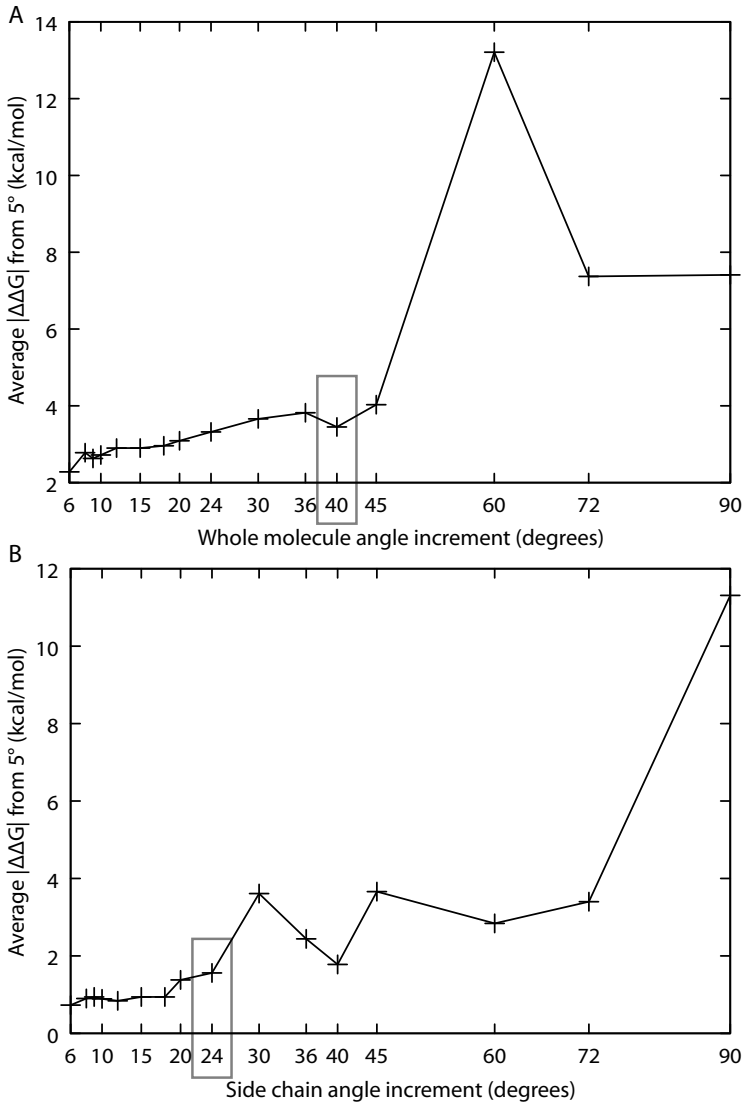

FIG. S1: Comparison of different rotation increments for amino acid rotations. (A) Whole molecule rotations (relative to the DNA) and (B) side chain χ_1 angle rotations. Rotations were done in various increments from 5° to 90° (for 360° rotations; increments were adjusted to give the same period for 180° rotations), and the average free energy difference (over the entire 108-point grid) against the 5° increment was computed. Optimal rotation angles were chosen by considering the largest angle (fewest structures to generate, so fastest computation time) that retained robust results (indicated in boxes). For (A), 40° ($2\pi/9$) was chosen, corresponding to a period of 9. The calculations for (A) were performed for glycine around the GGG DNA 3-mer, as this amino acid does not have a side chain so its results are independent of the subsequent analysis of side-chain rotation increment presented in part B. For (B), 24° ($2\pi/15$) was chosen, corresponding to a period of 15. The calculations for (B) were performed for aspartic acid around GGG using the 40° whole-molecule rotation determined from (A).

