iScience, Volume *27*

Supplemental information

Grammar rules and exceptions for the language of transcriptional activation domains

David G. Cooper, Tamara Y. Erkina, Bradley K. Broyles, Caleb A. Class, and Alexandre M. Erkine

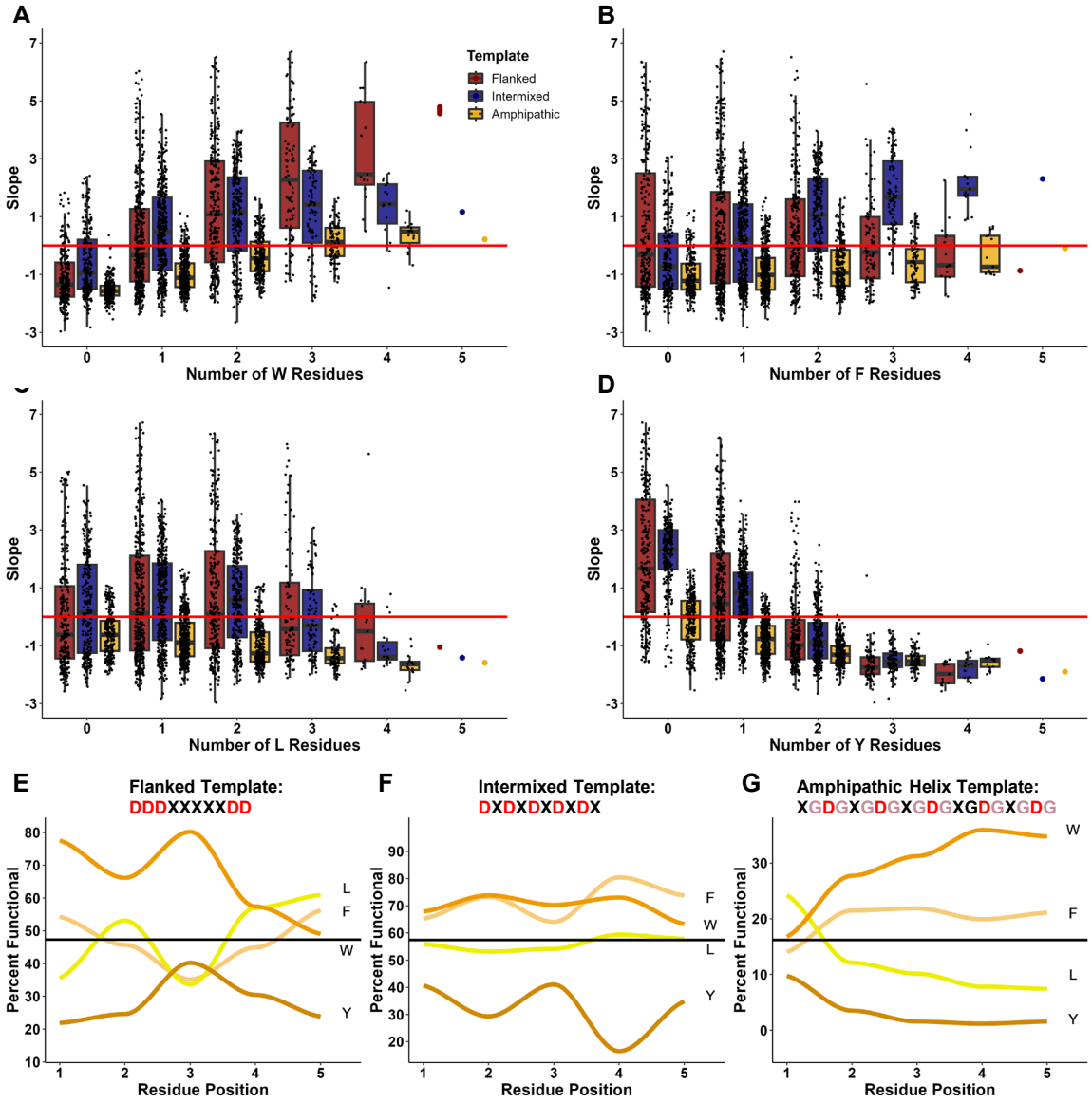

Figure S1. Compositional and positional analysis of functional AD sequences in context of flanked, intermixed, and amphipathic helix templates, related to Figure 1. A-D – Effect of number of hydrophobic residues: W (panel A), F (panel B), L (panel C), or Y (panel D) within three sequence templates: flanked template (DDDXXXXXDD, red), intermixed template (DXDXDXDXDX, blue), and amphipathic helix template (XGDGXGDGXGDGXGDGXGDG, yellow). X-axis: number of residues. Y-axis: growth slope of cells carrying the corresponding sequences. **E-G** – Effect of position of hydrophobic residues: W, F, L, or Y within three sequence templates: flanked template (panel E), intermixed template (panel F), and amphipathic helix template (panel G). X-axis: position of W, F, L, or Y residues within the 5 X positions of the template sequence. Y-axis: percent of cells carrying the corresponding sequences that have a growth slope above the functionality threshold. Horizontal lines correspond to the total percent functionality of each template dataset.

