SUPPLEMENTARY MATERIAL

DeepGSR: An optimized deep-learning structure for the recognition of genomic signals and regions

Manal Kalkatawi[1], Arturo Magana-Mora[2,1], Boris Jankovic[1] and Vladimir B. Bajic[1,*]

[1]King Abdullah University of Science and Technology, Computational Bioscience Research Center, Thuwal 23955-6900, Kingdom of Saudi Arabia.

[2]National Institute of Advanced Industrial Science and Technology, Computational Bio Big-Data Open Innovation Laboratory, Tokyo, 135-0064, Japan.

1 Datasets

In DeepGSR, we used the cDNA data of four genomes to extract both the polyadenylation signals (PAS) and translation initiation sites (TIS). For *homo sapiens* (human) genome, we used the human assembly GRCh37 (also known as hg19); and for *Mus musculus* (mouse) genome, we used the primary assembly GRCm38; the cDNA data for these genomes were downloaded from the Mammalian Gene Collection (MGC) (Strausberg, et al., 1999; Team, et al., 2009). For *Bos taurus* (bovine) genome, we used the assembly Bos_taurus_UMD_3.1.1 and the cDNA data was downloaded from Ensembl organization (Aken, et al., 2016). For *Drosophila melanogaster* (fruit fly) genome, we used Release_6 – annotation release Dmel_Release_6.01 and the cDNA data was downloaded from FlyBase (Gramates, et al., 2017).

The data extraction workflow is depicted in Figure S1. Its automation and simplicity are the main advantage and strength for such a pipeline. In the following subsections, we describe the data extraction procedure in detail.

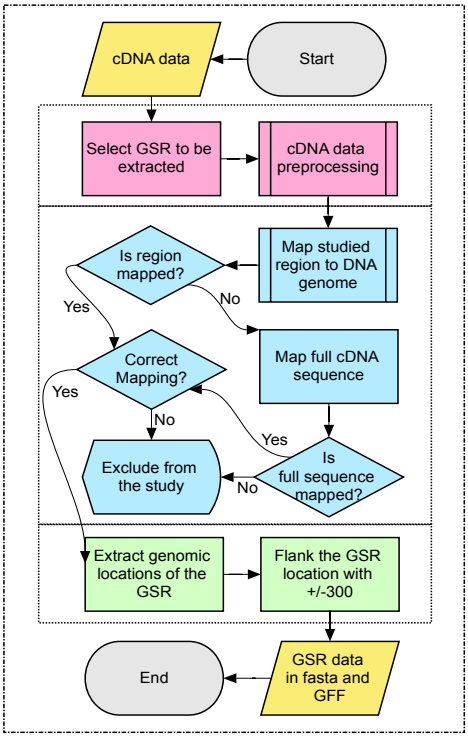

Fig. S1: Data extraction workflow

1.1 cDNA Data Preprocessing

cDNA data should be preprocessed to determine the region of study where signals are located. For example, in the case of PAS motifs, they are located in the 3'UTR, which is the region between the stop codon and the start of polyadenylation tail or (polyA tail). Sequences with 3'UTR of length less than nine nucleotides are excluded, since the PAS in such cases could be located in the internal exons or within introns (Tian, et al., 2007). Regarding the dangling polyA tail, it should be removed up to the first non-A nucleotide that represents the end of the 3'UTR. The tail should contain at least eight consecutive A's, and if there is a non-A nucleotide, then it should be followed by eight consecutive A's to be considered part of the tail, in which a non-A nucleotide could be an error during sequencing or due to internal priming (Proudfoot, 2011). For those sequences that do not have a polyA tail, the latter might be removed before submitting the sequences to public cDNA/EST databases because of their low sequencing quality or low complexity; being consecutive A's (Lee, et al., 2008). Then, we locate PAS, the most frequent hexamer AATAAA and the next top 11 canonical hexamers which differ with one base only than AATAAA, which are: ATTAAA, TATAAA, AGTAAA, AAGAAA, AATATA, AATACA, CATAAA, GATAAA, AATGAA, ACTAAA and AATAGA. If none of these 12 motifs are found in the 3'UTR, we check for the next four most frequent non-canonical PAS, which are: TTTAAA, AAAAAG, AAAACA and GGGGCT. For the cases where more than one motif variants were found within the same 3'UTR we considered the following two solutions:

1) If only the same motif is found multiple times, we select the most 3' motif.
2) If different motifs are found in the same 3'UTR, we calculate a score for each of the candidate motifs using Equation (1) and select the one with the highest score. Assuming M is a motif found at location L, the resulting score is the probability of M at L multiplied by the frequency of M in all cDNA sequences. The score is calculated as:

$$Score = \frac{count\ of\ M\ at\ L}{count\ of\ M} \times \frac{count\ of\ M}{count\ of\ all\ motifs} = \frac{count\ of\ M\ at\ L}{count\ of\ all\ motifs} \qquad (1)$$

Our pipeline extracted in this way 20933, 18693, 12082, and 27203 PAS data in total for all 16 motifs; for human, mouse, bovine and fruit fly, respectively. Table S1 illustrates the detailed number of all variants of true PAS signals we extract from different genomes.