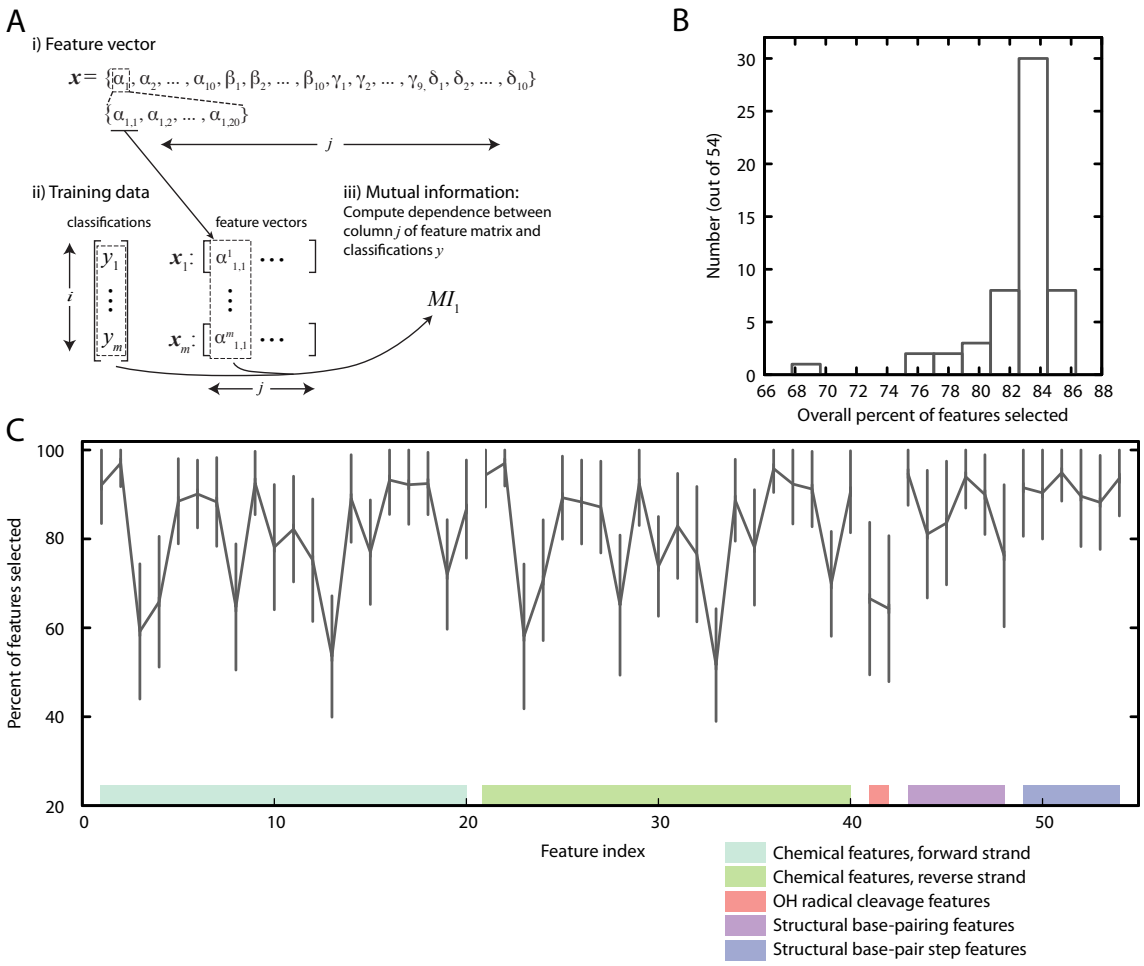

FIG. S2: Schematic and statistics of the feature selection step. (A) Schematic of the mutual information procedure, starting from the feature vector obtained at the end of Fig. 2 of the text. (i) Each component of the feature vector $\mathbf{x}$ has a number of associated scalar values. For example, each α is drawn from the list of chemical features and has 20 values: $\alpha_{k,l}$ is the lth chemical feature associated with the kth nucleotide in the binding site, and j runs over both k and l. (ii) The training data consists of the classification vector y and the m training sequences, indexed by i. The ordered list of training vectors can be considered a feature matrix, each column corresponding to a particular feature. (iii) Mutual information is computed between column j of the feature matrix and the classification vector by Eq. 9. (B) Distribution of the percent of features remaining after feature selection (maximum possible is 90%). (C) Analysis of the features retained by feature selection, as a function of the type of feature; percentages are averaged over the length of the binding sequences.