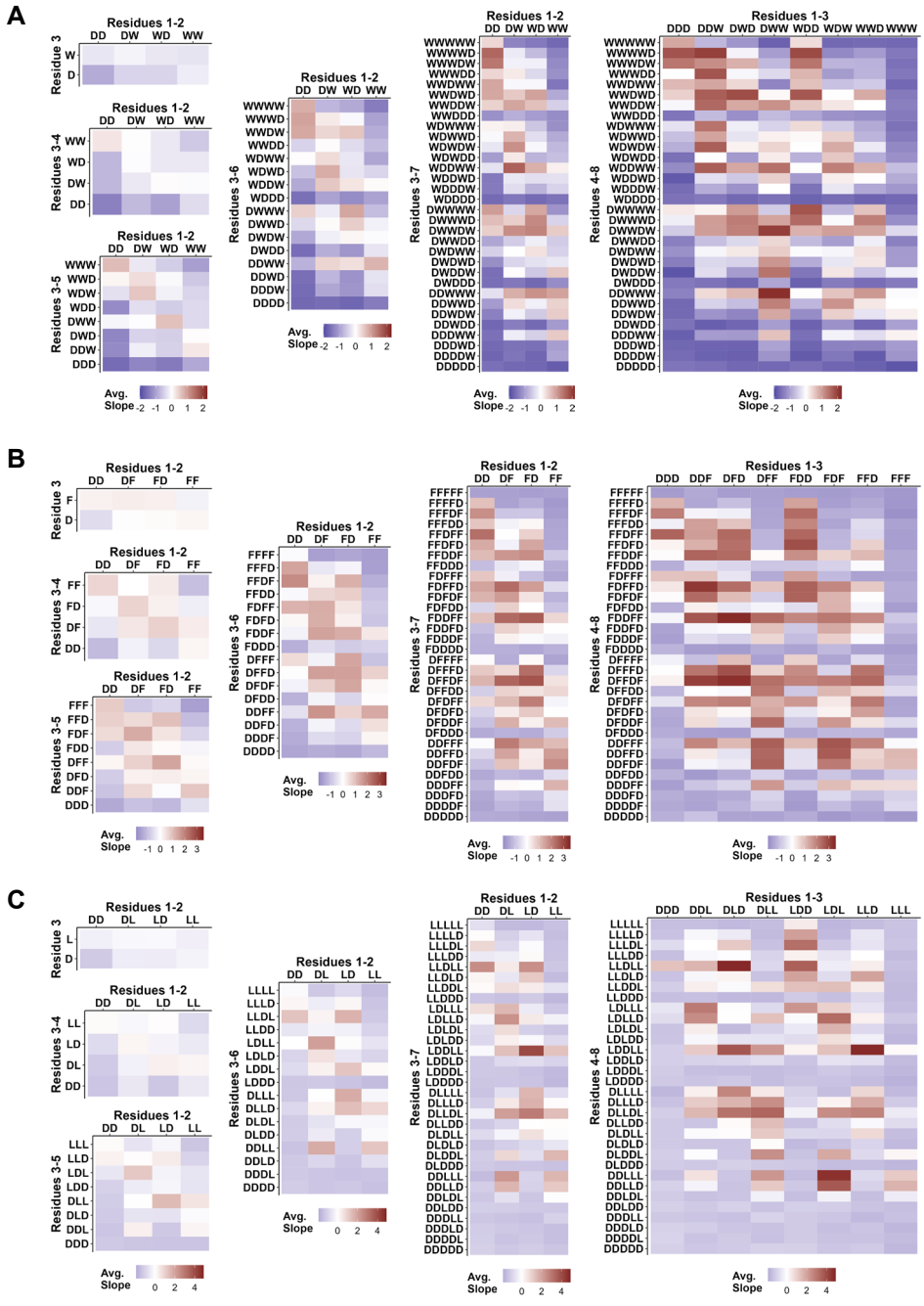

Figure S2. Search for SLiMs in WD10, FD10, and LD10 datasets identifies a spectrum of short sequences enriched for *in vivo* active ADs, related to Figure 4. A-C – AD activity heat maps for all possible 3 to 8-residue long sequences for WD10 (panel A), FD10 (panel B), and LD10 (panel C) datasets. Cell shading corresponds to average growth slope for all sequences containing the designated mini-motif short sequence.

