Improving the Accuracy of Classifiers for the Prediction of Translation Initiation Sites in Genomic Sequences

Experimental Results

Machine Learning and Knowledge Discovery Group
Department of Informatics,
Aristotle University of Thessaloniki,
Thessaloniki 54124, Greece
http://mlkd.csd.auth.gr

Abstract. The prediction of the Translation Initiation Site (TIS) in a genomic sequence is an important issue in biological research. Although several methods have been proposed to deal with this problem, there is a great potential for the improvement of the accuracy of these methods. Due to various reasons, including noise in the data as well as biological reasons, TIS prediction is still an open problem and definitely not a trivial task. We follow a three-step approach in order to increase TIS prediction accuracy. In the first step, we use a feature generation algorithm we developed. In the second step, all the candidate features, including some new ones generated by our algorithm, are ranked according to their impact to the accuracy of the prediction. Finally, in the third step, a classification model is built using a number of the top ranked features. We experiment with various feature sets, feature selection methods and classification algorithms and we compare with alternative methods.

This paper presents the detailed experimental results of our study on the prediction of TISs in genomic sequences.

Table 1. Measures of cross validation performance (TP: True Positives, FP: False Positives, TN: True Negatives, FN: False Negatives)

Sensitivity (TP Rate)	$\frac{TP}{TP + FN}$
Specificity (TN Rate)	$\frac{TN}{TN + FP}$
Precision	$\frac{TP}{TP + FP}$
Accuracy	$\frac{TP + TN}{TP + FP + TN + FN}$
Adjusted Accuracy	$\frac{Sensitivity + Specificity}{2}$

Table 2. The basic features considered in our study

Features in [1]	*New Features Proposed*	*Best Features Selected*
POS_-3	DOWN_IN_POS_2_T	POS_-3
UP_ IN_ATG	DOWN_IN_POS_3_C	UP_ATG
DOWN_ IN_CTG	DOWN_ IN_POS_1_G	UP_ IN_ATG
DOWN_ IN_TAA	UP_DOWN_A/G_DIF	DOWN_IN_STOP
DOWN_ IN_TAG	UP_DOWN_C/T_DIF	DOWN_IN_POS_2_T
DOWN_ IN_TGA		DOWN_IN_POS_3_C
DOWN_ IN_GAC		DOWN_ IN_POS_1_G
DOWN_ IN_GAG		UP_DOWN_A/G_DIF
DOWN_ IN_GCC		UP_DOWN_C/T_DIF

The following pages present a table (Table 3) and a number of graphs (Figure 1 – Figure 5) comparing the performance of the three classifiers we used. Three feature sets are included: the features proposed in [1] (denoted as *ZENG02*), the features proposed in [1] along with the new features we propose (denoted as *ZENG02 + Extra*) and the best features selected (denoted as *Best*). The experiments were repeated, once including the distance feature (*DIST*), once including the order feature (*ORDER*) and once including none of the above two features (in the graphs is denoted as *Simple*).

Table 3. Classification accuracy of the classifiers using 10-fold cross validation for a window length of 189 nucleotides and the features presented in Table 2

Features	*Algorithm*	*Sensitivity*	*Specificity*	*Precision*	*Adj. Accuracy*	*Accuracy*
ZENG02	C4.5	93.78	72.79	91.38	83.29	88.63
	RIPPER	92.52	75.36	92.03	83.94	88.31
	Naïve Bayes	85.77	83.49	94.11	84.63	85.21
ZENG02 + Extra	C4.5	94.95	80.64	93.78	87.80	91.44
	RIPPER	94.83	83.74	94.72	89.29	92.11
	Naïve Bayes	85.75	91.17	96.76	88.46	87.08
Best	C4.5	97.09	85.65	95.42	91.37	94.28
	RIPPER	96.66	86.77	95.74	91.71	94.23
	Naïve Bayes	90.58	90.32	96.64	90.45	90.52
ZENG02 + DIST	C4.5	96.33	88.48	96.26	92.40	94.40
	RIPPER	95.83	88.95	96.39	92.39	94.14
	Naïve Bayes	87.49	87.52	95.57	87.50	87.50
ZENG02 + Extra + DIST	C4.5	96.73	89.11	96.47	92.92	94.86
	RIPPER	96.15	90.23	96.80	93.19	94.70
	Naïve Bayes	85.73	91.54	96.89	88.63	87.15
Best + DIST	C4.5	98.07	93.07	97.75	95.57	96.84
	RIPPER	97.62	93.08	97.75	95.35	96.51
	Naïve Bayes	89.41	90.65	96.71	90.03	89.72
ZENG02 + ORDER	C4.5	95.08	76.29	92.50	85.69	90.47
	RIPPER	94.89	76.56	92.57	85.72	90.39
	Naïve Bayes	85.40	87.77	95.55	86.59	85.98
ZENG02 + Extra + ORDER	C4.5	95.71	81.12	93.98	88.42	92.14
	RIPPER	95.34	83.55	94.69	89.44	92.45
	Naïve Bayes	85.56	91.20	96.76	88.38	86.94
Best + ORDER	C4.5	97.04	85.63	95.41	91.34	94.24
	RIPPER	96.56	86.89	95.77	91.72	94.19
	Naïve Bayes	87.59	90.23	96.50	88.91	88.24