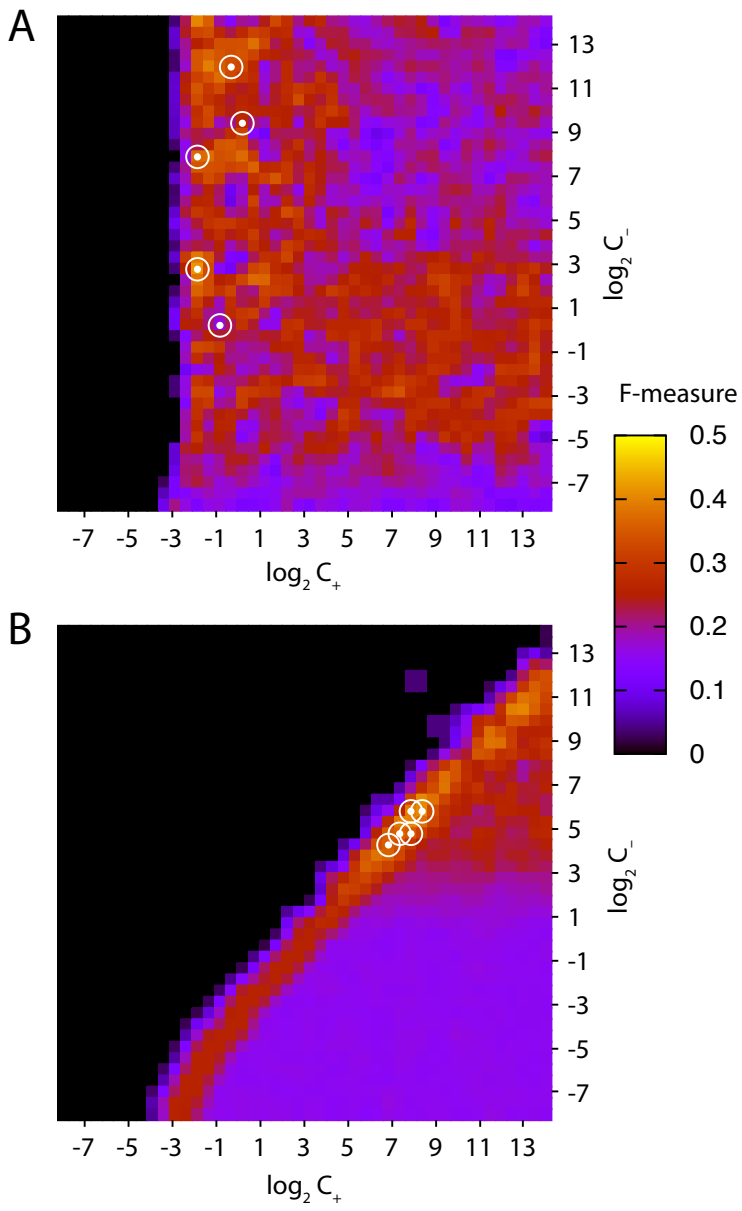

FIG. S3: Illustration of the parameter space searched by the optimization-through-cross-validation step. Two examples are shown: (A) DnaA, and (B) NanR, using SVM-PMM. Plotted is the F-measure at each possible pair of (C_+, C_-) over the fine grid spacing of $2^{0.5}$. White $\odot$ symbols indicate the values obtained in the 5 training runs performed. The optimal parameter space is less broadly distributed for NanR, reflected in more consistent parameter choices. Also, note that although the initial coarse grid is from 2^{-5} to 2^{11} for C_+ and C_-, the maximal grid is from 2^{-8} to 2^{14}. For example, if 2^{-5} is the optimal value for C_+ in the first search, the second grid is from 2^{-7} to 2^{-3}, and if 2^{-7} is the optimal choice in the second search, the third grid is from 2^{-8} to 2^{-6}, so C_+ could be chosen as low as 2^{-8}.

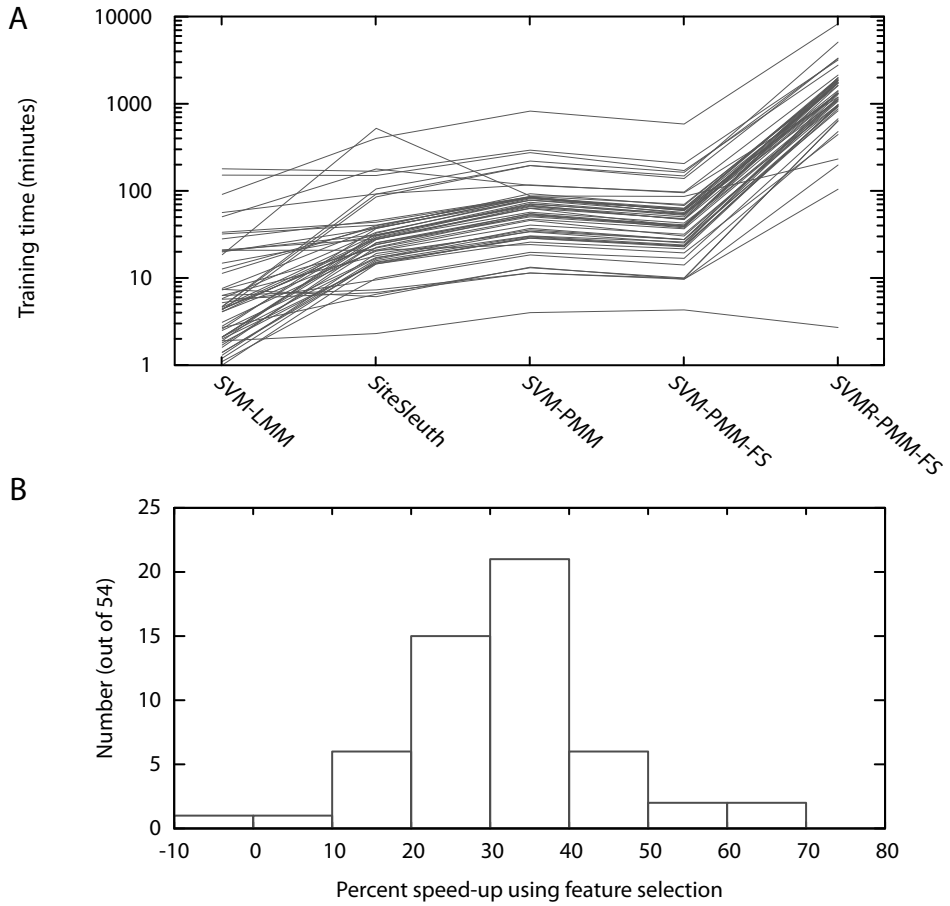

FIG. S4: Training time results. (A) Training times showing the trend across SVM method for each TF. (B) Histogram of speed-ups from the feature selection step, comparing training times for SVM-PMM and SVM-PMM-FS. In addition to increasing accuracy, the feature selection step also reduced training time for all but 1 of the 54 TFs studies, in some cases by over 60%. Average speed-up is 32%. Percent speed-up is defined as $100 \times (t_{\text{SVM-PMM}}/t_{\text{SVM-PMM-FS}} - 1)$.