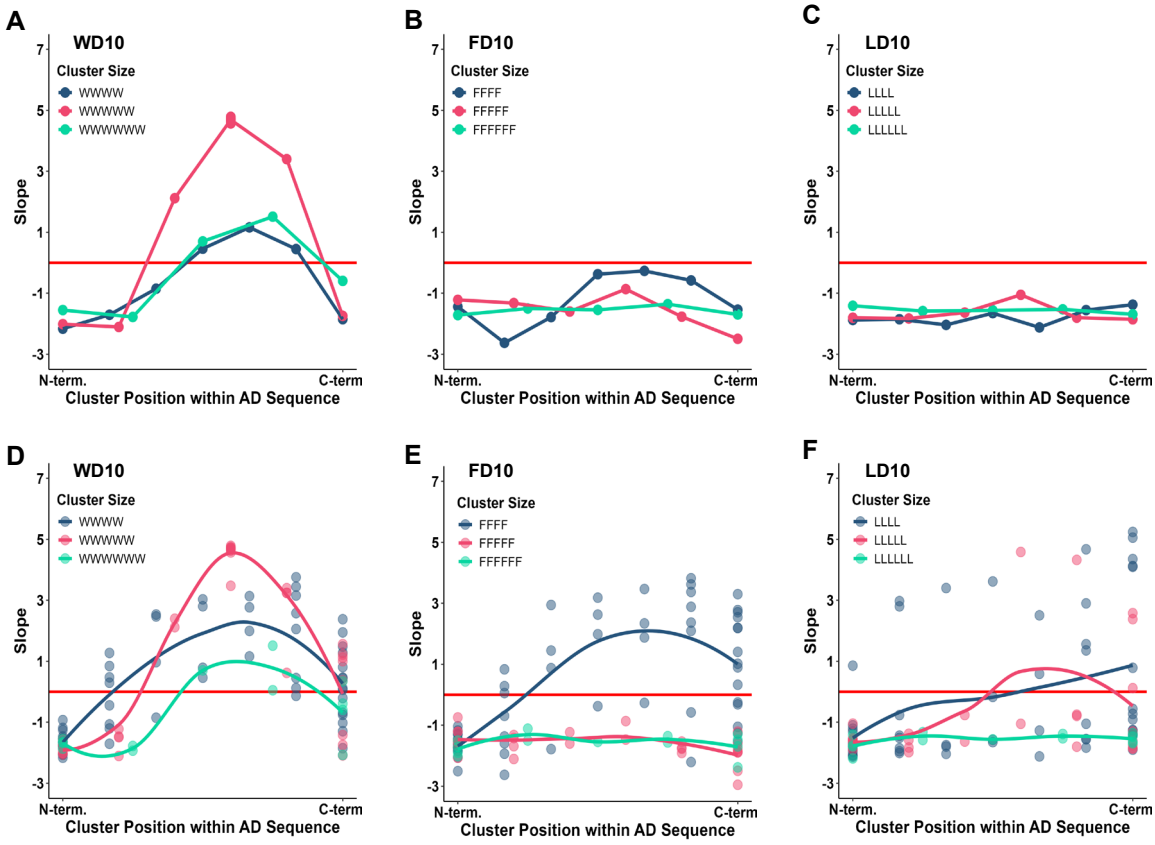

Figure S3. Activity of sequences containing large clusters of hydrophobic residues, related to Figure 4 and 5. A-C – Sequences with a single hydrophobic cluster flanked entirely by D residues. **D-F** – Sequences with a hydrophobic cluster [defined as a set number of hydrophobics flanked by 2 D residues, or flanked by 1 D residue directly before the edge, or present at either edge of the sequence] with the remaining residues being all possible combinations of hydrophobic and D residues. (extracted from the WD10, FD10, and LD10 sets).

