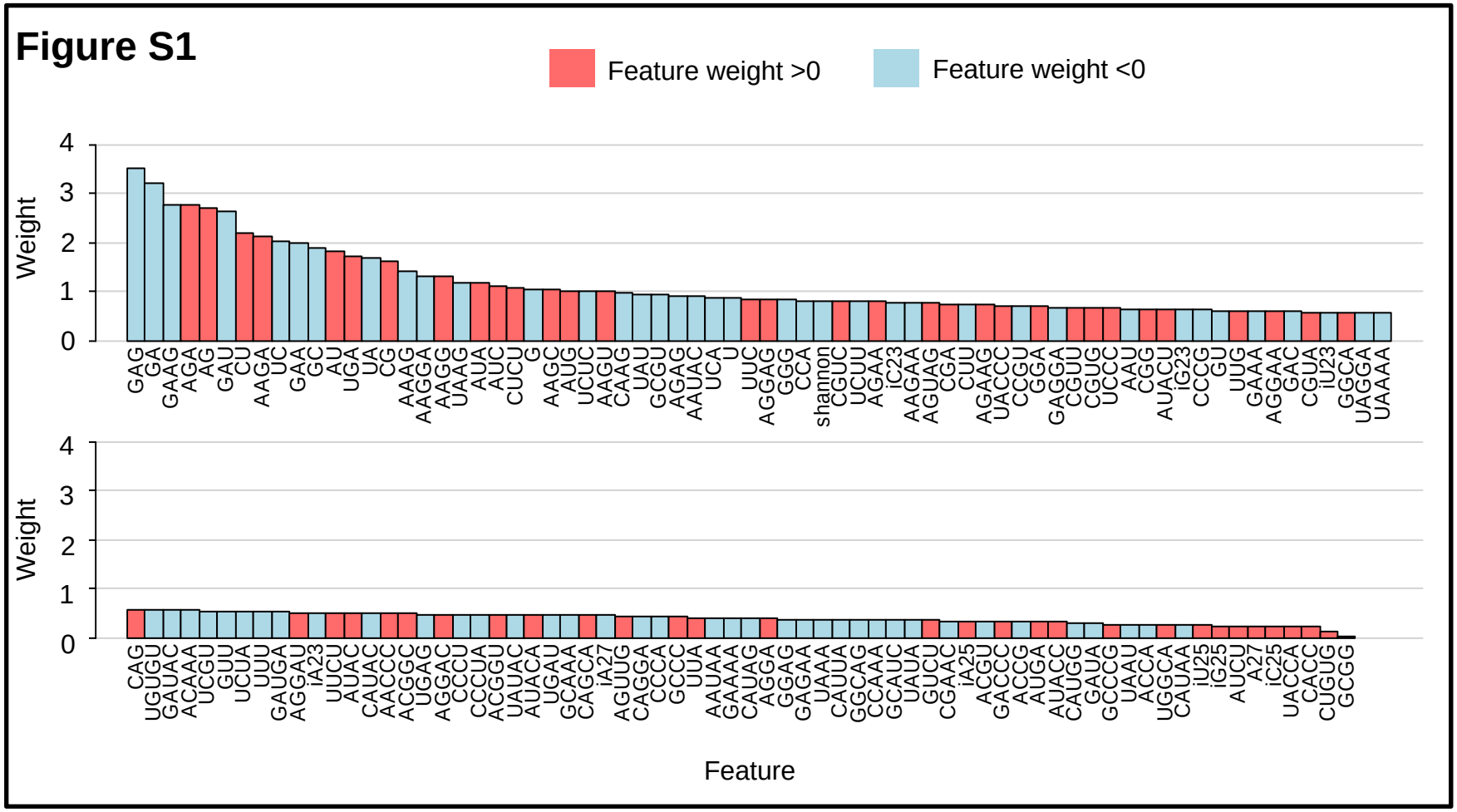

Figure S1. Features ranked by absolute value for their weights in the optimized classifier from layer 1 ("AGO" vs "noAGO"). Position specific base composition (PSBC) features are represented in the format LP, where $L \in \{A, C, G, U\}$ and P is the position of the nucleotide in the sequence starting the count at 5' or 3' in the case of PSBC v2, which is represented by iLP; Shannon: sequence information content given by Shannon entropy.