Table S1. PAS data extracted for different genomes

PAS variant	Human	Mouse	Bovine	Fruit fly
AATAAA	11,302	11,393	7,862	18,641
ATTAAA	3,016	2,447	1,604	3,510
TATAAA	573	402	283	540
AGTAAA	730	424	334	263
AAGAAA	1,379	1,430	662	264
AATATA	378	192	87	2,087
AATACA	634	262	160	827
CATAAA	451	227	183	303
GATAAA	291	172	115	118
AATGAA	641	523	136	144
ACTAAA	288	117	61	115
AATAGA	179	127	40	66
TTTAAA	278	292	165	178
AAAAAG	179	129	63	42
AAAACA	206	212	95	100
GGGGCT	408	344	232	5
Total	20,933	18,693	12,082	27,203

The PAS data that we extracted from four eukaryotic cDNA are located/distributed differently in the 3'UTR based on the motif type. As already known, PAS is located upstream of the polyadenylation cleavage sites (polyA CS), that is represented by position 0 in Figure S2, and it is located close to the end of the 3'UTR. However, there are some of the less common motifs located far from the polyA tail. The distribution of PAS in the 3'UTR within different genomes is illustrated in Figure S2. From this figure, we can observe that the same PAS variant across genomes of different organisms is located in the same region, with a big overlap between them.

In case of TIS data extraction, we followed the same procedure, but we determined the 5'UTR, which is the region directly upstream the start codon. Since the non-canonical TIS are rarely found in eukaryotic genomes (Lee, et al., 2008), The canonical TIS which is ATG variant was the focus of our study. Our pipeline in this way extracted 28244, 25205, 17558, and 30283 TIS data with the ATG signal for human, mouse, bovine and fruit fly, respectively. In Figure S3, the distribution of TIS for human and mouse is illustrated. The TIS data distribution of bovine and fruit fly was not shown, since these data were extracted from annotation files not cDNA.