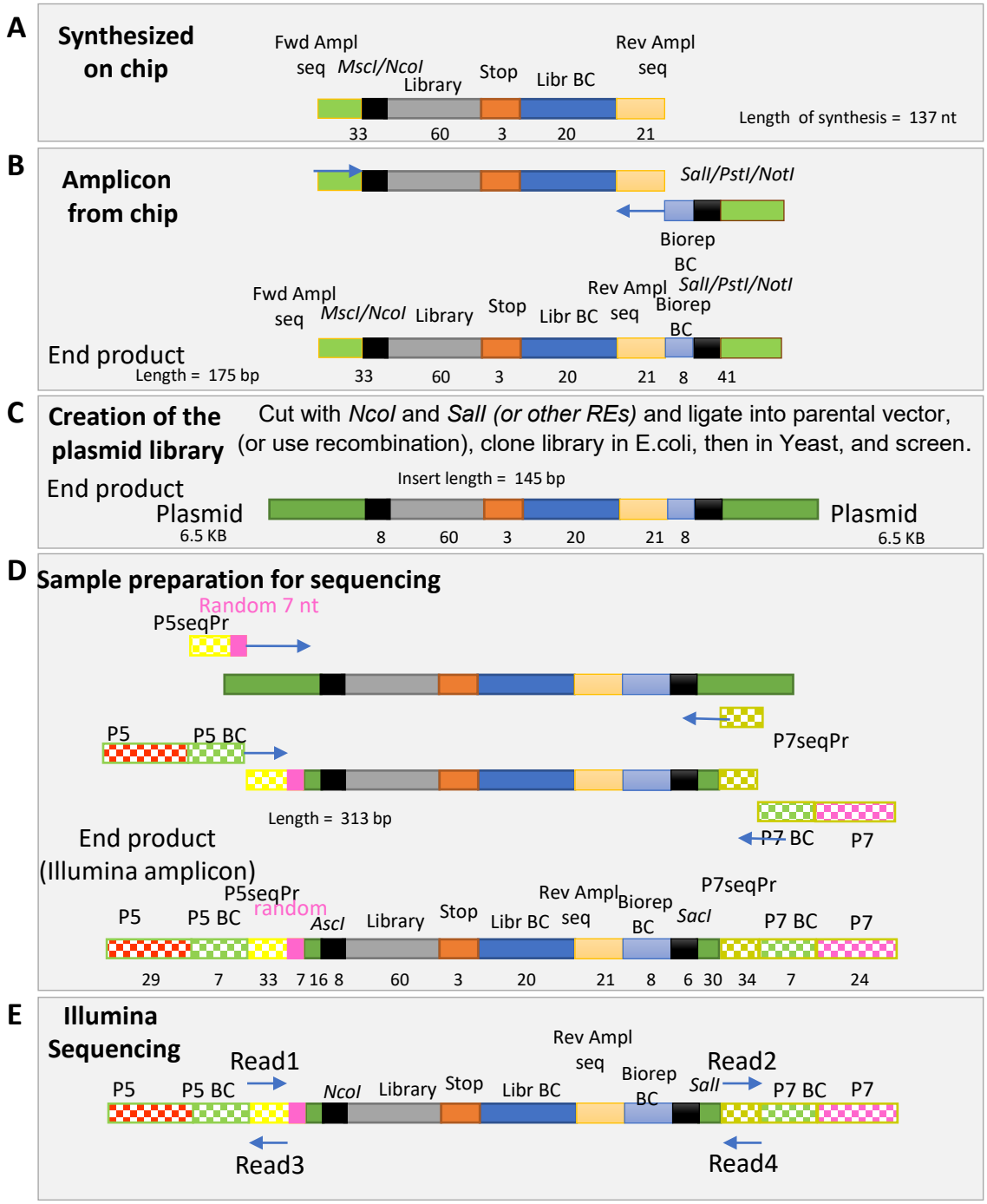

Figure S4. Schematic representation of wet lab steps for library sample preparations, related to Star Methods.: **A** – massive parallel synthesis of the design library; **B** – BioRep barcodes appending**; C** – cloning into parental yeast shuttle vector; **D** – sample preparation for NGS Illumina sequencing; **E** – sequencing at Illumina sequencing facility.

Group	Flanked		Intermixed		Amphipathic Helix		
	Percent Functional	Average Slope	Percent Functional	Average Slope	Percent Functional	Average Slope	Number in Group
W5L0F0Y0	100.0	4.69	100.0	1.17	100.0	0.22	1
W4L1F0Y0	100.0	4.35	100.0	1.81	100.0	0.55	5
W4L0F1Y0	100.0	3.29	100.0	2.03	100.0	0.66	5
W4L0F0Y1	100.0	2.25	60.0	-0.01	20.0	-0.18	5
W3L2F0Y0	90.0	3.80	100.0	2.14	60.0	0.29	10
W3L1F1Y0	95.0	3.44	95.0	2.31	95.0	0.64	20
W3L1F0Y1	90.0	2.59	75.0	0.36	40.0	-0.27	20
W3L0F2Y0	90.0	2.45	100.0	2.97	100.0	0.85	10
W3L0F1Y1	85.0	1.50	70.0	0.70	30.0	-0.08	20
W3L0F0Y2	50.0	0.11	0.0	-1.11	0.0	-0.76	10
W2L3F0Y0	80.0	2.45	100.0	1.91	40.0	-0.45	10
W2L2F1Y0	90.0	2.78	100.0	2.40	43.3	0.10	30
W2L2F0Y1	70.0	1.81	80.0	0.68	3.3	-0.73	30
W2L1F2Y0	83.3	2.49	100.0	2.89	83.3	0.45	30
W2L1F1Y1	80.0	1.72	85.0	1.08	25.0	-0.38	60
W2L1F0Y2	46.7	0.22	13.3	-0.86	0.0	-1.02	30
W2L0F3Y0	80.0	1.32	100.0	3.37	100.0	0.74	10
W2L0F2Y1	66.7	0.62	80.0	1.59	23.3	-0.15	30
W2L0F1Y2	30.0	-0.32	43.3	-0.60	0.0	-0.84	30
W2L0F0Y3	0.0	-1.42	0.0	-1.46	0.0	-1.20	10
W1L4F0Y0	60.0	1.18	60.0	-0.17	0.0	-1.48	5
