Surfactant mechanism of gene activation by transcriptional activation domains of sequence-specific factors

Bradley K. Broyles[1], Tamara Y. Erkina[1], Theodore P. Maris[1], Andrew T. Gutierrez[1], Daniel A. Coil[1], Thomas M. Wagner[1], Caleb A. Class[1], Alexandre M. Erkine[1]**

[1] College of Pharmacy and Health Sciences, Butler University,

Indianapolis, IN 46208, USA

**Corresponding Author and Lead Contact:

Tel: 317-940-8569; Fax: 317-940-6172; Email: aerkine@butler.edu

Summary

Gene expression in all eukaryotes depends critically on the function of transcriptional activation domains of gene activator proteins. The conventional model for activation domain (AD) function is the direct physical recruitment of specific coactivators and transcriptional machinery components. However, ADs are short and astronomically variable sequences, with up to 10^{24} possible interchangeable sequence variants for a single gene activator; each variant is intrinsically disordered in structure and interacts with its targets with low specificity and affinity. How these peptides recruit their targets is becoming increasingly difficult to explain, exposing a massive knowledge gap in molecular biology. Here, we show that the single required characteristic of ADs—consistent with their extreme variability, intrinsic structural disorder, and near-stochastic interaction mode—is an amphiphilic aromatic–acidic surfactant-like property. We propose that the AD surfactant, by triggering the local gene-promoter chromatin phase transition, catalyzes the formation of “transcription factory” condensates. We demonstrate that the presence of tryptophan and aspartic acid residues in the AD sequence is sufficient for *in vivo* functionality, even when present only as a single pair of residues within a 20-amino-acid sequence containing nothing more than additional 18 glycine residues. We demonstrate that the amphipathic α-helix structure, suggested previously as beneficial for AD function, is actually detrimental, and breaking this helix by inserting prolines significantly increases activation domain functionality. The proposed surfactant action mechanism based on near-stochastic interactions implied by the minimalistic activation domains changes not only the paradigm for the explanation of gene activation but also the fundamental biochemistry paradigm based on the specificity of sequence-to-structure-to-functional-interaction. The mechanism of activity regulation by near-stochastic allosteric interactions could easily be applied to other biological processes.

Introduction

The expression of genetic information written in nucleotide sequences of DNA is the fundamental function of all living entities. The process is initiated by the binding of gene activators to corresponding gene promoters. In eukaryotes, gene activator proteins typically contain two obligatory parts: the DNA-binding domain (DBD), which is responsible for recognition and specific binding to the cognate promoter DNA sequence, and the activation domain (AD), which is responsible for the initiation of gene transcription. The two domains are drastically different in nature and in function. The DBD for each activator contains a certain conserved amino acid sequence and structure and binds to a specific consensus DNA sequence [1,2]. Contrarily, ADs are typically short, ranging from six to several dozen amino acids, and extremely variable in sequence and it has been estimated that up to 10^{24} sequences are able to substitute for each other within the context of the same activator molecule [3-7]. In addition, ADs are intrinsically disordered in structure and bind to a wide variety of proteins and binding regions on the surface of these targets via “fuzzy”, low-specificity interactions [8-10]. The mechanism of the function of ADs is a long-standing enigma, as their features defy the traditional specificity of sequence-to-structure-to-function paradigm of molecular biology.

Two models have been proposed to explain the mechanism of AD function. First is the traditional and widely accepted model of direct physical recruitment by ADs of coactivators and components of enzymatic transcriptional machinery. This model implies a certain level of specificity for AD sequences, such as a consensus sequence, or at least Short Linear Motifs (SLiMs); some specific structural features, such as an amphipathic α-helix; and specific targets, for instance, the Mediator complex. However, recent high-throughput experimental data and the results of bioinformatics analysis increasingly contradict the recruitment model, showing the lack of specificity of ADs at the sequence, structure, and target interaction levels (for review see [co-submitted manuscript]). The second, newer and not yet widely accepted model [7,11] considers