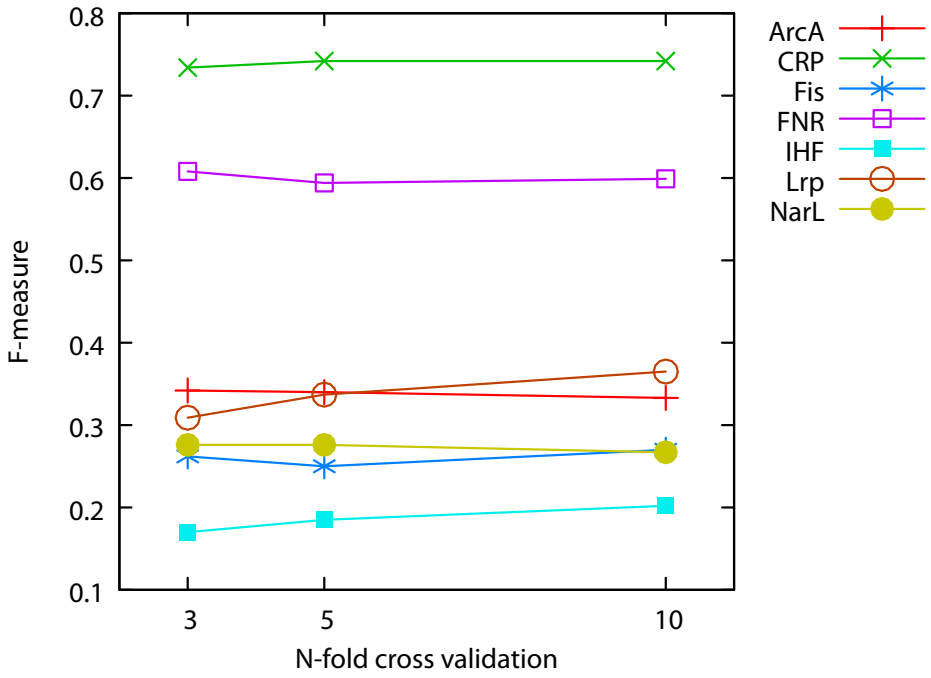

FIG. S5: Comparison of 3-fold, 5-fold, and 10-fold cross validation for the seven TFs with at least 80 positive training sequences. Values are the average F-measure over five SVM-PMM-FS training runs. See Table S1 for the number of positive training sequences for each TF.

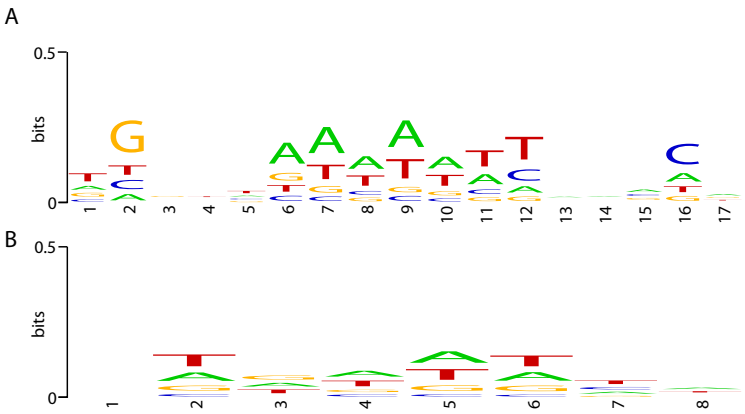

FIG. S6: Sequence logo representations of Fis (A) and Lrp (B) PWM motifs. Neither TF's binding sites are well captured with a PWM: note that the y-axis in both panels has been expanded for better visibility (typical motif representations range up to 2 bits). PWMs were generated from our training data for these TFs using WebLogo, as noted in the Results.

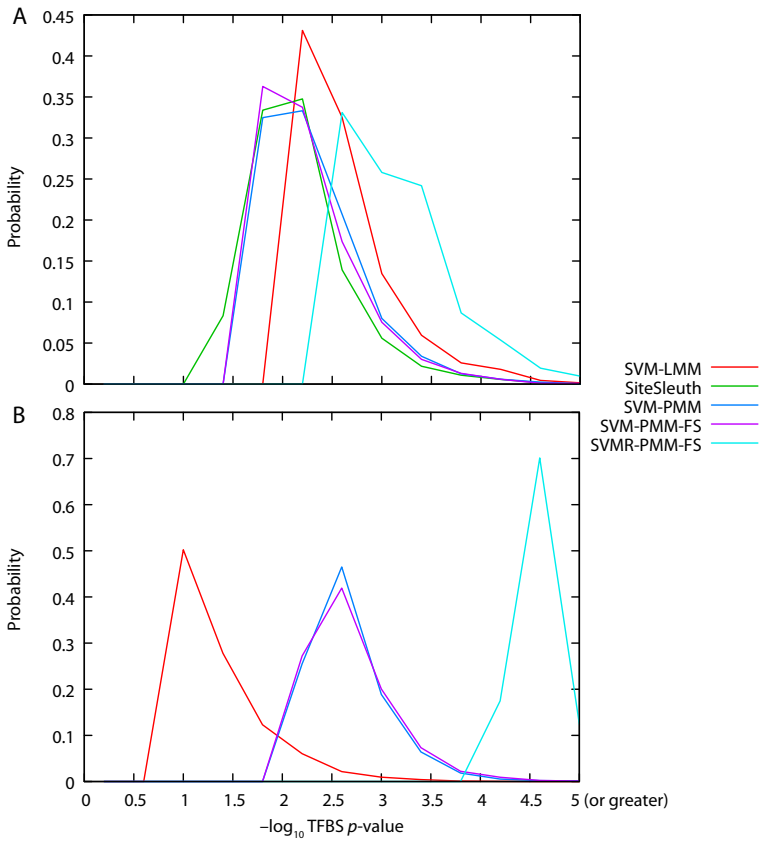

FIG. S7: Distribution of TFBS p-values from different SVM methods. Shown are histograms of p-values for Fis (A) and Lrp (B) in negative log-scale (horizontal axis). There is no histogram for SiteSleuth (green curve) for Lrp because the method predicted 0 TFBSs. p-values were computed using 10,000 random DNA sequences as the null model, so the smallest p-value we can estimate is 10^{-5}. Distributions are pooled over the 5 prediction runs for each TF and model pair.