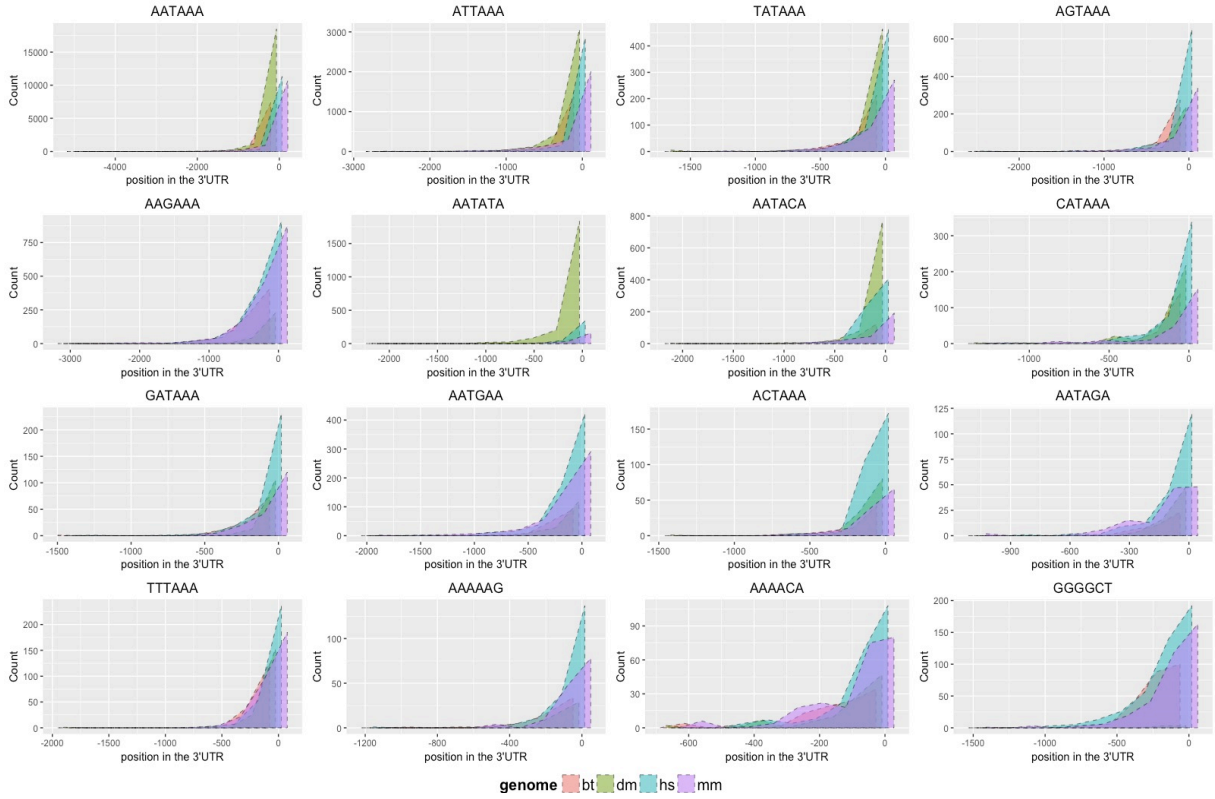

Fig. S2: PAS distribution in the 3'UTR across multiple genomes; *homo sapiens* (hs), *mus musculus* (mm), *bos taurus* (bt) and *drosophila melanogaster* (dm).

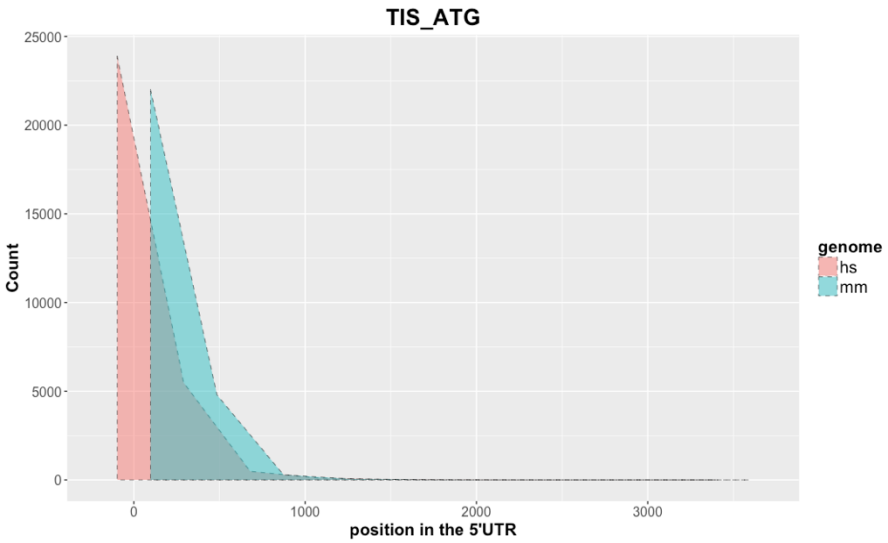

Fig. S3: TIS signals distribution in the 5'UTR across *homo sapiens* (hs) and *mus musculus* (mm) genomes

1.2 Mapping to Genome

After mapping the 3'UTR and the 5'UTR to the DNA genome, in the case of PAS and TIS data, respectively, some post-processing of the mapped data is necessary. First, we identify the sequences (the 3'UTR and the 5'UTR regions) that could not be mapped because of their short sequence length, then map the full cDNA sequence including the coding region. Then, we exclude all sequences that are mapped to different chromosomes or were incorrectly mapped. After excluding these sequences, we further process these samples that have some insertion, deletion or any modifications during the mapping to find the correct genomic locations. However, we exclude some of those samples that have big differences in mapping and contribute to the miscalculation of the genomic locations of signals under study. Such differences between the untranslated regions in the cDNA and the genomic data is caused by internal introns within these regions (Cenik, et al., 2010).

Finally, we use bedtools (Quinlan and Hall, 2010) to flank the locations of the signals with 300 bases in both upstream and downstream regions, then extracts the genomic sequences of each motif in fasta file format. It is worth mentioning that it is important to consider the strand where the signals are found during sequence extraction.

1.3 Data Properties

The frequency of the extracted PAS variants in the 3'UTR for human indicate different frequency-based ranking for the motifs than the published literature (Beaudoing, et al., 2000) (Tian, et al., 2005). This difference can be attributed to the more recent and more complete data. Despite the differences in ranking, all three studies recognize that AATAAA and ATTAAA are the two most frequent PAS. Table S2 and Figure S4 demonstrate the differences in ranking of PAS variants.