W1L3F1Y0	65.0	1.42	90.0	1.37	5.0	-0.99	20
W1L3F0Y1	45.0	0.53	25.0	-0.44	0.0	-1.38	20
W1L2F2Y0	83.3	1.66	100.0	2.26	33.3	-0.44	30
W1L2F1Y1	58.3	0.89	83.3	0.66	0.0	-1.09	60
W1L2F0Y2	26.7	-0.36	3.3	-1.04	0.0	-1.52	30
W1L1F3Y0	80.0	1.29	100.0	2.77	40.0	-0.02	20
W1L1F2Y1	60.0	0.74	93.2	1.43	3.3	-0.75	60
W1L1F1Y2	31.7	-0.33	26.7	-0.44	1.7	-1.26	60
W1L1F0Y3	5.0	-1.35	0.0	-1.52	0.0	-1.55	20
W1L0F4Y0	60.0	0.23	100.0	3.20	100.0	0.47	5
W1L0F3Y1	40.0	-0.22	100.0	1.94	10.0	-0.32	20
W1L0F2Y2	6.7	-0.90	50.0	-0.06	0.0	-1.02	30
W1L0F1Y3	0.0	-1.68	0.0	-1.28	0.0	-1.39	20
W1L0F0Y4	0.0	-1.98	0.0	-1.88	0.0	-1.55	5
W0L5F0Y0	0.0	-1.05	0.0	-1.42	0.0	-1.59	1
W0L4F1Y0	20.0	-0.66	0.0	-1.26	0.0	-1.91	5
W0L4F0Y1	20.0	-1.15	0.0	-1.54	0.0	-1.63	5
W0L3F2Y0	30.0	-0.20	70.0	0.21	0.0	-1.44	10
W0L3F1Y1	10.0	-0.84	0.0	-1.15	0.0	-1.66	20
W0L3F0Y2	0.0	-1.47	0.0	-1.68	0.0	-1.52	10
W0L2F3Y0	40.0	-0.03	100.0	1.36	0.0	-1.33	10
W0L2F2Y1	26.7	-0.61	53.3	0.06	0.0	-1.58	30
W0L2F1Y2	10.0	-1.29	0.0	-1.41	0.0	-1.67	30
W0L2F0Y3	0.0	-1.77	0.0	-1.74	0.0	-1.67	10
W0L1F4Y0	40.0	-0.17	100.0	2.02	0.0	-0.77	5
W0L1F3Y1	25.0	-0.56	95.0	1.02	0.0	-1.26	20
W0L1F2Y2	6.7	-1.20	13.3	-0.68	0.0	-1.59	30
W0L1F1Y3	0.0	-1.86	0.0	-1.69	0.0	-1.68	20
W0L1F0Y4	0.0	-1.87	0.0	-1.67	0.0	-1.60	5
W0L0F5Y0	0.0	-0.87	100.0	2.31	0.0	-0.10	1
W0L0F4Y1	20.0	-1.02	100.0	1.36	20.0	-0.71	5
W0L0F3Y2	0.0	-1.60	40.0	-0.11	0.0	-1.38	10
W0L0F2Y3	0.0	-1.88	0.0	-1.27	0.0	-1.52	10
W0L0F1Y4	0.0	-2.05	0.0	-1.74	0.0	-1.57	5
W0L0F0Y5	0.0	-1.19	0.0	-2.14	0.0	-1.90	1