ADs as surfactants that act on promoter nucleosomes, triggering chromatin phase transition, freeing promoter DNA, and promoting condensation of necessary enzymes and co-factors. The surfactant model postulates the extremely high variability of AD sequences, because the requirement for the functional sequence is the presence of a limited number of acidic and aromatic amino acids. The model requires no specific structure but spatial and structural flexibility, and presumes near-stochastic interactions with targets at the anchoring site (i.e., the gene promoter). The absence of a specific sequence or structure, and a near-stochastic target interaction mechanism, of a protein domain that is critical for a fundamental biological function presents a new paradigm, challenging the traditional foundational biochemical principle of specificity rooted in the sequence-to-structure-to-functional-interaction triad.

Here, we provide experimental data that allow us to discriminate between these two proposed mechanisms. By synthesizing and testing large sets of specific AD sequences *in vivo*, we demonstrate that, in line with the surfactant model and contrary to the specificity required by the recruitment model, the presence of only one tryptophan and aspartic acid, each represented as a single side chain-bearing amino acid residue within the AD region, is sufficient for gene activation. Also contrary to the expectations of the recruitment model, we demonstrate that the presence of the amphipathic α-helix structure is detrimental for AD functionality, while breaking this structure by insertion of prolines significantly increases AD functionality, providing for the first time a rationale for the existence of the entire class of proline-rich ADs. With these and other data, and in line with the surfactant model, we hope to change the gene expression paradigm and establish near-stochastic interactions as required for, rather than detrimental to, this and possibly other crucial biological functions.

Results

To elucidate the mechanism of AD function, we developed a high-throughput assay to test the *in vivo* functionality of individual sequences within a library of 12,400 synthetically

generated sequences. These sequences were designed to address multiple questions regarding potential determinants of AD function, such as length, composition, structure preference and others. The synthesized library was then cloned into a yeast centromeric shuttle vector, fusing each AD sequence to the Gal4 DBD. Then, the library was transformed into the yeast two-hybrid Y2HGold tester strain, and the transformed yeast were screened for growth, which was dependent on the expression of the *Gal*-Aureobasidin antibiotic resistance reporter gene (Fig. 1A). The library DNA for different growth time points was isolated and sequenced to determine the number of reads for each individual sequence and its change over time. The growth slope for each sequence was calculated and served as a measure of the AD functionality. Since the results were obtained within the scope of the whole library pool screening, the results for individual sequences and for all different sequence sets can be considered to be obtained under identical experimental conditions and thus to be accurately comparable. In addition, each sequence was labeled by multiple individual barcodes; thus, the results for each individual sequence are the mean of multiple (typically five) independent experimental repeats. As part of the library we used the sequences of known AD regions, such as Gal4(840-857), Gal4(860-872) and VP16 sequences [12,13], as internal positive controls (Fig. 1B). As negative controls, we used a null sequence that contained a stop codon after the DBD and a sequence containing a stretch of 20 glycines (G), which was shown to be neutral for AD functionality under similar experimental conditions [7]. To ensure high stringency, the cutoff for a "functional AD" was defined as the mean of the five highest values produced by 50 independent stop codon-null sequences.

A single aspartic acid and single tryptophan residue within the AD region are sufficient for functionality