Table S2. Rank of the most common PAS data in human

PAS variant	DeepGSR	(Beaudoing, et al., 2000)	(Tian, et al., 2005)
AATAAA	1	1	1
ATTAAA	2	2	2
AATGAA	3	11	10
AGTAAA	4	4	4
AATATA	5	5	6
TATAAA	6	3	3
AAGAAA	7	10	5
CATAAA	8	6	8
AAAAAG	9	12	-
AATACA	10	8	7
GGGGCT	11	16	-
GATAAA	12	7	9
ACTAAA	13	14	12
AATAGA	14	13	13
TTTAAA	15	9	11
AAAACA	16	15	-

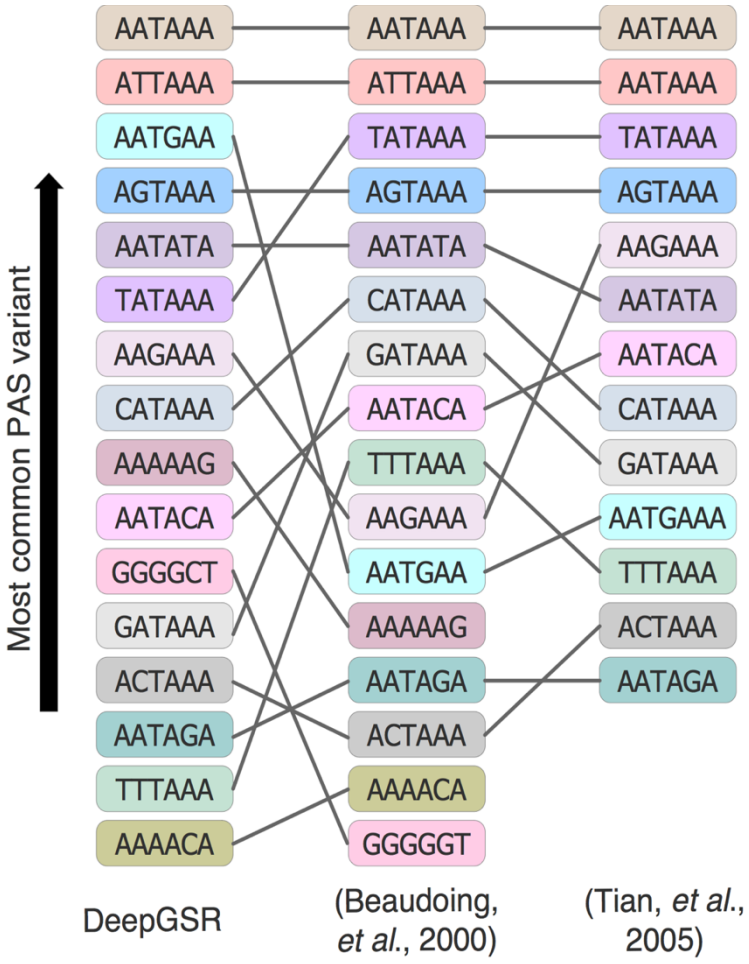

Fig. S4: Rank of the most common PAS data in human

2 Model structure and results

2.1 Model structure

Table S3. DeepGSR structure

Layer	Type of layer	Output dimensions	Connected to
Convolutional layer 1	Convolution	50 x 598 x 64	DNA input data
	Activation ReLU	50 x 598 x 64	Convolution
	Maxpooling	50 x 598 x 32	Activation ReLU
Convolutional layer 2	Convolution	100 x 589 x 25	Maxpooling
	Activation ReLU	100 x 589 x 25	Convolution
	Maxpooling	100 x 589 x 12	Activation ReLU
Fully connected layer	Dropout	100 x 589 x 12	Maxpooling
	Flattening	706,800 x 1	Dropout
	Activation tanh	256 x 1	Flattening
	Dropout	256 x 1	Activation tanh
Output layer	Activation softmax	2 x 1	Dropout

2.2 Effect of data representation on model performance

Deep learning considerably relies on the proper representation of the raw data, for this, we considered different approaches to represent the data and assessed them using a simplified CNN structure with fixed parameters. The DeepGSR represents each sequence in a two-dimensional (2D) space to mimic images as the most common input of CNN. We used one-hot vector representation that is corresponding to k-mer where k=1, 2 or 3; individually. Moreover, we could exploit the biological information represented in the mono/di/tri nucleotides. As such, instead of using the one-hot vector, the electron ion interaction pseudo potentials (EIIP) (Nair and Sreenadhan, 2006) of nucleotides can be used to substitute the numeric value of the corresponding nucleotide as such, nucleotides A, G, C, and T are replaced by 0.1260, 0.0806, 0.1340, and 0.1335, respectively. Similarly, we could use thermodynamic feature (Friedel, et al., 2009) or the base stacking (BS) energy values (Abeel, et al., 2008) or both for dinucleotides numerical representation. The results of the different data representation for 2D-CNN is depicted in Figure S5.

Moreover, we also considered a 1D-CNN model with word embedding. For this, we divided each sequence into 'words' of different number of nucleotides. The size of the words may vary, but they are overlapped to avoid missing some internal information in the sequence. Figure S6 shows the results obtained with the different word sizes for 1D-CNN.