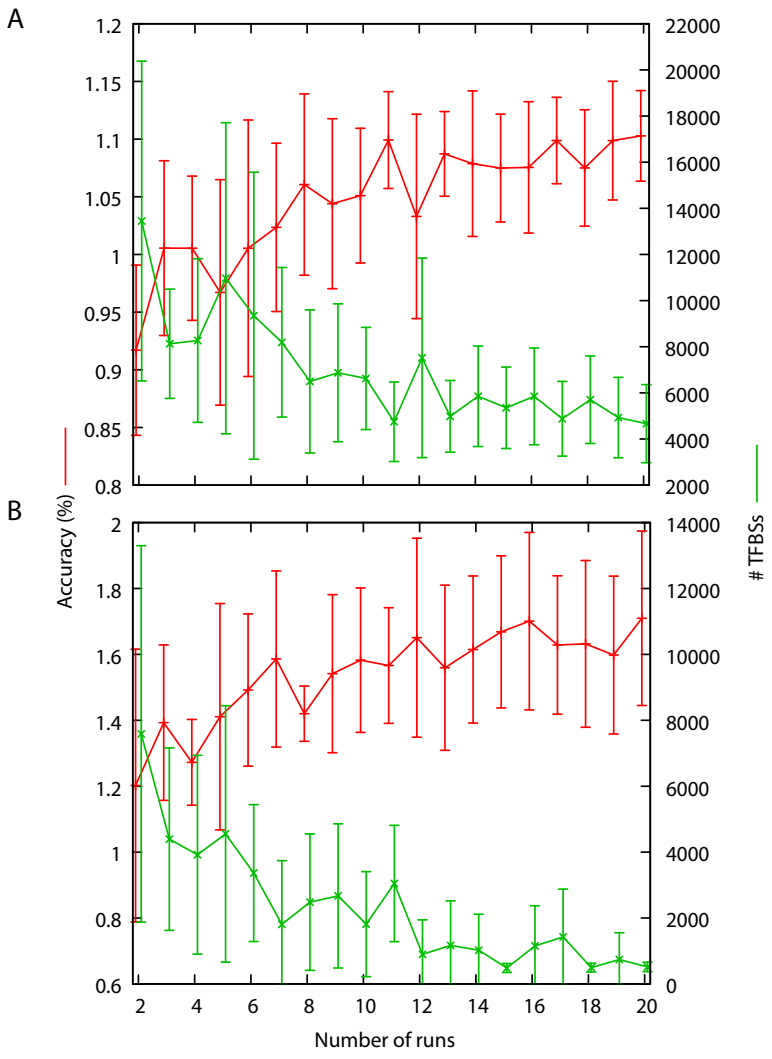

FIG. S8: Accuracy and number of TFBS from consensus predictions using different numbers of runs. For Fis (A) and Lrp (B), we plot the mean ± standard deviation of accuracy and number of predicted TFBSs for each n from 2 to 20, computed as described in the text.

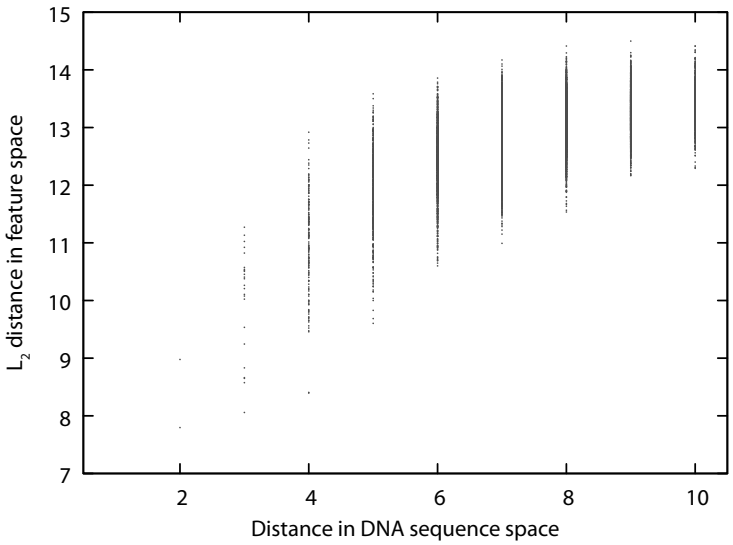

FIG. S9: Comparison of DNA sequence space with feature space. We generated 10,000 random pairs of DNA sequences of length 10 and compared their distances in DNA sequence space and feature space. Three G flanking nucleotides were added to each sequence, so that variations in the flanking sequences would not affect distance measures. DNA sequence space distance was defined as the number of positions with different nucleotides between the sequences (so maximum value is 10). Feature space distance was defined using the L_2 norm. While there is some correlation between the distance measures, it is clearly not enough to be considered predictive: that is, the DNA letter sequence cannot fully account for the physiochemical properties of DNA.