Table S1. Diverse sequence compositions produce functional ADs across three templates, related to Figure 2. Templates – Flanked (DDDXXXXXDD), Intermixed (DXDXDXDXDX), and Amphipathic Helix (XGDGXGDGXGDGXGDGXGDG). **Group** – Description of sequence composition with numbers of each residue (W, L, F, and Y). **Number in Group** – Number of unique sequences with a given sequence composition. **Percent Functional** – Percent of cells carrying the corresponding sequences that have a growth slope above the functionality threshold. Cells shaded with a gradient from 100% (green) to 0% (red). **Average Slope** – Average growth slope across cells carrying the corresponding sequences. Cells shaded with a gradient from high positive slope (blue) to low negative slope (red).

	Number of hydrophobics	Deviation from Balance	Average position of hydrophobics	Clustering (Patterning Parameter)
WD10	10.67	-9.67	11.18	-4.37
FD10	11.48	-18.19	16.13	-16.42
LD10	17.76	-8.93	7.87	-8.16

Table S2. Logistic regression coefficient estimates for grammar rules on WD10, FD10, and LD10 sequences, related to Figure 7. Logistic regression models were computed for each set (WD10, FD10, LD10) to confirm the effect of the rules on functionality of ADs. Input values for each rule were calculated for each sequence and scaled as necessary to values between 0 and 1 (see methods, Logistic Regression). Output values represent average change in the log odds of a sequence being functional per unit increase in each rule value. All rules contribute to the functionality prediction for all three sets ($p<0.001$). Negative values for balance and clustering represent that sequences with larger values for these rules are less likely to be functional.