Based on our experiments, we confirmed that there is a significant positive correlation between the proper data representation and the model capability to learn directly from data. Some of these data representations prevented network to learn at all in which cases the performance accuracy was similar to random prediction, e.g., thermodynamic feature and base stacking.

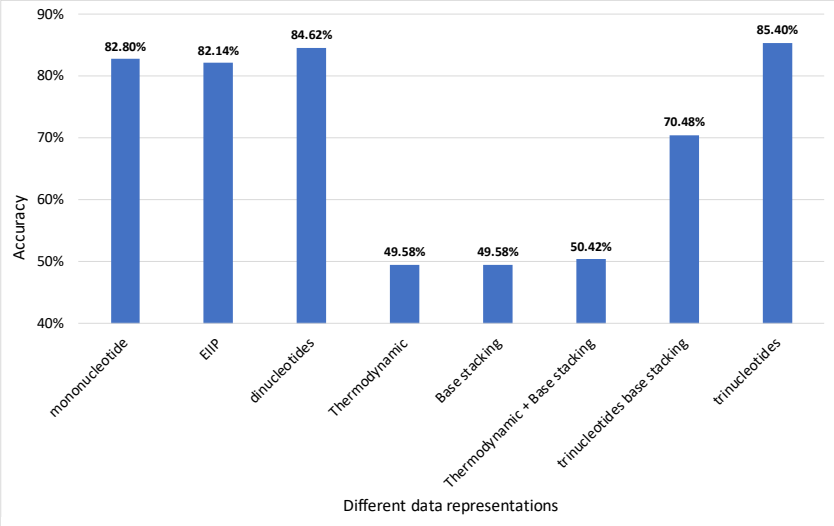

Fig. S5: 2D-CNN performance on PAS data (AATAAA) variant using various data representation

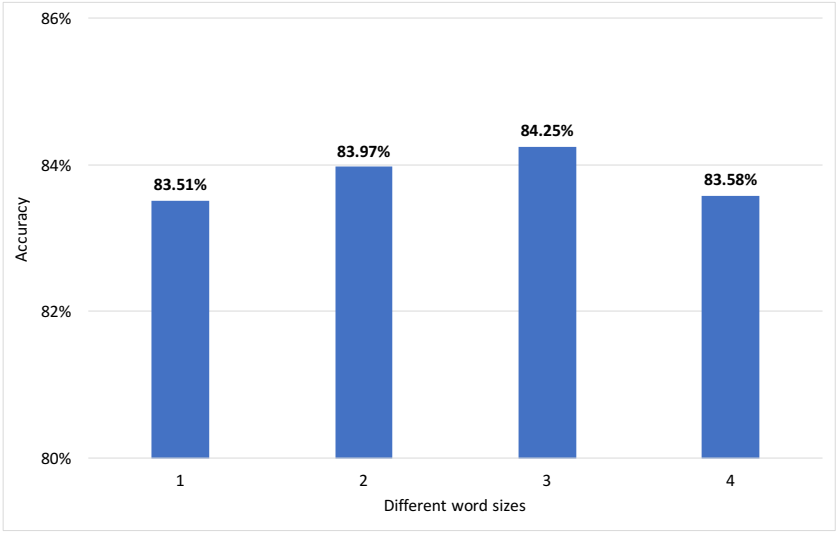

Fig. S6: 1D-CNN with word embedding performance on PAS data (AATAAA) variant using various word sizes

2.3 GSR recognition and cross-organism conservation detailed results

Table S4. Human_AATAAA_DeepGSR and Human_ATG_DeepGSR were used to test genomes of other organisms (cross-organism tests)

Testing data		Se(%)	Sp(%)	Acc(%)	AUPR(%)
Human	PAS_AATAAA	87.08	86.80	86.94	90.10
	PAS_all*	65.87	74.49	70.18	77.91
	TIS_ATG	94.76	93.89	94.32	95.89
Mouse	PAS_AATAAA	80.25	78.63	79.44	84.62
	PAS_all	76.36	78.03	77.19	82.88
	TIS_ATG	90.04	92.53	91.28	93.27
Bovine	PAS_AATAAA	81.63	77.87	79.75	84.92
	PAS_all	80.69	76.91	78.80	84.19
	TIS_ATG	91.71	88.63	90.17	92.88
Fruit fly	PAS_AATAAA	37.63	80.47	59.05	73.28
	PAS_all	40.07	79.25	59.66	73.28
	TIS_ATG	81.12	76.96	79.04	84.39

*For human data only, PAS_all represents all variants except AATAAA + only the testing portion of AATAAA (25%) that was not included in the training.