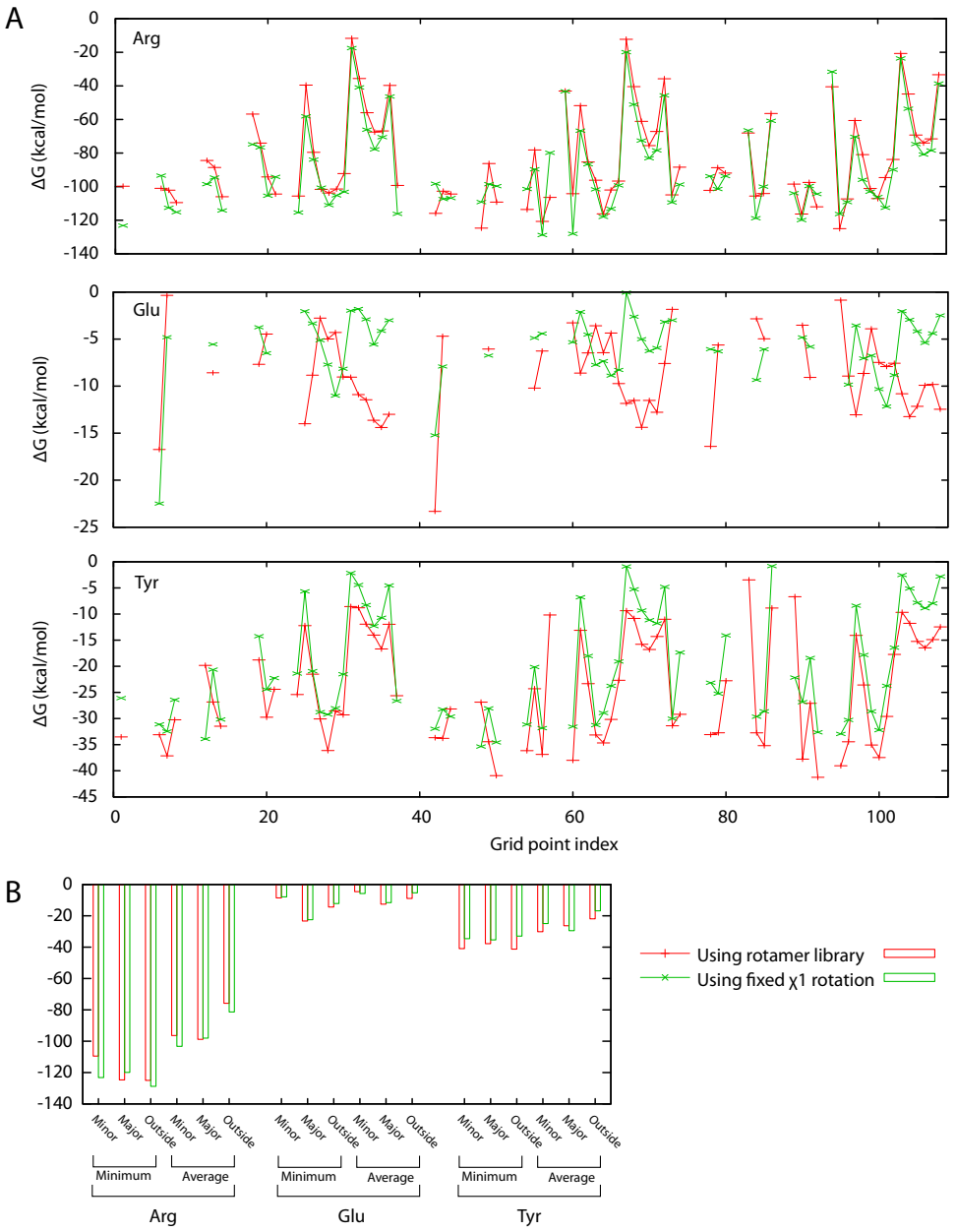

FIG. S10: Comparison of chemical feature calculations using different models for generating rotamers. In addition to performing the calculations as described in the Methods (fixed χ_1 rotation), we also used rotamers generated a rotamer library as described in the Discussion. (A) Free energies across the 108 grid points for arginine, glutamic acid, and tyrosine around the GGG DNA 3-mer. We only plot those points with $\Delta G < 0$. (B) Comparison of the six values (minimum and average ΔG over the minor groove, major groove, and outside DNA sub-grids) obtained for each amino acid. These values for the fixed χ_1 rotation were used along with similar values for other amino acids in PCA.

TABLE S1: Training results: F-measure and training time averaged over 5 training runs. "PTS" is positive training sequences.

