**s70\_scanner.pl**

A sigma 70 promoter finding algorithm using PWMs and UP-element scoring systems derived for ECF sigma promoters based on (Rhodius et al. 2011).

**Usage** (Perl script; requires Perl):

s70\_scanner.pl -f <seqfile> -o <outputfile> -z <z-score> (options)

**Required options with values:**

-f <seqfile> Input sequence file name. Sequence input either FASTA format or as a single line of sequence.

-o <outputfile> Output filename. Outputs as a tab delimited text file. Best viewed in Excel.

-z <z-score> Motif search cutoff threshold (must be a number, e.g. -1). This ensures a minimum match for the core -10 and -35 motifs, based on the average and standard deviation (z-score) PWM scores of these motifs in 674 experimentally determined sigma 70 promoters. Using a z-score of -1 (recommended) identifies 75% of known sigma 70 promoters. Given that many sigma 70 promoters are activator-dependent, this is probably a reasonable cutoff to identify constitutive sigma 70 promoters.

**Options with values:**

-p <z-score> Promoter cutoff threshold (must be a number, e.g. -1). Ensures a minimum score for the whole promoter. This will reduce false-positives and likely restrict hits to just constitutive promoters. Using a value of -1 identifies 84% of known sigma 70 promoters (that have a minimum -10 and -35 scores of -1 [-z -1]).

**Options with no values:**

-b Search both strands (default is forward strand only). In output file, hits on the reverse strand are denoted by "\_r" after the sequence name.

-h Output highest score only for each sequence. Useful If you only want to identify the highest scoring promoter on a particular sequence.

-u Scores upstream sequence for UP-element (default only scores core promoter motifs [-35, spacer, -10, discriminator and start]). Scores UP-element as a 22 nt window directly upstream of the -35 (approx -57 to -36) by counting the number of overlapping 3 nt A- or T-tracts. Scoring the UP-element will more accurately stimulate promoter strength.

-a Outputs all predictions (default removes lower scoring predictions that share -10 or -35 motifs). Sigma 70 predictions often identify multiple overlapping sites that often share one motif. Therefore to simplify the predictions the default setting only outputs the highest scoring overlapping prediction. However, if the aim is to re-engineer a sequence to remove all potential sigma 70 promoters, it is useful to output all predictions so that adjacent overlapping weaker promoters are also removed.

**Recommended settings to identify constitutive promoters:**

s70\_scanner.pl -f <seqfile> -o <outputfile> -z -1 -p -1 -u

**Output file:**

Tab delimited text file, best viewed in Excel or equivalent. Header section (proceeded by #) summarizes input settings. Column headers as follows:

#Sequence Sequence name (if FASTA input file) or file name (if single line sequence input). <Sequence name\_r> denotes reverse strand if -b option selected.

UP sequence Upstream sequence from -57 to -37. Sequence only displayed if -u option selected.

UP-score Score of upstream sequence (counts of overlapping 3 nt A- and T-tracts). Only displayed if -u option selected.

-35 pos Distance (nt) of 5' end of -35 element from 5' end of sequence. Note for reverse strand sites (-b option) this is distance from 5' end of reverse strand.

-35 seq Sequence of -35 element (-36 to -30: consensus aTTGaca).

-35 score PWM score of -35 element.

spacer sequence Sequence of spacer region (-21 to -13; consensus tttatntGg). This includes the putative β' zipper contact (-21 to -18) and the extended -10 (TG) motif at -15/-14. The remainder of the spacer sequence has not been included since this has not been implicated in RNAP contacts.

spacer score PWM score of spacer sequence.

-10 pos Distance (nt) of 5' end of -10 element from 5' end of sequence. Note for reverse strand sites (-b option) this is distance from 5' end of reverse strand.

-10 seq Sequence of -10 element (-12 to -7: consensus TAtaaT).

