

## Sensitivity and Specificity of FIRSTEF for Promoter Prediction

FIRSTEF is a first-exon prediction program published in Ref. [1], and since the 5' end of the first exon corresponds to the transcription start site (TSS), FIRSTEF may be used as a promoter prediction program as well. While the first-exon prediction accuracy of FIRSTEF is studied in detail in Ref. [1], the promoter prediction accuracy of FIRSTEF is studied only marginally in that publication. Here (i) we present details of the analysis of the promoter prediction accuracy of FIRSTEF carried out in Ref. [1] (analysis A), and (ii) we present two additional analyses of the promoter prediction accuracy of FIRSTEF with varying cutoff values [2] for the *a posteriori* probabilities  $P(\text{promoter})$ ,  $P(\text{exon})$ , and  $P(\text{promoter})$  (analyses B and C). Analyses B and C serve the goal of obtaining an estimate of the FIRSTEF promoter prediction sensitivity ( $S_n$ ) and specificity ( $S_p$ ) as a function of those cutoff values.

All three analyses are based on the following two criteria of measuring the promoter prediction accuracy:

- Criterion 1 defines a predicted promoter region as true positive (TP) if the transcription start site (TSS) is located within or up to 200 bp downstream of the predicted promoter region, and it defines a predicted promoter region as false positive (FP) otherwise [3].
- Criterion 2 defines a promoter prediction as TP if the annotated TSS falls within the  $-2000$  bp to  $+500$  bp region of the predicted promoter, and it defines a promoter prediction as FP if the predicted promoter falls within the ATG + 500 bp to STOP codon region [4].

Analyses A and B are based on criterion 1, and analysis C is based on criterion 2. Analyses A and B are based on a set of 58 experimentally verified promoters (see Appendix A) on human chromosome 22, and analysis C is based on the Sanger Center annotation of human chromosome 22.

For analysis A, we follow the protocol presented in appendix B, and we obtain that FIRSTEF predicts 86 promoters in total, out of which 46 (40) predictions are TPs (FPs), yielding a sensitivity ( $S_n$ ) of 79.3% and a specificity ( $S_p$ ) of 53.5% [6]. In order to compare FIRSTEF to the best currently available promoter prediction program, PROMOTERINSPECTOR [3], we follow the same protocol, and we obtain that PROMOTERINSPECTOR predicts 65 promoters in total, out of which 28 (37) predictions are TPs (FPs), yielding a sensitivity ( $S_n$ ) of 48.3% and a specificity ( $S_p$ ) of 43.1%. These numbers are published in Ref. [1].

For analysis B, we consider the FIRSTEF predictions on both strands and post-process the output of FIRSTEF such that only the locations of the promoters, but not their orientations, are predicted. By eliminating the prediction of the promoter orientation of FIRSTEF we make the promoter predictions of FIRSTEF directly comparable to the promoter predictions of PROMOTERINSPECTOR, which predicts only the locations of the promoters, but not their orientations. Aside from this modification of step 4, all other steps of the protocol presented in appendix B are identical. Table I(a) and Figure 1(a) show that for FIRSTEF  $S_n$  decreases from 74% to 57% and  $S_p$  increases from 30% to 67% as  $p$  is varied from 0.4 to 1.0.

For analysis C, we simply apply criterion 2 to the promoter predictions of FIRSTEF in order to determine the number of TPs and FPs for different values of  $p$ . Table I(b) and Figure 1(b) show that for FIRSTEF the TP percentage increases from 24% to 64% and the FP percentage decreases from 14% to 3% as  $p$  is varied from 0.4 to 1.0.

Table I(a)

Program	Cutoff Probability $p$	Number of Predictions	TP	FP	$S_n$	$S_p$
PROMOTERINSPECTOR		65	28	37	48.3	43.1
FIRSTEF	1.00	49	33	16	56.9	67.3
FIRSTEF	0.99	59	36	23	62.1	61.0
FIRSTEF	0.80	72	38	34	65.5	52.8
FIRSTEF	0.70	80	38	42	65.5	47.5
FIRSTEF	0.60	91	40	51	69.0	44.0
FIRSTEF	0.50	105	40	65	69.0	38.1
FIRSTEF	0.40	143	43	100	74.0	30.1

Table I(b)

Program	Cutoff Probability $p$	Number of Predictions	TP	FP	TP%	FP%
PROMOTERINSPECTOR		497	204	84	41.0	16.9
FIRSTEF	1.00	389	251	15	64.5	3.8
FIRSTEF	0.99	495	283	27	57.2	5.4
FIRSTEF	0.90	617	299	43	48.5	7.0
FIRSTEF	0.80	758	318	59	42.0	7.8
FIRSTEF	0.70	929	332	89	35.7	9.6
FIRSTEF	0.60	1135	347	128	30.6	11.3
FIRSTEF	0.50	1448	374	187	25.8	12.9
FIRSTEF	0.40	1631	388	221	23.8	13.6

TABLE I. Analysis of the promoter prediction accuracy of PROMOTERINSPECTOR and FIRSTEF, based on two different criteria of measuring the accuracy of promoter predictions. The results of Table I(a) are based on criterion 1, and the results of Table I(b) are based on criterion 2.