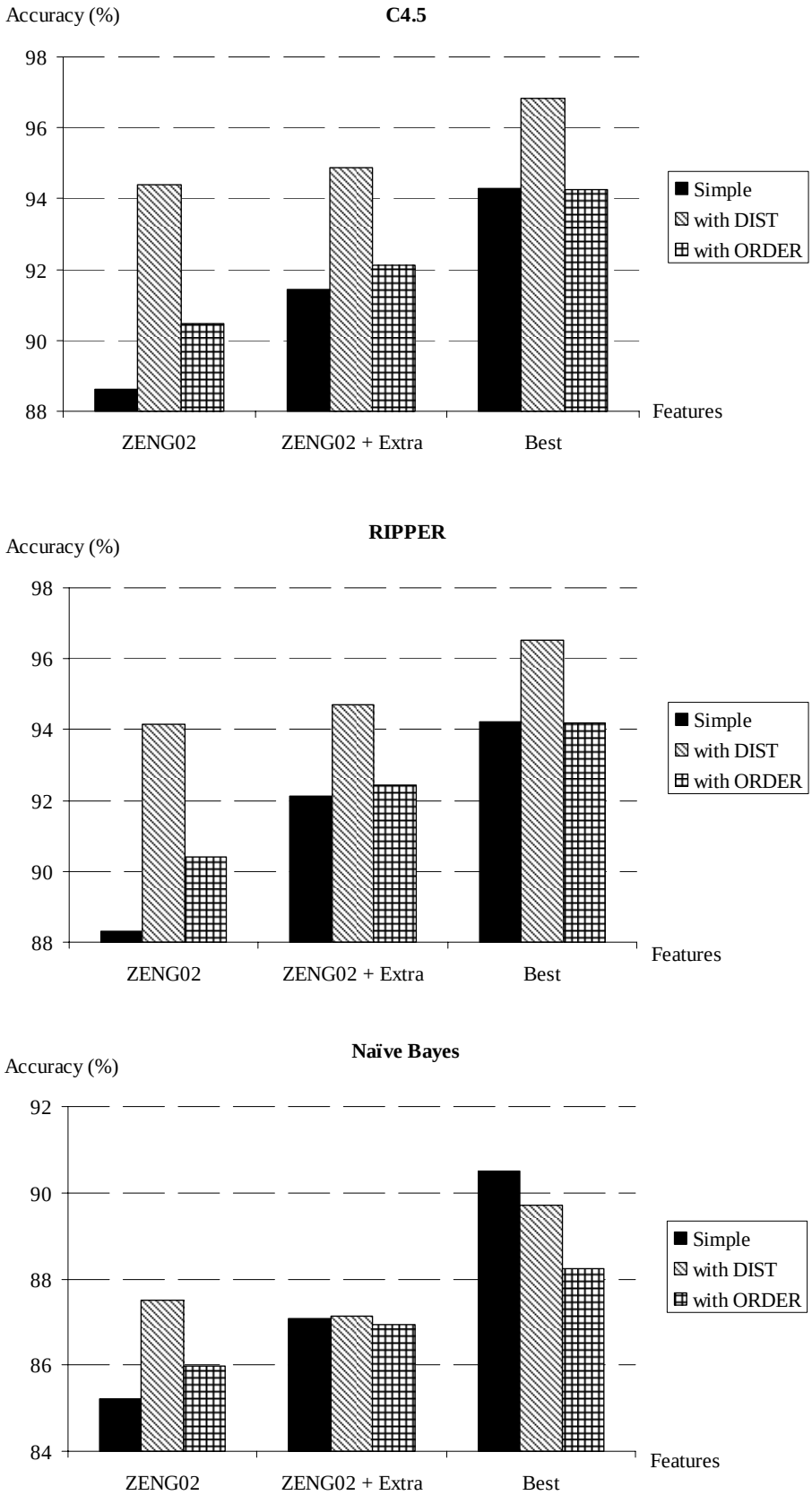

Figure 1. The accuracy graphs

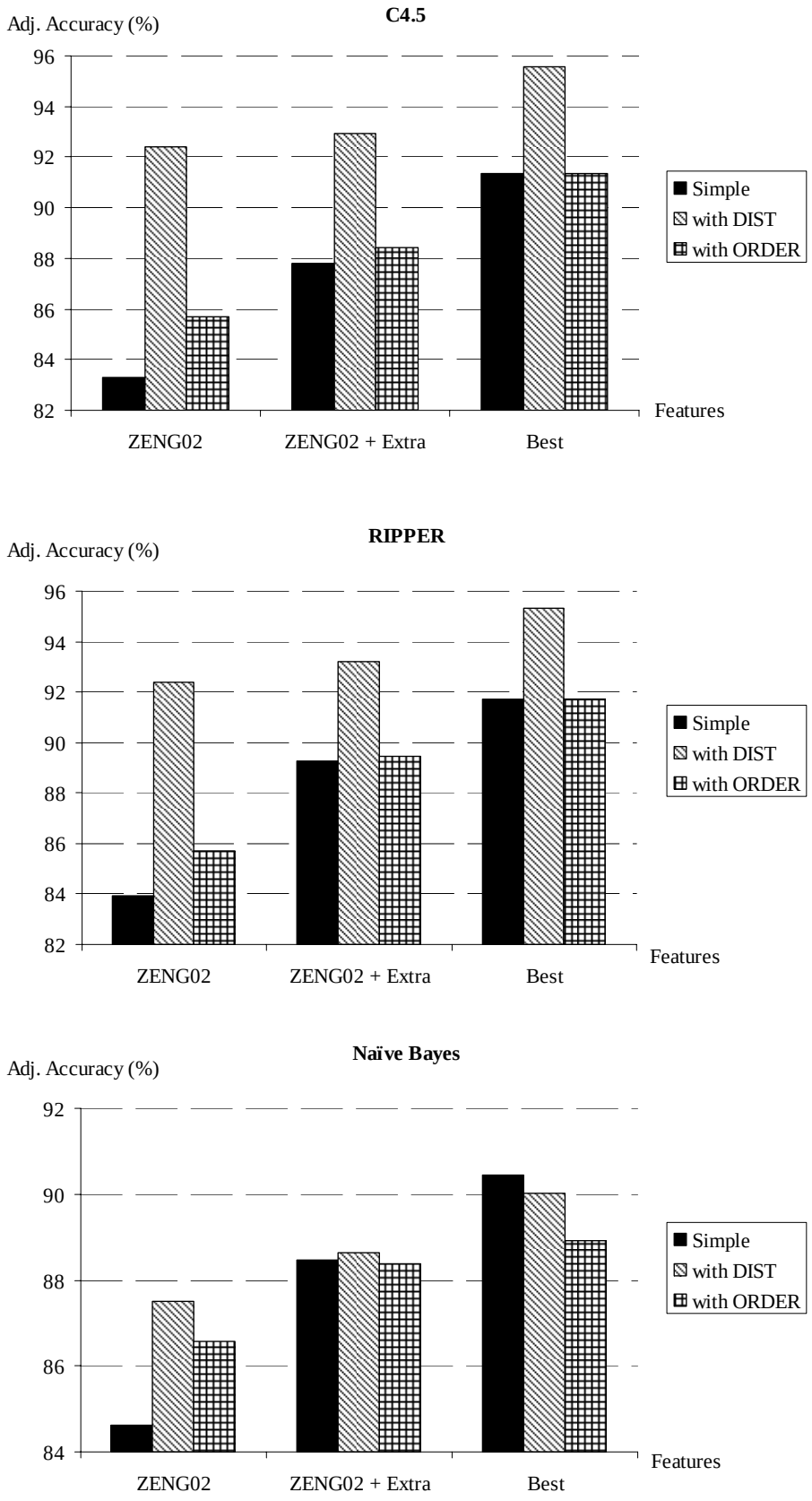