Previous reports indicated that the most beneficial residues for AD functionality are aromatic and acidic amino acids, and the highest gain in prediction of AD functionality using machine learning on large >10^6 set of diverse random sequences was derived from the presence of amino acids W and D [7]. In addition, high AD functionality was repeatedly demonstrated for the monotonous WDWDWDWDWDWDWDWDWDWD sequence[5], indicating that for high functionality, it is sufficient to have only W and D. Here, we analyzed sequences with different numbers of WD repeats to determine how functionality depends on this feature and found that decreases in WD repeat number generally correlate with decreases in functionality (Fig. 1B). Unexpectedly, the sequence GGGGGGGGGGGGGGGGGGDW, containing just one W and one D, showed residual functionality (1.11±0.02 [95% CI]) comparable to that of the VP16 minimal AD module (0.55±0.28). Similarly, analysis of WD sequences surrounded by G residues or WD blocks separated by a stretch of G residues (Fig. 1C) revealed functionality for sequences containing a single W and a single D. The lowest functionality was demonstrated for sequences containing four or five Ws and Ds, and the highest functionality was demonstrated for the DWx9 sequence (2.12±0.22). The answer to the same question about the amount of Ws and Ds sufficient for AD functionality within the context of sequences with non-alternating clustered Ws and Ds (Fig. 1D) is that one W and one D is sufficient for at least a low level of functionality. Sequences with a single W and single D separated by different numbers of Gs (e.g., WGD, WGGD, WGGGD, see Fig. 1E and F) are also functional. By contrast, sequences containing more than two adjacent Ws or Ds (e.g., GGGGGGGGGGGGGGGGWWWDDD or GGGGGGGGGGGGGGGGDDDWWW) are nonfunctional, and the sequence functionality generally drops with increasing lengths of homo-W and homo-D stretches (Fig. 1D). These results are in good agreement with previous observations that clustering of separate acidic and separate aromatic residues within the AD region is detrimental to functionality [7]. Interestingly, some functionality was observed for sequences containing both W homostretches and D homostretches separated from each other

by larger numbers of Gs (Fig. 1E, green dots). In examining why the sequence containing the WDWDWDWD block is nonfunctional, we found that DWWDDWDW, a different sequence with an identical composition, is functional (Fig. 1G). A possible explanation for this observation is that while forming a 3D α-helix, Ws within the WDWDWDWD sequence are forming pi-pi interacting pairs, similar to those in a tryptophan zipper [14], while in DWWDDWDW, at least one W remains free. That in turn suggests that solvent exposure of at least one aromatic residue in the AD sequence is required for functionality and is consistent with the finding that a single D and single W is sufficient for functionality (Fig. 1F).

Balance and intermixing of acidic and aromatic residues underlies AD function

To expand the analysis, we turned to the other part of our library, which contains a set of sequences with all possible combinations of W and D at 12 positions (3968 quantified out of 4096 possible sequences). The *in vivo* screening revealed 1330 functional and 2638 nonfunctional sequences within this set. Analyzing this set as a whole, we found that in general, two features are important for functionality: the balance between W to D residues in the sequence, and how they are intermixed (Figs. 2A and 2B). Similar patterns were noticed when the balance and intermixing between aromatic and acidic residues were analyzed within a much larger set of ADs in the context of the Gcn4 activator or HSF activator (see reference[7] and Fig. 2C).

To examine whether these functional sequences might contain a specific mini-motif, we tested all 16 WD tetrapeptide variants and found that tetrapeptide sequences containing an excess of Ds, especially Ds clustered together, are generally detrimental, while D and W intermixing is beneficial and creates a number of functional tetrapeptide sequences, with DDWW as the top performing sequence (Fig. 2D). A very similar trend of aromatic-acidic tetrapeptide motif distribution is obvious from our analysis of two independent *in vivo*-tested AD and previously published sequence datasets created on the basis of natural AD sequences in

the context of artificial DBD-ER fusions [9] and a sampling of large unbiased random sequence ADs in the context of Gcn4 [6]. This similarity of the trends suggests an activator-independent general mechanism for AD function.

To test whether the position of the tetrapeptides within the AD sequence is important, we calculated the probability of each tetrapeptide contributing to functionality when positioned in different parts of the AD region (Fig. 2E). This analysis indicated that while W and D intermixing is beneficial, W-rich sequences are generally more beneficial at the spatially freer end of the molecule, while D-rich sequences are beneficial internally. Similar trends are observed for the broader set of acidic-aromatic tetrapeptides in the Gcn4 context (Fig. 2F and reference [7]). When we used individual tetrapeptides, their position within the sequence, and the balance between acidic and aromatic residues as features for regression ML model training, we observed that each feature had a positive value for the prediction of AD functionality, and the combination of all these features produces the most accurate prediction (Fig. 2G).