In order to visualize the results of Table I, we show in Figure 1 two plots of TP versus FP for both PROMOTERINSPECTOR and FIRSTEF.

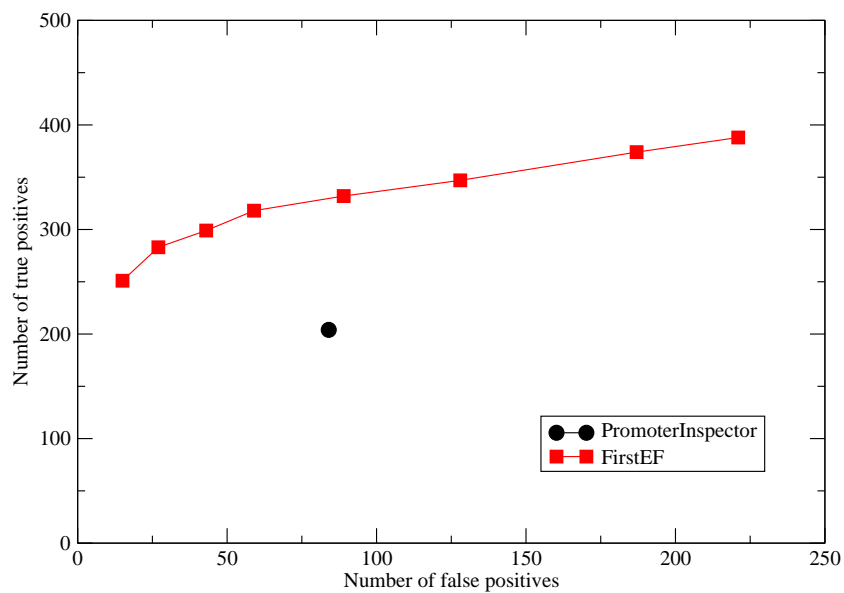
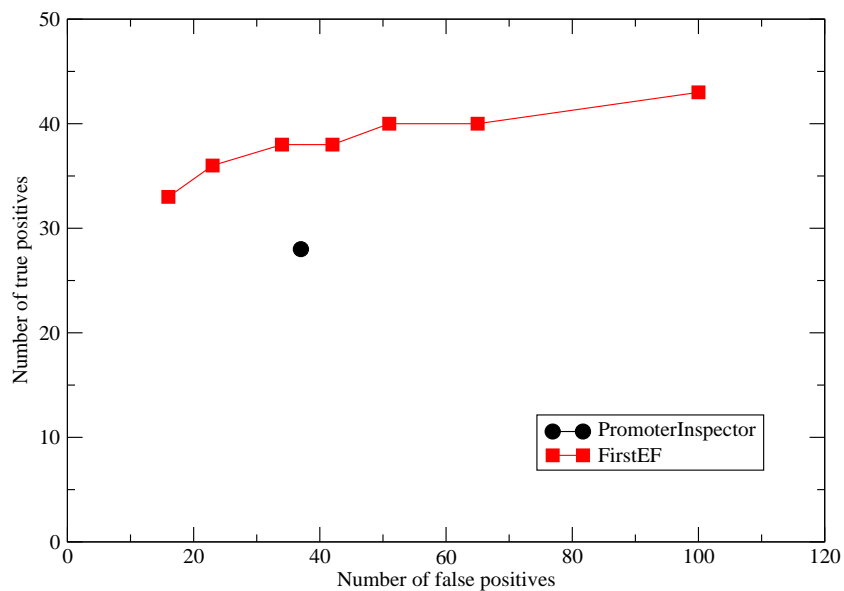


FIG. 1. TP versus FP based on criterion 1 (Figure 1(a)) and based on criterion 2 (Figure 1(b)) for PROMOTERINSPECTOR (circles) and FIRSTEF (squares).

The FIRSTEF program as well as all of the output parsers used in this study are freely available at <http://www.cshl.org/mzhanglab>.

- 
- [1] Davuluri, R. V., Grosse, I. and Zhang, M. Q. Computational identification of promoters and first exons in the human genome. *Nature Genetics* **29**, 412–417 (2001).
- [2] Specifically, we vary the three cutoff values for  $P(\text{promoter})$ ,  $P(\text{exon})$ , and  $P(\text{donor})$  as follows: we denote by  $p$  a real number ranging from 0.4 to 1, we set the cutoff values for  $P(\text{promoter})$  and  $P(\text{exon})$  equal to  $p$ , and we keep the cutoff value for  $P(\text{donor})$  constant at 0.4.
- [3] Scherf, M., Klingenhoff, A. and Werner, T. Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. *J. Mol. Biol.* **297**, 599–606 (2000); Scherf, M., Klingenhoff, A., Frech, K., Quandt, K., Schneider, R., Grote, K., Frisch, M., Gailus-Durner, V., Seidel, A., Brack-Werner, R. and Werner, T. First pass annotation of promoters on human chromosome 22. *Genome Research* **11**, 333–340 (2001).
- [4] Scherf, M. and Werner, T. Private communication (2002).
- [5] TSS location means the TSS location on the April-2001 freeze of the Golden-Path assembly of human chromosome 22.
- [6] We use the following definitions for sensitivity and specificity:  $S_n = \frac{TP}{TP+FN}$  and  $S_p = \frac{TP}{TP+FP}$ .