-10 score PWM score of -10 element.

spacer len Distance (nt) between -35 and -10 element (spacer length). Most common distance is 17 nt.

spacer pen Penalty score for suboptimal spacer lengths. Based on observed frequencies of spacer lengths.

start seq Sequence of start motif (-1/+1: consensus c/t(-1) a/g(+1)).

start score PWM score of start motif.

disc len Distance (nt) between -10 element and start motif (discriminator length). Most common distance is 5 nt. Note this is 1 nt shorter than in the literature, which calculates discriminator length from the -10 to +1.

disc pen Penalty score for suboptimal discriminator lengths. Based on observed frequencies of discriminator lengths.

Total score Total promoter score. Calculated as:

Total score = (UP model) + (PWM-35) + (PWMspacer (-21 to -13)) + (PWM­-10) + (PWMstart(-1/+1)) + (Spacer penalty) + (Disc penalty)

Note, UP model score is only included if -u option selected.

Total zscore Z-score of total promoter score, based on mean () and standard deviation (σ) scores of experimentally determined promoters. The z-score provides a relative measure of promoter score (e.g. -1 is considered a weak hit; +1 a strong hit). Only promoters that contained -35 and -10 elements > zscore -1 were used to calculate z-scores (506 in total). For core promoters (no UP-element score):  = 5.75, σ = 1.86. For complete promoters (including UP element score -u option):  = 7.70, σ = 2.80.

Note: The motif positions are *relative* due to the flexible spacer and discriminator lengths. E.g., the -35 element described as -36 to -30 is assuming the most common spacer and discriminator lengths (17 nt and 5 nt, respectively).

**Brief model notes:**

Sigma 70 promoter database from RegulonDB 11-28-07. Contains 674 experimentally confirmed sigma 70 promoters (mapped +1 starts). Used (Shultzaberger et al. 2007) as a guide for fitting -10 and -35 hexamers. Performed independent searches for -10 and -35 using Bioprospector 1 block function: -10 searched as a 6mer within window -16 to -5 (allows for discriminators from 4-10, allowing for start site mapping errors). -35 searched for 6 mer 15-20 nt upstream of spacer

**Datafiles:**

PWMs

35PWM.txt is PWM scores for 7mer block of -35 motif (aTTGaca). 2.77 bits

10PWM.txt is PWM scores for 6mer block of -10 motif (TAtaaT). 4.52 bits

SpPWM.txt is PWM scores for 10 mer block from -21 to -13 and is aligned with the -10. Includes putative Zn finger contact (-21 to -18; (Yuzenkova et al. 2011)), -17/-16 dyad and -15/-14 TG motif (Burr et al. 2000; Mitchell et al. 2003). 0.32 bits

StartPWM.txt is PWM scores for 2mer block of -1/+1. 0.23 bits

Spacer histograms

Alls70\_spacer\_his.txt is a histogram of distances between -35 and -10 motifs for 674 promoters. Column 1 is distance (nt), column 2 is observed frequencies.

Alls70\_disc\_his.txt is a histogram of distances between -10 and start (-1/+1) motifs for 674 promoters. Column 1 is distance (nt), column 2 is observed frequencies.

References

Burr, T., J. Mitchell, et al. (2000). "DNA sequence elements located immediately upstream of the -10 hexamer in Escherichia coli promoters: a systematic study." Nucleic Acids Res **28**(9): 1864-70.

Mitchell, J. E., D. Zheng, et al. (2003). "Identification and analysis of 'extended -10' promoters in Escherichia coli." Nucleic Acids Res **31**(16): 4689-95.

Rhodius, V. A., V. K. Mutalik, et al. (2011). "Predicting the strength of UP-elements and full-length E. coli sigmaE promoters." Nucleic Acids Res.

Shultzaberger, R. K., Z. Chen, et al. (2007). "Anatomy of Escherichia coli sigma70 promoters." Nucleic Acids Res **35**(3): 771-88.

Yuzenkova, Y., V. R. Tadigotla, et al. (2011). "A new basal promoter element recognized by RNA polymerase core enzyme." EMBO J **30**(18): 3766-75.