The position effect is much clearer when the functionality contributions of W and D within the AD sequence space are analyzed directly (Fig. 3A). Generally, the positive contribution of W increases when it is positioned toward the end of the molecule, while D shows the exactly opposite behavior. This observation is consistent with our previous analysis [7]; however, the trend breaks at the last two positions within a 12-amino-acid AD. In examining what sequences display higher functionality with generally detrimental terminal D(s), we found that it is especially beneficial if a cluster of Ws precedes the D(s) (Fig. 3B). A similar trend was observed when we analyzed the acidic and aromatic amino acid distributions within a previously published [6] dataset of random AD sequences in the Gcn4 context (Fig. 3C). Comparing the different sequences containing a cluster of 5 Ws, which is usually detrimental for functionality, we found that flanking such clusters with Ds is generally beneficial, with the highest functionality observed if the majority of Ds are situated internally (Fig. 3D). Molecular modeling suggested that the D-

flanking effect likely occurs because the repelling charges of aspartic acid residues prevent the tryptophan moieties from forming a hydrophobic mini-globule or a disordered aggregate supported by hydrophobic and pi-pi interactions between aromatic rings (Fig. 3E).

Formation of an amphipathic α-helix is not beneficial for AD functionality, while breaking the helix with proline increases the gene activation potential

Since the 3D structure within the AD microenvironment (Fig. 1-3) seems to play an important role, and because the amphipathic α-helix is specifically considered as an important structural feature of ADs [9,10], we created a library of sequences all containing 5 Ws and 5 Ds interspersed with random amino acids (W.D.W.D.W.D.W.D.W.D, henceforth called the WD5 library). The WD5 backbone sequence, if folded, always creates an amphipathic α-helix (Fig. 4A). WD5 library screening followed by DNA sequencing, normalization of the number of reads for each sequence to that at the 0 time point, and AD functionality cutoff based on redundant stop codon null sequences, as described above, confirmed that of 107,975 distinct sequences tested, 19.6% were functional ADs. We binned the sequences by their predicted percent α-helix formation and found that the percentage of functional sequences in each bin decreased as the prediction of the α-helix fraction increased (Fig. 4B). By analyzing sets of sequences enriched with individual amino acids within the WD5 library (Fig. 4C), we found that increasing the number of basic amino acids between Ws and Ds was detrimental, which was consistent with the highly negative effect of K and R on AD functionality [7]. A similar negative effect was observed for sequences enriched with additional (> 5) aromatic residues and to a lesser degree among sequences enriched with additional acidic residues, which is consistent with the results of Fig. 2A and the previously observed negative effect of shifting the balance between acidic and aromatic residues [7]. Unexpectedly, a progressive increase in the number of proline residues (P) within the WD5 sequences was correlated with a significant increase in the probability of functionality, from 15.7% with zero prolines to 55.6% with five prolines (Fig. 3C). Proline

residues are known to be potent helix breakers. Thus, breaking the amphipathic α-helix of the WD5 sequence is generally beneficial for functionality of the sequence within the WD5 context. This proline effect was not observed for sequences from random-sequence libraries [5,6], suggesting that the positive proline effect may be specific to the amphipathic helix context of the WD5 library. The presence of prolines in this case likely prevents tryptophan rings from interacting with each other on one side of the amphipathic helix, thus keeping the Ws exposed to the solvent (Fig. 4D). Fig. 4E confirms the phenotypes of the key sequences discussed above.