Figure S2

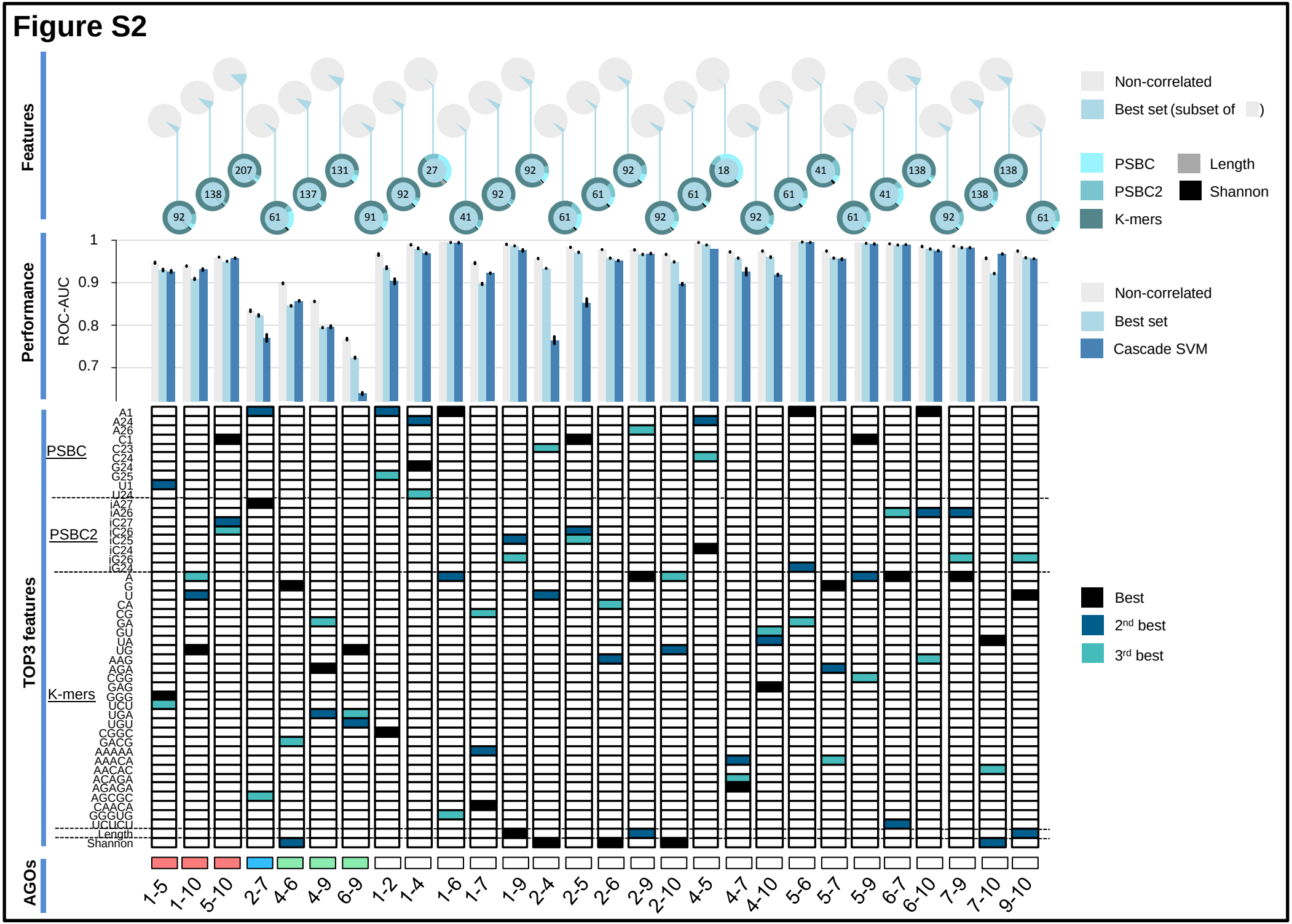
Features
Performance
TOP3 features
AGOs
ROC-AUC
1
0.9
0.8
0.7
PSBC
PSBC2
K-mers
A1
A24
A26
C1
C23
C24
G24
G25
U1
U24
iA27
iA26
iC27
iC26
iC25
iC24
iG26
iG24
A
G
U
CA
CG
GA
GU
UA
UG
AAG
AGA
CGG
GAG
GGG
UCU
UGA
UGU
CGGC
GACG
AAAAA
AAACA
AACAC
ACAGA
AGAGA
AGCGC
CAACA
GGGUG
UCUCU
Length
Shannon
1-5
1-10
5-10
2-7
4-6
4-9
6-9
1-2
1-4
1-6
1-7
1-9
2-4
2-5
2-6
2-9
2-10
4-5
4-7
4-10
5-6
5-7
5-9
6-7
6-10
7-9
7-10
9-10
Non-correlated
Best set (subset of)
PSBC
PSBC2
K-mers
Length
Shannon
Non-correlated
Best set
Cascade SVM
Best
2nd best
3rd best