Table S5. The results on PAS data using Human_pooled-PAS_DeepGSR for predicting PAS in other organisms

Testing data		Se(%)	Sp(%)	Acc(%)	AUPR(%)
Human	PAS_all*	82.75	84.67	83.71	87.72
Mouse	PAS_AATAAA	78.23	77.07	77.65	83.25
	PAS_all	76.65	77.15	76.90	82.67
Bovine	PAS_AATAAAg	79.30	77.25	78.27	83.74
	PAS_all	79.65	76.40	78.02	83.58
Fruit fly	PAS_AATAAA	42.72	78.89	60.81	73.69
	PAS_all	46.43	77.80	62.11	74.06

*For human data only, PAS_all represents the testing portion (25%) of all PAS variants

Table S6. The results for PAS and TIS data using DeepGSR organism specific models

Testing data		Se(%)	Sp(%)	Acc(%)	AUPR(%)
Human	PAS_AATAAA	87.08	86.80	86.94	90.10
	PAS_all	82.75	84.67	83.71	87.72
	TIS_ATG	94.76	93.89	94.32	95.89
Mouse	PAS_AATAAA	84.18	84.40	84.29	87.88
	PAS_all	88.77	76.75	82.79	87.70
	TIS_ATG	94.78	93.90	94.34	95.82
Bovine	PAS_AATAAA	83.75	82.50	83.12	87.41
	PAS_all	84.81	82.08	83.44	87.79
	TIS_ATG	92.01	93.33	92.67	94.39
Fruit fly	PAS_AATAAA	87.05	87.43	87.23	89.93
	PAS_all	88.50	87.38	87.94	91.03
	TIS_ATG	94.07	92.25	93.16	95.04

*PAS_AATAAA means the model was trained on the AATAAA variant only. PAS_All means the model was trained on all PAS variant data pooled together. TIS_ATG means the model was trained on TIS data of the canonical signal ATG.

2.4 An analysis of robustness of the results to the order of the trinucleotides in the data representation

In order to assess the robustness of the DeepGSR model when considering different ordering for the input data, we shuffled the trinucleotide ordering for both PAS and TIS in the human genome. Results shown in Table S7 demonstrate that DeepGSR was able produce comparable results for the different ordering of the input data. We attribute these results to the similarity of the abstract features extracted from the input data.

Table S7. Results for human PAS and TIS data using DeepGSR with random trinucleotides ordering.

	Model	Alphabetical order				Random shuffling 1				Random shuffling 2				Random shuffling 3			
		Se (%)	Sp (%)	Acc (%)	AUPR (%)	Se (%)	Sp (%)	Acc (%)	AUPR (%)	Se (%)	Sp (%)	Acc (%)	AUPR (%)	Se (%)	Sp (%)	Acc (%)	AUPR (%)
Human	PAS_AATAAA	87.08	86.80	86.94	90.10	86.09	85.92	86.01	89.19	85.88	84.75	85.33	88.71	86.33	85.18	85.76	89.05
	PAS_all	82.75	84.67	83.71	87.72	83.27	85.43	84.33	87.83	88.17	79.32	83.83	88.02	86.18	81.39	83.83	87.75
	TIS_ATG	94.76	93.89	94.32	95.89	94.77	93.37	94.07	95.75	94.58	93.35	93.96	95.65	94.14	94.05	94.09	95.64

3 Computational Validation

To prove the efficacy of our deep learning approach presented in DeepGSR, we derived some DNA hand-crafted features in an attempt to have a unified framework using traditional machine learning techniques. The workflow of this approach is shown in Figure S7.

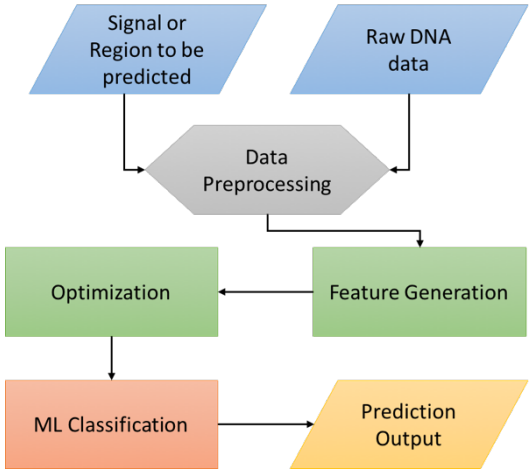

Fig. S7: Workflow of hand-crafted features approach

After engineering a large number of features, we used an artificial neural network (ANN) for assessing the performance of each of these hand-crafted features individually. Then, we conducted several experiments with different feature combinations and selected the top three best performing features and feature combinations. We then used stacked auto-encoders (AE) as an aid to choose the suitable number of hidden nodes in each layer of the deep artificial neural network (DANN). The number of neurons in each layer is a proportional function of the number of neurons in the previous layer, for example, the number of neurons in the first hidden layer of the AE was a percentage of the number of neurons in the input layer (number of features); this percent ranges from 20 to 90. The same considerations were considered to determine the number of neurons of the second hidden layer of the AE but relative to the number of neurons of the first hidden layer. Thus, we had 64 combinations of the number of neurons in the two-layers AE. For the output layer that is used for classification, we applied both ANN and softmax layer. We used MATLAB for this implementation and reported the results using a 3-folds cross-validation.