Figure 2. The adjusted accuracy graphs

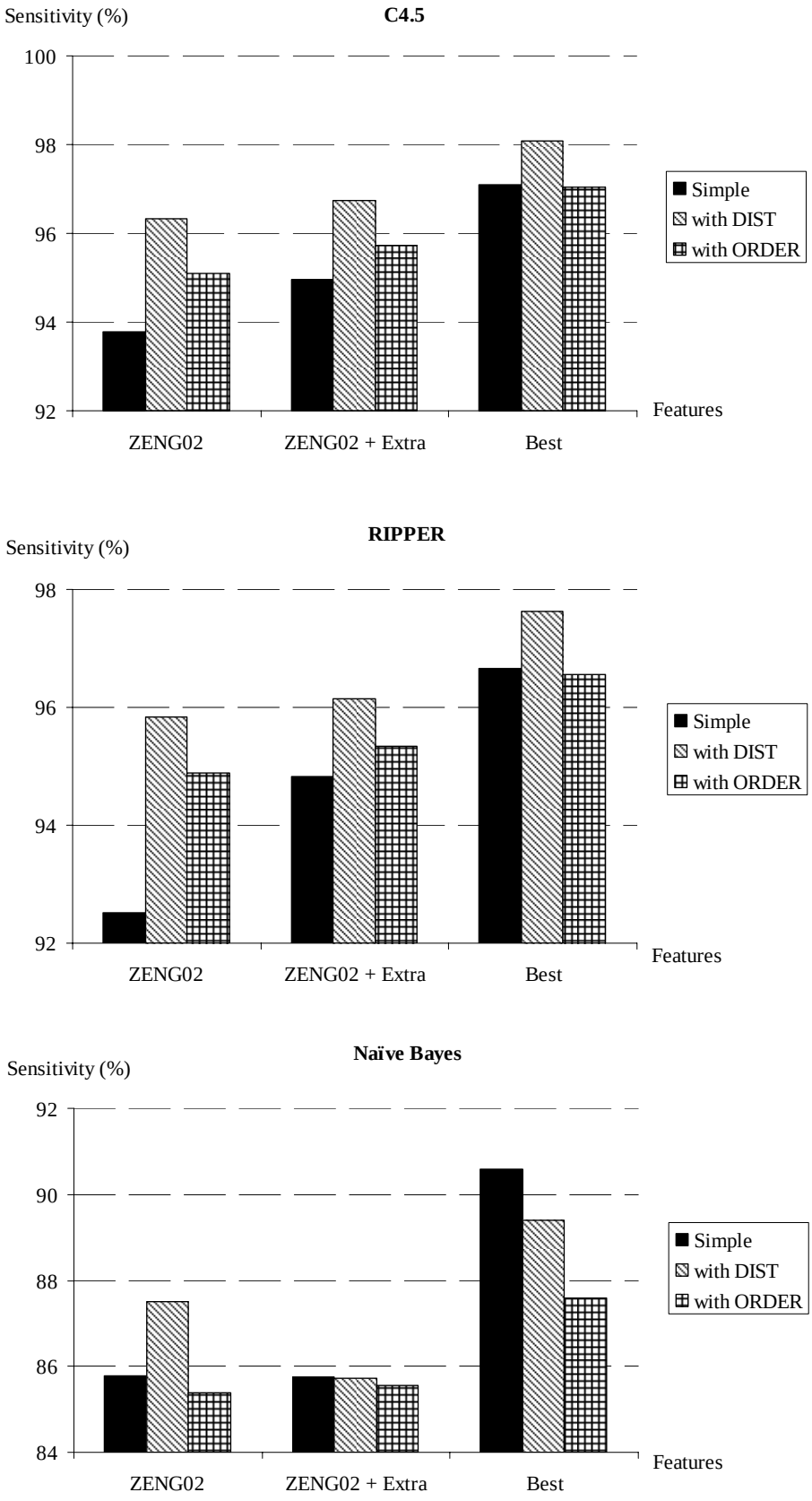

Figure 3. The sensitivity graphs

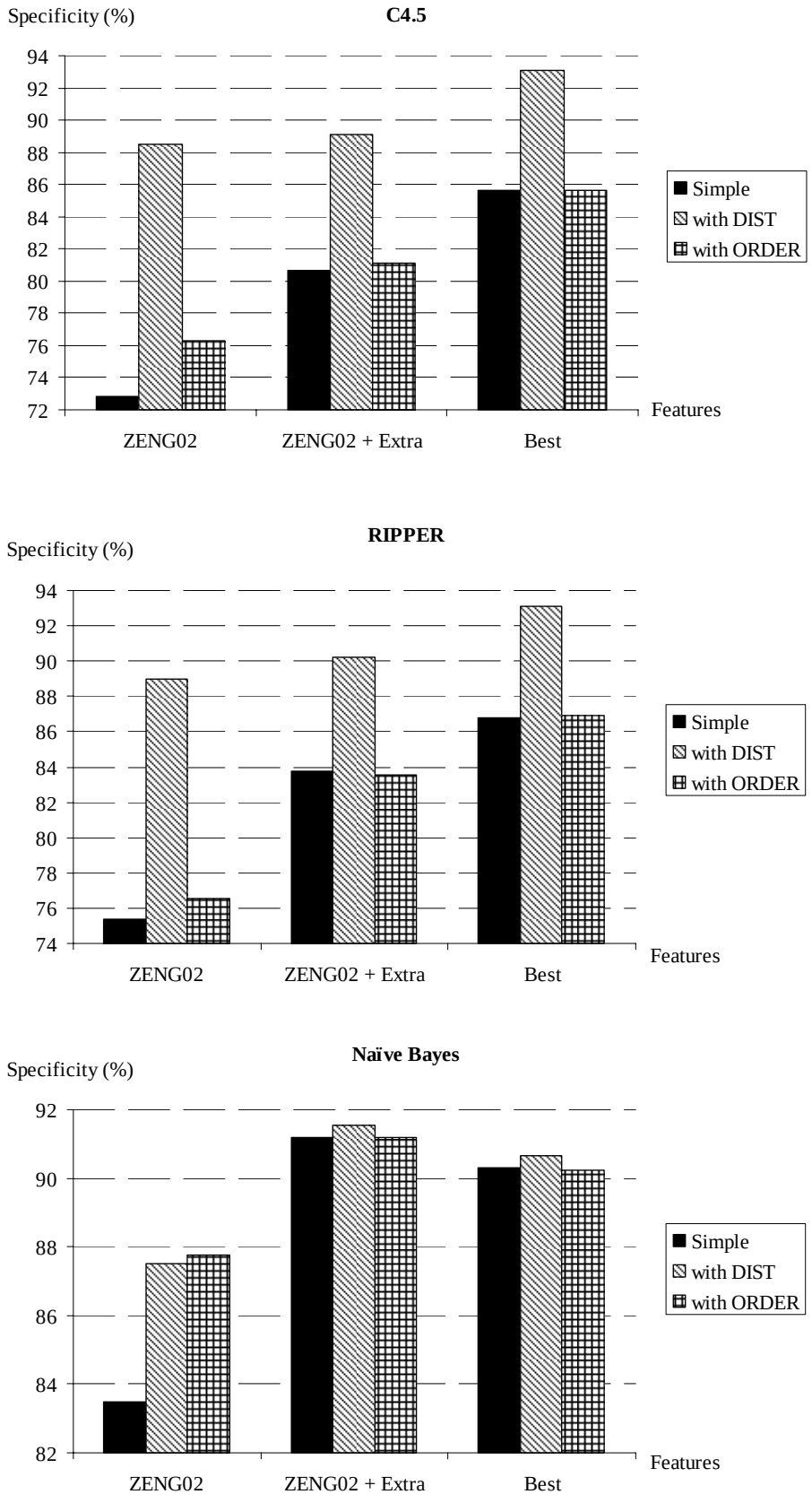