	Average F-measure ± standard deviation						Average training time (minutes)					PTS
TF	BvH	SVM-LMM	SiteSleuth	SVM-PMM	SVM-PMM-FS	SVMR-PMM-FS	SVM-LMM	SiteSleuth	SVM-PMM	SVM-PMM-FS	SVMR-PMM-FS	
AgaR	0.224	0.347 ± 0.0871	0.470 ± 0.0513	0.491 ± 0.0513	0.533 ± 0.0000	0.512 ± 0.0419	2.8	30.3	71.2	52.5	1838.6	11
AraC	0.057	0.109 ± 0.0282	0.400 ± 0.0305	0.426 ± 0.0114	0.439 ± 0.0231	0.449 ± 0.0246	14.8	30.7	74.0	53.5	1320.5	20
ArcA	0.205	0.224 ± 0.0193	0.324 ± 0.0166	0.333 ± 0.0294	0.346 ± 0.0101	0.329 ± 0.0072	91.2	400.8	824.2	585.5	8299.9	91
ArgR	0.561	0.646 ± 0.0438	0.519 ± 0.0064	0.494 ± 0.0322	0.505 ± 0.0165	0.532 ± 0.0146	4.4	24.1	54.9	39.7	1738.4	24
CpxR	0.095	0.140 ± 0.0082	0.326 ± 0.0206	0.402 ± 0.0301	0.404 ± 0.0317	0.462 ± 0.0310	20.3	31.3	68.3	46.8	1311.2	33
CRP	0.715	0.719 ± 0.0060	0.701 ± 0.0191	0.711 ± 0.0154	0.722 ± 0.0228	0.752 ± 0.0048	151.7	150.2	274.9	171.2	2776.9	260
CysB	0.000	0.000 ± 0.0000	0.000 ± 0.0000	0.000 ± 0.0000	0.000 ± 0.0000	0.000 ± 0.0000	4.3	85.3	195.3	138.7	3344.9	8
CytR	0.189	0.439 ± 0.0366	0.600 ± 0.0000	0.600 ± 0.0000	0.600 ± 0.0000	0.600 ± 0.0000	4.7	91.9	195.6	148.4	5102.5	14
DeoR	0.033	0.114 ± 0.0628	0.289 ± 0.0778	0.444 ± 0.0000	0.444 ± 0.0000	0.444 ± 0.0000	2.5	17.2	30.4	25.9	925.0	7
DgsA	0.089	0.435 ± 0.0435	0.516 ± 0.0582	0.545 ± 0.0000	0.545 ± 0.0000	0.516 ± 0.0582	1.7	22.5	53.0	39.3	1615.2	8
DnaA	0.006	0.012 ± 0.0052	0.403 ± 0.0226	0.513 ± 0.0369	0.517 ± 0.0459	0.682 ± 0.0192	6.3	7.3	11.4	9.9	639.2	10
FadR	0.068	0.282 ± 0.0278	0.253 ± 0.0586	0.200 ± 0.0435	0.304 ± 0.0176	0.308 ± 0.0151	4.1	20.3	40.5	31.8	975.5	10
Fis	0.185	0.290 ± 0.0079	0.219 ± 0.0181	0.264 ± 0.0119	0.258 ± 0.0125	0.330 ± 0.0106	179.3	167.8	293.5	205.9	3181.3	133
FlhDC	0.024	0.061 ± 0.0084	0.176 ± 0.0236	0.153 ± 0.0190	0.194 ± 0.0292	0.141 ± 0.0076	21.0	28.0	56.9	42.7	973.7	20
FNR	0.412	0.459 ± 0.0101	0.583 ± 0.0118	0.599 ± 0.0231	0.607 ± 0.0076	0.617 ± 0.0097	33.5	43.7	84.2	54.7	1124.6	85
FruR	0.340	0.772 ± 0.0295	0.722 ± 0.0251	0.771 ± 0.0409	0.733 ± 0.0256	0.716 ± 0.0258	1.2	15.8	30.1	23.9	1388.9	13
Fur	0.295	0.376 ± 0.0165	0.463 ± 0.0176	0.454 ± 0.0384	0.453 ± 0.0294	0.555 ± 0.0158	32.0	40.4	84.1	60.6	1630.6	54
GadE	0.324	0.335 ± 0.0250	0.571 ± 0.0000	0.571 ± 0.0000	0.571 ± 0.0000	0.571 ± 0.0000	1.4	16.4	34.4	27.6	1190.7	5
GalR	0.528	0.555 ± 0.0205	0.750 ± 0.0000	0.761 ± 0.0136	0.824 ± 0.0000	0.794 ± 0.0360	2.0	14.5	29.5	23.2	934.3	10
GalS	0.267	0.480 ± 0.1000	0.716 ± 0.0315	0.658 ± 0.0167	0.675 ± 0.0157	0.714 ± 0.0000	1.8	14.9	28.6	22.8	952.1	9
GcvA	0.000	0.000 ± 0.0000	0.000 ± 0.0000	0.000 ± 0.0000	0.000 ± 0.0000	0.000 ± 0.0000	1.9	32.7	81.4	60.9	1907.1	5
GlpR	0.085	0.121 ± 0.0187	0.212 ± 0.0187	0.216 ± 0.0132	0.251 ± 0.0235	0.216 ± 0.0501	7.6	37.7	78.1	62.7	1975.2	23
GntR	0.318	0.399 ± 0.0701	0.662 ± 0.0184	0.640 ± 0.0000	0.656 ± 0.0210	0.650 ± 0.0209	4.6	31.3	66.8	50.3	1751.6	17
H-NS	0.033	0.042 ± 0.0057	0.034 ± 0.0064	0.068 ± 0.0139	0.059 ± 0.0084	0.047 ± 0.0131	28.1	45.7	85.4	59.3	1329.2	34
IclR	0.002	0.005 ± 0.0004	0.072 ± 0.0099	0.060 ± 0.0043	0.075 ± 0.0037	0.291 ± 0.0155	5.7	6.8	11.4	9.7	104.5	10
IHF	0.086	0.098 ± 0.0058	0.168 ± 0.0152	0.166 ± 0.0119	0.177 ± 0.0107	0.214 ± 0.0186	56.4	92.1	117.2	95.2	1191.6	87