The best results of all the features tested individually and in combination are illustrated in Figure S8 and Figure S9, respectively. Finally, Figure S10 shows the comparison of the results obtained by the best performing ANN and AE.

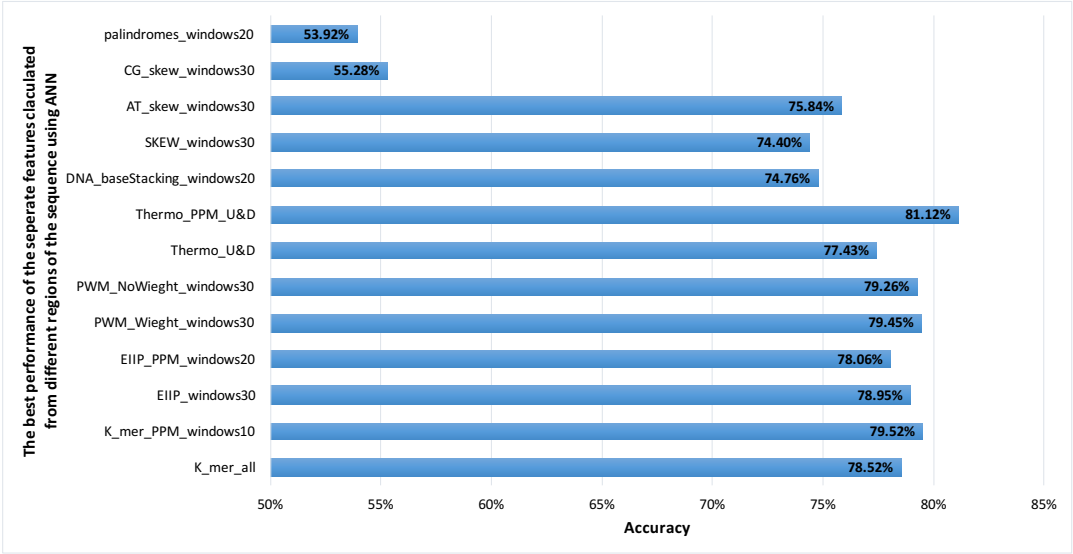

Fig. S8: ANN performance on PAS data (AATAAA) variant using different hand-crafted features calculated from different regions of the sequence

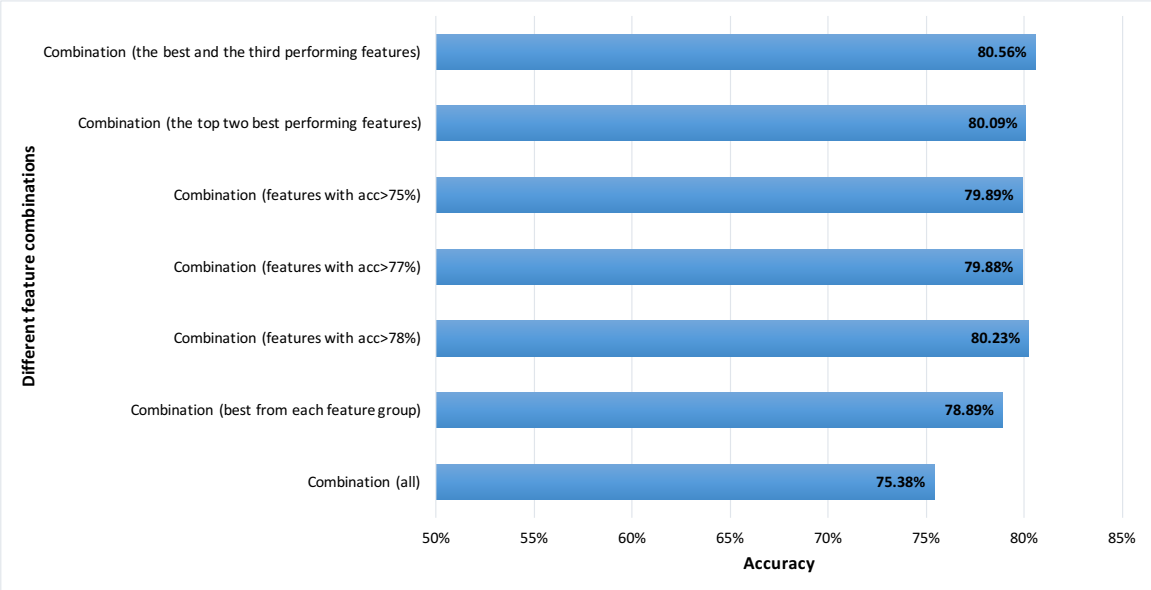

Fig. S9: ANN performance on PAS data (AATAAA) variant using different feature combination

