer A and T stretches in the flanks and how to better distinguish which sequences would be bound to which TFs in the E2F family without the use of the models. One change in the methods would have been to normalize the log signal intensity scores to allow a better comparison of the data.

References

[1] Michael F Berger and Martha L Bulyk. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nature protocols*, 4(3):393–411, January 2009.

[2] Chih-Wei Hsu, Chih-Chung Chang and Chih-Jen Lin. A Practical Guide to Support Vector Classification. *BJU international*, 101(1):1396–400, 2008.

[3] Chih-Chung Chang et al. Libsvm – a library for support vector machines.

[4] Gordân Raluca et al. Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell reports*, 3(4):1093–104, April 2013.

[5] Jolma et al. DNA-binding specificities of human transcription factors. *Cell*, 152(1-2):327–339, 2013.

[6] GeneCards. E2f transcription factor 1, 2008.

[7] Alex J Smola, Bernhard Sch, and B Scholkopf. A Tutorial on Support Vector Regression. *Statistics and Computing*, 14(3):199–222, 2004.

[8] D Wang, J L Russell, and D G Johnson. E2F4 and E2F1 have similar proliferative properties but different apoptotic and oncogenic properties in vivo. *Molecular and cellular biology*, 20(10):3417–3424, 2000.