Figure S2. Results for the models trained to discriminate sRNA from two different AGO-IP sets (layer 2): features selected, performance and top3 features. The top three rows of pie charts represent the features kept after the correlation analysis (grey section), a fraction of which comprises the optimal sets obtained by feature selection (light blue section). The three following rows of pie charts represent the composition of the features kept after feature selection, with the number in the middle representing the total number of features that survived SVM-RFE. For each binary model, ROC-AUC is shown for classifiers trained: using all non-correlated features (grey bars), after feature selection (light blue bars) and using the optimized feature set in a non-linear learning with the cascade SVM scheme (dark blue bar). Each bar plot has on top the standard deviation for the performance calculated from the 5-fold cross-validation procedure. The top3 features found with the feature selection procedure (features kept in the last 3 SVM-RFE iterations) are indicated in a 3rd section. The bottom section contains an indication of the AGO-IP sets used in each of the 28 models, plus a colour scheme that signs inter- (white) and intra-clade models (red: clade 1; blue: clade 2; green: clade 3).

Figure S3

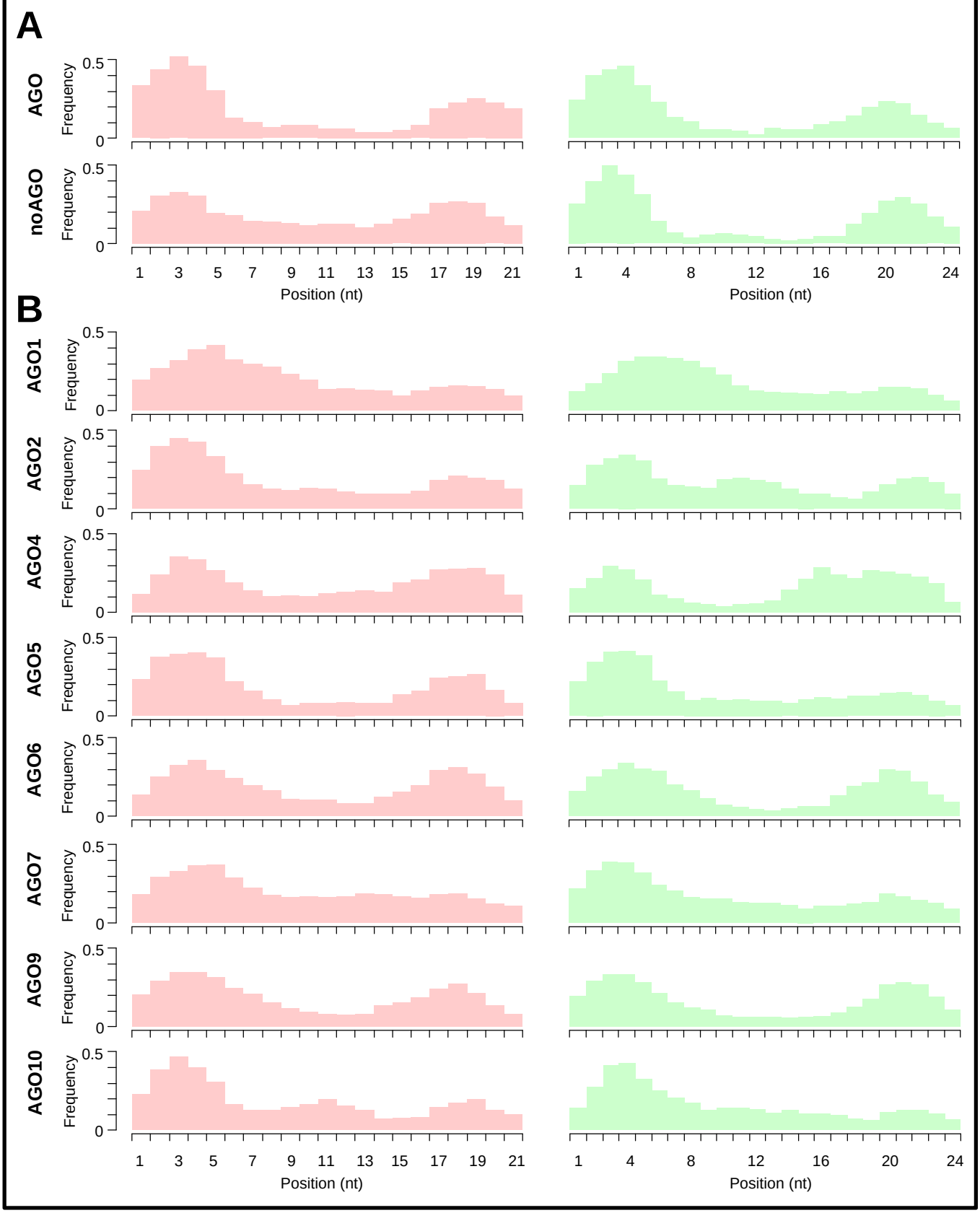

Figure S3. Distribution of k-mers retained in the classifiers according to their density peaks in 21 nt and 24 nt sRNA: (A) layer 1: AGO vs noAGO; (B) layer 2: mean of the distributions for classifiers involving a given AGO.

