

# Issues on Sampling Negative Examples for Predicting Prokaryotic Promoters


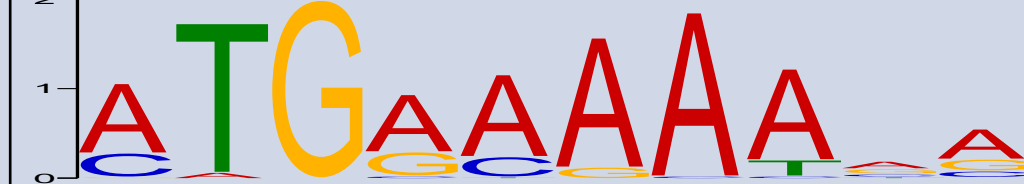






Eduardo G. Gusmão\* and Marcilio C. P. de Souto<sup>+</sup>\* IZKF Aachen Comp. Bio. Research Group/RWTH Aachen Univ. , Germany <sup>+</sup> LIFO/Univ. Orléans, France

**Issues on sampling negative examples.** In many research fields, experimental identification of negative examples can be laborious, expensive or unfeasible. Particularly in bioinformatics, this problem appears in several areas such as prediction of mRNAs that are target of miRNAs, regulatory networks, protein-protein interactions, non-coding RNA finding, among others. In the context of prokaryotic promoter prediction, various definitions of negative examples have been made. Here, we study the impact of different negative dataset definitions in the context of prokaryotic promoter prediction.

**Experimental Design.** First, we obtained an experimentally verified positive dataset for E. coli in RegulonDB. Then, we defined/created common representations of negative datasets in the literature. Finally, we made several experiments using a representative set of classification techniques in two different scenarios:

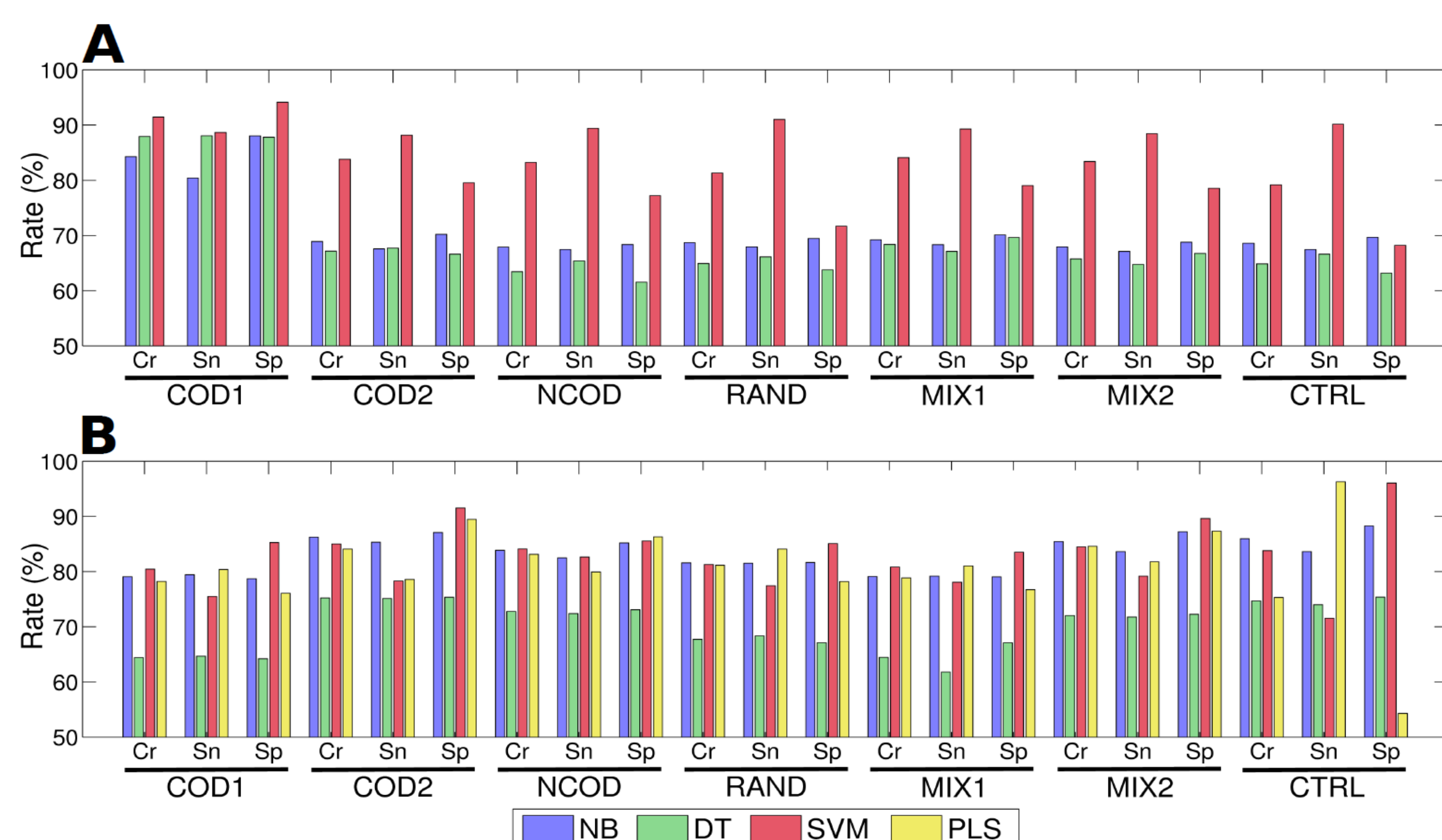
- **Sequence:** DNA sequences (categorical attributes).
- **vw Z-curve:** Numerical features extracted from the DNA sequences using the variable-window (vw) Z-curve method.