Figure 4. The specificity graphs

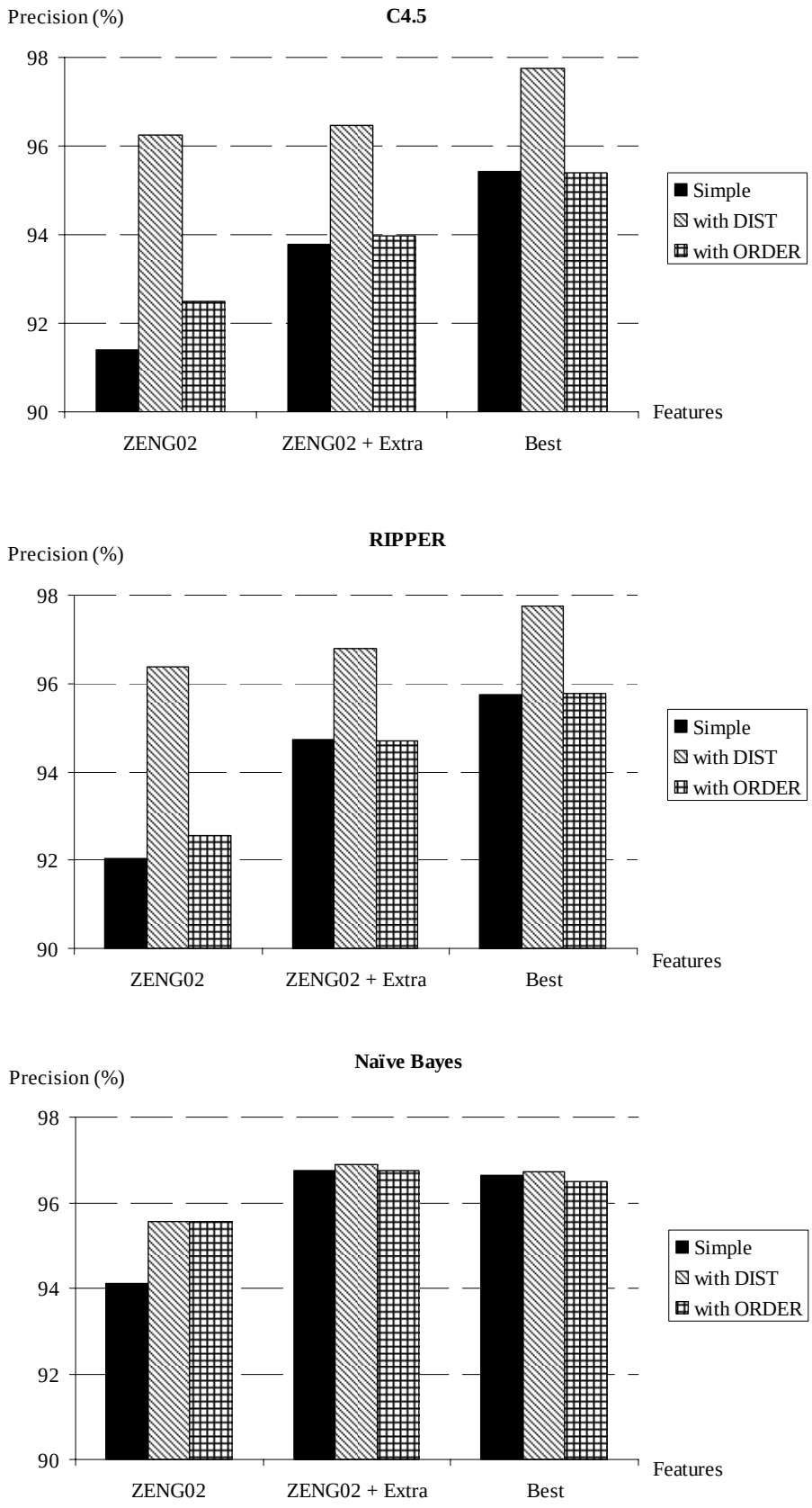

Figure 5. The precision graphs

References

1. Zeng F., Yap H., Wong, L.: Using Feature Generation and Feature Selection for Accurate Prediction of Translation Initiation Sites. In Proceedings of the 13th International Conference on Genome Informatics, Tokyo, Japan (2002) 192-200