Figure S4

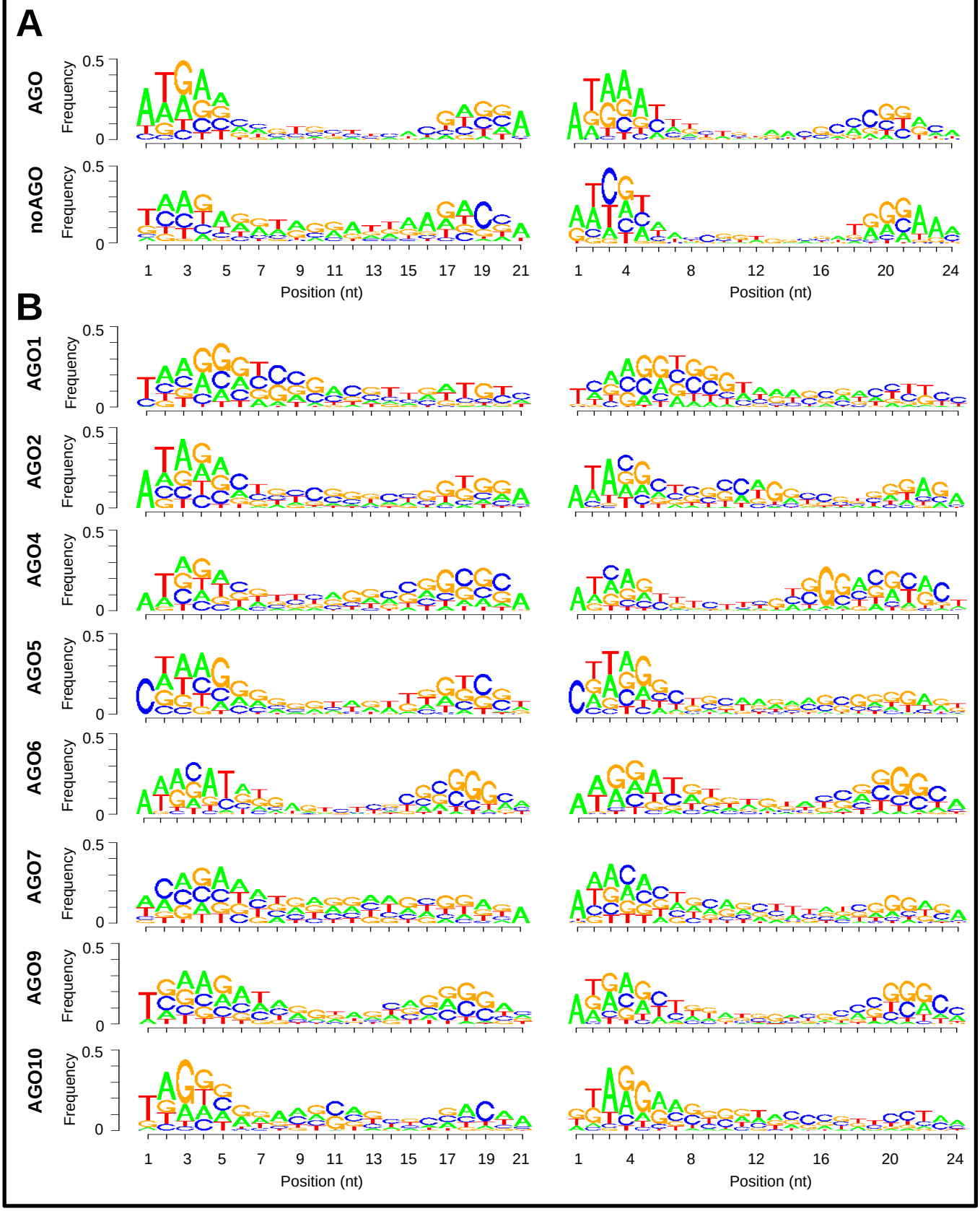

Figure S4. Frequency of nucleotides in k-mer across the sRNA libraries, determined from the distribution of k-mers retained in the classifiers according to their density peaks in 21 nt and 24 nt sRNA: (A) layer 1: AGO vs noAGO; (B) layer 2: mean of the frequencies for classifiers involving a given AGO.