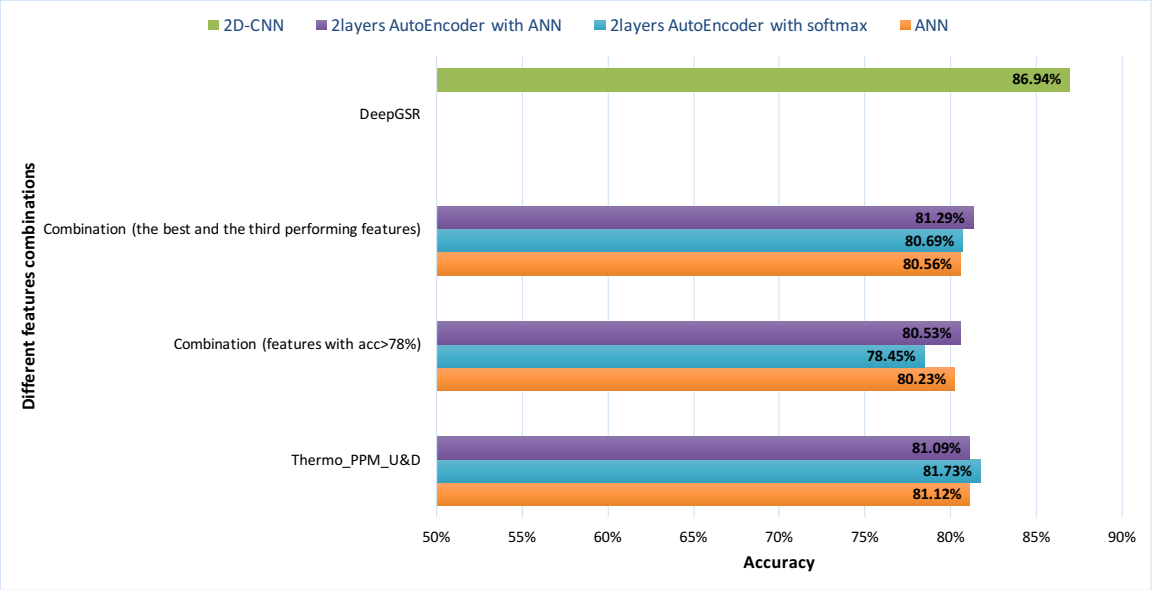

Fig. S10: Performance comparison on PAS data (AATAAA) variant between ANN and AE on the best performing features and feature combinations

Supplementary References

Abeel, T., *et al.* (2008) ProSOM: core promoter prediction based on unsupervised clustering of DNA physical profiles. *Bioinformatics* 24(13):i24-31.
Aken, B.L., *et al.* (2016) The Ensembl gene annotation system. *Database (Oxford)* 2016.
Beaudoing, E., *et al.* (2000) Patterns of variant polyadenylation signal usage in human genes. *Genome research* 10(7):1001-1010.
Cenik, C., *et al.* (2010) Genome-wide functional analysis of human 5' untranslated region introns. *Genome Biol* 11(3):R29.
Friedel, M., *et al.* (2009) DiProDB: a database for dinucleotide properties. *Nucleic acids research* 37(Database issue):D37-40.
Gramates, L.S., *et al.* (2017) FlyBase at 25: looking to the future. *Nucleic acids research* 45(D1):D663-D671.
Lee, J.Y., Park, J.Y. and Tian, B. (2008) Identification of mRNA polyadenylation sites in genomes using cDNA sequences, expressed sequence tags, and Trace. *Methods Mol Biol* 419:23-37.
Nair, A.S. and Sreenadhan, S.P. (2006) A coding measure scheme employing electron-ion interaction pseudopotential (EIIP). *Bioinformation* 1(6):197-202.
Proudfoot, N.J. (2011) Ending the message: poly(A) signals then and now. *Gene Dev* 25(17):1770-1782.
Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26(6):841-842.
Strausberg, R.L., *et al.* (1999) The mammalian gene collection. *Science* 286(5439):455-457.
Team, M.G.C.P., *et al.* (2009) The completion of the Mammalian Gene Collection (MGC). *Genome research* 19(12):2324-2333.
Tian, B., *et al.* (2005) A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic acids research* 33(1):201-212.
Tian, B., Pan, Z. and Lee, J.Y. (2007) Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing. *Genome research* 17(2):156-165.