Discussion

Several key results of our study help to discriminate between the recruitment and surfactant models for the AD mechanism of function. An important observation in our study is that a variety of sequences containing a single W and a single D, interspersed with glycine residues, which lack a side chain, are able to serve as functional ADs when fused to the Gal4 DBD and activate the reporter gene expression *in vivo.* This indicates that interactions of ADs with targets have extremely low affinity and specificity, at the level of a single salt bridge and a single amino acid hydrophobic contact. This level of interactions is easily compatible with the action of a surfactant triggering the local promoter-chromatin phase transition, while for the coactivator recruitment and selection by AD among variety of possible targets, a significantly higher level of specificity for AD sequence is necessary. This conflict with the specificity concept, used by the conventional sequence-to-structure-to-function mentality, was noticed previously [6,9,15], and is reflected in invoking of a consensus sequence for ADs, or at least short linear motifs (SLiMs). However, the consensus sequence and SLiMs upon inspection and machine learning (ML) analysis of large AD datasets are demonstrated to be absent or not contributing to ML predictions for a sequence to be functional AD [6,7]. In contrast, the surfactant model does not require high level of specificity and can even explain the otherwise puzzling

early reports of nonnatural acidic-aromatic chemical compounds and even RNA fragments acting as ADs in the context of the Gal4 DBD [16,17]. It is also worth noting that in our study, although we analyzed designed AD sequences created by the cutting-edge massive parallel synthesis method, and thus the sequences could be characterized as “synthetic” or “artificial”, all of the sequences were verified *in vivo* and thus represent *bona fide* ADs. In addition, naturally occurring AD sequences such as VP16 and Gal4 AD modules were included in the library as internal controls.

The conventional recruitment model favors the idea that the amphipathic α-helix is a structural element that is involved in the recruitment of coactivators and transcriptional machinery by the AD [9,10]. The amphipathic α-helix in this case fits the valley or even a tunnel of the AD binding site on the coactivator surface, thus ensuring multiple bonds necessary for the recruitment event [9]. However, the results of our analysis of thousands of sequences suggest that the presence of the amphipathic α-helix in AD is, if not detrimental, then at least not beneficial for function, and breaking this structure by prolines increases the probability of AD functionality proportionally to the increase in proline content (Fig. 4C). Contrary to the expectation based on the recruitment model concept, breaking the structure and thus making Ws and Ds more solvent-exposed and more available for interaction with target(s) is functionally beneficial. Following the same logic, adding a surplus of Ws in a WD5 sequence increases the likelihood of pi-pi interactions between neighboring Ws, thus leading to the formation of a locked noncanonical structure or a structure similar to that of a tryptophan zipper [14], which might contribute to the decrease in functionality observed for sequences containing 4 WDs (Fig. 1G). The gain of functionality for sequences with identical composition correlates with the ability to adopt a structure that ensures an individual tryptophan maintaining a solvent exposed configuration.

The positive effect of proline, demonstrated for 830 sequences with ≥3 prolines within the WD5 amphipathic α-helix context, suggests the explanation for existence of the entire proline-rich class of ADs. This class was described several decades ago [8,18], but the reason for the functional preference of proline in ADs has remained obscure. Considering the surfactant model, proline residues simply ensure the exposure of key residues, such as W and D, for interactions with the target. Following the same logic, 2579 functional AD sequences out of 8094 sequences with ≥2W and ≥2D in our design library maintain the functionality of the domain through the internal repulsion of similarly charged Ds, which disrupts conventional structures or non-conventional aggregation of Ws, maintaining aromatic residues in the solvent exposed configuration.

Whether the hydrophobic residues in AD are aliphatic or aromatic should be irrelevant for the recruitment mechanism, but this is incompatible with the previously published results of ML analyses[7] , which suggest that aromatic residues are beneficial, while aliphatic residues, although also highly hydrophobic (i.e. I, V), are not [7]. Our study shows a variety of highly functional sequences containing only W as a hydrophobic residue, which is consistent with previous findings of the ML analysis [7]. Although the exclusive role of aromatic residues is not fully compatible with the recruitment model, the surfactant model proposes the initial step of AD function as “fuzzy” interaction with DNA via intercalation [7,11], which requires aromatic residues to be exclusively important.

