FProm

Human promoter prediction

Method description:

Program predicts potential transcription start positions by linear discriminant function combining characteristics describing functional motifs and oligonucleotide composition of these sites. FProm uses file with selected factor binding sites from currently supported functional site data base.

For approximately 50-55% level of true promoter region recognition, FProm program will give one false positive prediction for about 4000 bp.

Another promoter recognition program, TSSG, uses promoter.dat file with selected factor binding sites (TFD, Ghosh,1993).

Prediction accuracy for each promoter type Promoter Type A: TATA-less promoter

Sensitivity	Specificity	Threshold*	Length**
1.000000	0.198215	-9.496	1.32975
0.990000	0.646996	-6.025	3.02029
0.950000	0.917724	-2.414	12.9585
0.900000	0.968909	+0.0467	34.2921
0.800000	0.992493	+3.329	142.028
0.700000	0.997591	+5.342	442.657
0.600000	0.998801	+6.508	889.255
0.500000	0.999409	+7.621	1805.3
0.400000	0.999705	+8.596	3610.59
0.300000	0.999858	+9.598	7491.98
0.200000	0.999911	+10.66	11987.2
0.100000	0.999968	+12.14	33297.7

Promoter Type B: TATA promoter

Sensitivity	Specificity	Threshold*	Length**
1.000000	0.773441	-6.766	71.1151
0.990000	0.965914	-2.318	472.68
0.950000	0.996183	+1.117	4220.83
0.900000	0.998333	+2.528	9667.06
0.800000	0.999570	+4.613	37459.9
0.700000	0.999785	+6.41	74919.8/td>
0.600000	0.999839	+7.963	99893
0.500000	0.999946	+9.586	299679
0.400000	0.999946	+11.21	299679
0.300000	0.999946	+12.5	299679
0.200000	1.000000	+14.14	1e+06
0.100000	1.000000	+16.54	1e+06

*Threshold value used by the program for a giver level of sensitivity

**Average length which contains 1 false-positive promoter.

References:

1. Solovyev V.V., Salamov A.A. (1997)
The Gene-Finder computer tools for analysis of human and model organisms genome sequences.
In Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology (eds.Rawling C.,Clark D., Altman R.,Hunter L.,Lengauer T.,Wodak S.), Halkidiki, Greece, AAAI Press,294-302.

2. Solovyev V.V. (2001)
Statistical approaches in Eukaryotic gene prediction.
In Handbook of Statistical genetics (eds. Balding D. et al.), John Wiley & Sons, Ltd., p. 83-127.

3. Solovyev VV, Shahmuradov IA. (2003)
PromH: Promoters identification using orthologous genomic sequences. Nucleic Acids Res. 31(13):3540-3545.

FProm output:

FProm output:

```
Sequence    1 of    1,  Name: Homo sapiens chromosome 21; range 31946321 - 31958321; length
12001
Length of sequence:     12001
     7 promoter/enhancer(s) are predicted
Promoter Pos:      6473 LDF:   +8.734
Promoter Pos:      3102 LDF:   +5.824
Promoter Pos:      6078 LDF:  +16.297 TATA box at       6049    +5.597 TATAAAGT Enchancer
at:      5942 Score:  +12.499
Promoter Pos:      1363 LDF:   +5.235 TATA box at       1336    +6.514 AATAAAAG
Promoter Pos:      7068 LDF:   +1.165 TATA box at       7039    +4.190 TAAAAATA
Promoter Pos:      9650 LDF:   +1.051 TATA box at       9618    +4.491 GTTAAAAA
Promoter Pos:      5541 LDF:   +0.455 TATA box at       5512    +7.353 TATAAAAA
```

Where:

7 promoter/enhancer(s) are predicted	Number of predicted promoters in this sequence.
Each line below defines an appropriate predicted promoter. Detailed description of a line from this list is shown further: `6078 LDF:  +16.297 TATA box at       6049    +5.597 TATAAAGT Enchancer at: 5942 Score:  +12.499`	
Promoter Pos: 6078	Position of TSS on DNA.
LDF: +16.297	value of Fisher's linear discriminant for the current promoter. A bigger value corresponds to more reliable promoter.
If a promoter belongs to class of TATA-containing promoters, the following fields are added:	
TATA box at 6049	TATA-box position in the current promoter
+5.597	Score of this TATA-box
TATAAAGT	Nucleotide sequence of this TATA-box
If there is an enhancer in proximity to the current promoter, the following fields are added:	
Enchancer at: 5942	The position of enhancer in this promoter
Score: +12.499	Score of this enchancer

Parameters:

Input	
Sequence	Input file with sequence in FASTA-format
Output	
Result	Name of the output file
Print programm info	Print information about program accuracy. First and second type errors for each threshold value for each promoter type.