**APPENDIX A: GENBANK IDS AND TSS LOCATIONS OF THE 58 FIRST EXONS USED FOR ANALYSES A AND B**

GenBank ID	TSS location	GenBank ID	TSS location	GenBank ID	TSS location
AF169360_CAR	15830549	L41944_IFNAR2	31460041	AF106656_ADSL	37271101
AF072557_ZNF74	17400176	J03171_IFNAR	31555037	NM_015705_dJ1042K10.2	37295232
S81003_UBE2L3	18570627	U05875_AF-1	31632965	Y14392_SH3BGR	37678661
Z31682_VL-IL	19101715	X63071_DBP-5	31773129	U53707_PCP4	38093523
X52828_BCR	20171250	AB028942_KIAA1019	31780609	X55639_MX1	39652625
Y17118_SNF5-INI1	20775336	AF027153_SLC5A3	32303634	AB038162_TFF1	40615086
NM_001355_DDT_(alt)	20968068	S58267_HMOX1	32387195	L14577_CBS	41324416
U40770_ADORA2A	21475134	X00371_MB	32570743	AB001523_TMEM1	42222745
X60069_GGT1_(alt)	21651965	U85267_DSCR1	32756882	AB001517_PWP2	42317781
AF049891_TPST2	23632057	S69002_AML1-EVI1	33118770	NM_015653_MGC4107	42359009
Y07848_GAR22	26349030	AF019225_APOL1	33206539	AB006684_AIRE1	42496270
AF165426_NF2	26645560	U54558_EIF3S7	33453797	AP001754_PFKL	42510441
AF129855_OSM	27272952	X51678_parvalbumin	33744118	AP001754_TRPC7	42563646
X85237_SF3A1	27363025	AF069543_CSF2RB	33838276	Z50022_PTTG1IP	43084149
AB029900_CST	27580610	X59434_TST	33948824	Z95976_hRED1	43284900
AF047576_TCH	27613272	X53093_IL2RB	34074869	X15879_COL6A1	44189235
AF115551_SMTN	28087416	Z93096_MFNG	34410980	M59486_S100B	44812579
AB016665_LIMK2b	28254499	AF115252_PLA2G6	35106358	NM_014577_BRD1	46722830
Z82248_YWHAH	28950619	AU143698_TTC3	35303133		
U14394_TIMP3	29806909	AF121002_GRAP2	36871414		

TABLE II. GenBank IDs and TSS locations [5] of the 58 genes used for the comparison of PROMOTERINSPECTOR and FIRSTEF.

## APPENDIX B: PROTOCOL FOR ANALYSIS A

The following protocol was used to evaluate the promoter prediction accuracies of PROMOTERINSPECTOR and FIRSTEF for analysis A (published in Ref. [1]).

1. Run FIRSTEF on both strands of chromosome 22 of the April-2001 freeze of the Golden Path assembly of the human genome.
2. Mark on chromosome 22 of the April-2001 freeze of the Golden Path assembly of the human genome the positions of the 58 experimentally verified TSSs (see Appendix A).
3. Extract all PROMOTERINSPECTOR and FIRSTEF predictions that fall into the 58 regions starting 20 kb upstream of the TSSs and ending 20 kb downstream of the first-exon splice-donor sites.
4. Since FIRSTEF predicts not only the location of the predicted promoter, but also its orientation, consider only those FIRSTEF predictions with the same orientation as the true promoter.
5. Apply criterion 1 to determine for each prediction of PROMOTERINSPECTOR and FIRSTEF if it is TP or FP.

## APPENDIX C: ANALYSIS OF CHROMOSOME 20

In order to test if the accuracy of FIRSTEF obtained for chromosome 22 can be reproduced for chromosome 20, we repeat the analysis yielding Table I(a) for chromosome 20.

Cutoff Probability	Number of Predictions	TP	FP	TP%	FP%
1.00	448	291	15	64.96	3.34
0.99	604	337	33	55.79	5.46
0.90	780	352	73	45.13	9.36
0.80	927	360	89	38.83	9.60
0.70	1120	378	128	33.75	11.43
0.60	1381	395	170	28.60	12.31
0.50	1758	419	231	23.83	13.13
0.40	1985	438	262	22.07	13.20

TABLE III. Analysis of human chromosome 20, based on criterion 2, using the same parameter values as for the analysis of chromosome 22 (see Table I). We find that the results of this Table are consistent with the results of Table I, indicating that the accuracy of FIRSTEF observed in Table I and Figure 1 for chromosome 22 is reproducible for chromosome 20.