## Datasets

Dataset	MEME Top-Enriched Motif	Nucleotide Freq. (%)				Description
		A	C	G	T	
POS		29.04	20.48	20.00	30.48	Known (experimentally verified) promoters from RegulonDB
COD1		26.62	22.29	24.88	26.21	Start of E. coli's coding regions
COD2		24.19	24.58	27.21	24.02	Random part within E. coli's coding regions
NCOD		23.94	25.01	26.78	24.27	E. coli's non coding regions (convergent intergenic spacers)
RAND		24.46	25.79	25.34	24.41	Random non-promoter regions within E. coli's genome
MIX1		25.47	23.35	25.83	25.35	50% COD1 + 50% NCOD
MIX2		24.02	24.69	27.14	24.15	50% COD2 + 50% NCOD
CTRL		24.62	25.42	25.37	24.59	Completely random sequences given E. coli's nucleotide frequencies

## First Case Study

Performance assessment with the usual 10-fold cross-validation procedure. Results are shown for sensitivity (Sn), specificity (Sp) and accuracy (Cr) of all classifiers.



## Second Case Study

Performance assessment when training/testing with different negative datasets. Results are shown for the specificity of the SVM classifier.

			Testing						
			COD1	COD2	NCOD	RAND	MIX1	MIX2	CTRL
			Sequence	Sequence	Sequence	Sequence	Sequence	Sequence	Sequence
Training	Sequence	COD1	94,14	39,43	38,30	35,83	68,80	41,93	31,26
		COD2	52,79	79,55	78,08	75,11	64,07	87,39	69,93
		NCOD	50,36	77,98	77,21	72,87	73,33	87,87	68,81
		RAND	47,98	75,73	73,67	71,67	58,39	74,80	66,06
		MIX1	95,25	67,61	82,39	63,88	79,04	75,76	58,40
		MIX2	54,90	87,79	87,83	73,88	70,40	78,55	69,47
		CTRL	48,22	73,74	74,76	71,93	59,75	74,27	68,23
Training	vw Z-curve	COD1	85,29	91,91	84,88	82,14	91,84	88,29	72,98
		COD2	71,44	91,51	85,16	80,11	77,84	92,66	62,13
		NCOD	61,69	84,84	85,58	80,50	79,72	92,08	56,43
		RAND	69,00	87,39	89,25	85,10	78,84	88,81	69,77
		MIX1	87,33	91,06	93,60	84,96	83,49	91,96	68,15
		MIX2	68,76	94,00	91,85	80,25	79,57	89,64	64,84
		CTRL	63,18	75,73	79,52	79,27	70,17	76,12	96,06