IscR	0.000	0.443 ± 0.0457	0.222 ± 0.0000	0.222 ± 0.0000	0.222 ± 0.0000	0.178 ± 0.0889	2.1	29.2	70.1	54.7	1750.2	8
LexA	0.666	0.776 ± 0.0119	0.812 ± 0.0155	0.821 ± 0.0097	0.835 ± 0.0111	0.829 ± 0.0000	4.1	27.8	62.9	39.0	1833.6	24
Lrp	0.035	0.040 ± 0.0023	0.267 ± 0.0131	0.327 ± 0.0066	0.315 ± 0.0078	0.657 ± 0.0119	50.5	178.5	115.7	96.9	2135.1	84
MalT	0.071	0.079 ± 0.0047	0.529 ± 0.0284	0.566 ± 0.0332	0.598 ± 0.0298	0.606 ± 0.0313	5.2	9.5	18.4	14.1	479.9	20
MarA	0.019	0.036 ± 0.0368	0.097 ± 0.0110	0.000 ± 0.0000	0.078 ± 0.0391	0.000 ± 0.0000	12.7	38.4	92.7	67.9	1858.9	16
MelR	0.147	0.316 ± 0.0200	0.667 ± 0.0000	0.667 ± 0.0000	0.667 ± 0.0000	0.667 ± 0.0000	2.1	16.9	37.1	27.6	1279.9	6
MetJ	0.005	0.007 ± 0.0003	0.155 ± 0.0141	0.205 ± 0.0121	0.205 ± 0.0185	0.376 ± 0.0237	20.8	20.8	24.2	19.4	443.9	27
MetR	0.094	0.344 ± 0.0415	0.500 ± 0.0000	0.500 ± 0.0000	0.500 ± 0.0000	0.500 ± 0.0000	2.1	27.6	65.8	47.3	1983.4	6
ModE	0.117	0.340 ± 0.0711	0.618 ± 0.0594	0.545 ± 0.0000	0.642 ± 0.0485	0.545 ± 0.0000	2.6	37.0	83.8	63.2	1881.9	8
Nac	0.009	0.023 ± 0.0037	0.000 ± 0.0000	0.000 ± 0.0000	0.000 ± 0.0000	0.000 ± 0.0000	6.3	18.0	35.7	27.6	830.4	10
NagC	0.271	0.745 ± 0.0441	0.803 ± 0.0249	0.749 ± 0.0271	0.771 ± 0.0356	0.764 ± 0.0306	1.6	25.4	54.3	37.6	1996.2	14
NanR	0.019	0.019 ± 0.0000	0.256 ± 0.0000	0.447 ± 0.0165	0.418 ± 0.0222	0.800 ± 0.0000	1.9	2.3	4.0	4.3	2.7	6
NarL	0.023	0.027 ± 0.0000	0.257 ± 0.0230	0.275 ± 0.0087	0.266 ± 0.0075	0.383 ± 0.0072	18.5	523.0	87.4	86.4	232.4	91
NarP	0.008	0.009 ± 0.0004	0.163 ± 0.0162	0.181 ± 0.0247	0.176 ± 0.0177	0.385 ± 0.0137	7.4	6.1	13.3	9.7	198.1	16
NtrC	0.349	0.442 ± 0.0490	0.656 ± 0.0205	0.689 ± 0.0082	0.674 ± 0.0214	0.674 ± 0.0154	4.4	24.3	50.4	38.8	1138.3	22
OmpR	0.139	0.177 ± 0.0221	0.281 ± 0.0288	0.262 ± 0.0207	0.290 ± 0.0052	0.253 ± 0.0181	5.8	33.4	78.3	56.5	1418.2	20
OxyR	0.000	0.000 ± 0.0000	0.000 ± 0.0000	0.000 ± 0.0000	0.000 ± 0.0000	0.000 ± 0.0000	4.7	105.4	220.1	163.3	3297.1	9
PhoB	0.378	0.496 ± 0.0378	0.486 ± 0.0201	0.484 ± 0.0219	0.509 ± 0.0441	0.499 ± 0.0299	4.4	28.7	63.3	48.0	1720.3	14
PhoP	0.517	0.581 ± 0.0333	0.581 ± 0.0116	0.582 ± 0.0241	0.615 ± 0.0215	0.613 ± 0.0162	3.1	20.9	45.8	30.3	1068.7	22
PspF	0.000	0.000 ± 0.0000	0.000 ± 0.0000	0.000 ± 0.0000	0.000 ± 0.0000	0.000 ± 0.0000	1.4	15.2	28.7	23.5	874.9	5
PurR	0.152	0.273 ± 0.0421	0.510 ± 0.0323	0.498 ± 0.0239	0.500 ± 0.0000	0.567 ± 0.0130	7.3	22.1	47.3	36.7	1095.0	18
RcsAB	0.000	0.000 ± 0.0000	0.000 ± 0.0000	0.000 ± 0.0000	0.000 ± 0.0000	0.000 ± 0.0000	1.1	9.9	19.7	16.8	673.3	5
Rob	0.000	0.000 ± 0.0000	0.000 ± 0.0000	0.000 ± 0.0000	0.000 ± 0.0000	0.000 ± 0.0000	2.1	25.2	51.8	40.8	1319.4	6
SoxS	0.001	0.017 ± 0.0067	0.000 ± 0.0000	0.000 ± 0.0000	0.000 ± 0.0000	0.000 ± 0.0000	19.9	37.2	82.9	59.7	1326.4	18
TorR	0.044	0.083 ± 0.0440	0.598 ± 0.0428	0.769 ± 0.0000	0.733 ± 0.0724	0.769 ± 0.0000	2.7	6.5	13.1	10.0	635.6	8
TrpR	0.329	0.734 ± 0.0617	0.610 ± 0.0467	0.629 ± 0.0467	0.648 ± 0.0381	0.590 ± 0.0381	1.3	19.1	33.9	25.2	1140.2	10
TyrR	0.091	0.152 ± 0.0133	0.453 ± 0.0230	0.417 ± 0.0000	0.412 ± 0.0174	0.403 ± 0.0275	11.3	39.8	86.5	70.0	1730.6	19
UxuR	0.000	0.853 ± 0.0435	0.889 ± 0.0000	0.889 ± 0.0000	0.889 ± 0.0000	0.889 ± 0.0000	1.0	14.5	25.8	21.6	972.7	5
Average	0.160	0.268	0.381	0.394	0.405	0.433	17	53	85	63	1571	