Another argument in favor of the surfactant model is the general preference for aromatic residues at the spatially free terminus of the molecule observed in this study (Fig. 3A) and previously [6,7]. While for the recruitment model, initial interactions with the target are based on scanning by the negatively charged acidic residue and establishing a strong initial salt bridge with the target [19], for the surfactant model the negative charge at the end is disadvantageous for function due to the repulsion of DNA phosphates and hence the interference with the required

initial intercalations of aromatic residues into DNA. Consistent with the surfactant model, we demonstrate that the exception from the aromatic end preference is observed only when a terminal acidic residue(s) is(are) required for the unraveling of aromatic clusters (Fig. 3).

While the role of acidic residues in exposing aromatics is important, the main function of acidic moieties in the surfactant model is interference in DNA-histone nucleosome salt bridges. This action of the amphiphilic AD triggers promoter chromatin remodeling, freeing the promoter DNA [11]. The attraction of multiple transcription machinery components to the exposed promoter DNA may explain the liquid–liquid phase separation (LLPS) observed in the eukaryotic nucleus upon induction of transcription [15].

The deficiency of the recruitment model in explaining the coactivator recruitment by the AD sequences, which have almost no sequence-structure specificity, has been noted many times [6,9,11], and an attempt to resolve this contradiction was recently made by invoking the LLPS model [15]. The transient condensates are proposed to bring together the potential interacting partners, such as gene activators, coactivators, and transcriptional machinery, which otherwise are unlikely to interact due to “fuzziness” and the overwhelming variability of the interactions. However, the initial trigger of condensation is supposed to be action of ADs. With the extremely high variability of AD sequences, absence of a specific structure, and lack of target selectivity, it is unclear how ADs could initiate this process by recruiting specific other factors. The surfactant model restores the logic for LLPS, suggesting that the initial trigger is the local chromatin phase transition, leading to the exposure of gene promoter DNA, which attract the general transcriptional machinery, including the Mediator complex, and thereby promotes transcriptional factory condensation.

While the surfactant model allows us to look at the most important function in biology – gene expression – from a different perspective, it is not the only biological function with an unexplained mechanism that requires close attention. Near-stochastic interactions and

intrinsically disordered protein regions have been shown to play important roles in such processes as mRNA processing, apoptosis, molecular transport within and between cells, glycolysis, and many others [20,21]. Breaking from the specific sequence-to-structure-to-function paradigm and considering near-stochastic interactions as fundamentally important and not detrimental opens the direction to the completely new branch of biochemistry and molecular biology [11].

Methods

Library construction, cloning, and screening

The parental library plasmid was constructed by cloning the fragment containing the *ADH1* promoter and the Gal4(1-147) DBD cassette, PCR amplified from the commercially available pGBKT7 vector. The PCR fragment was cloned into the *SacI* and *KpnI* sites of the centromeric yeast shuttle vector pRS314.

The design library containing 11,500 individual sequences, each with an individual 20-nucleotide barcode directly following the stop codon to improve alignment performance, was synthesized at the GenScript commercial facility, amplified by PCR five times (each time appending a unique BioRep barcode), quantitated for the DNA content, and mixed in equal proportions into a single pool. For description of more detailed steps, see supplementary Fig.1. The pool was cloned into the *NcoI* and *SalI* restriction sites remaining from the pGBKT7 fragment of the parental library plasmid. The library complexity was estimated by individual colony counts after transformation for a fraction of the total transformation mix, then multiplying by the fraction factor. Total complexity was estimated to be ~10^6. The total content of individual sequences within the library was determined by NGS at GenScript. The NGS sequencing also confirmed the in-frame fusion of AD sequences to the Gal4 DBD region. After the bacterial cloning and verification, the plasmid library was isolated for the following yeast transformation.

At the Butler University research lab, the isolated plasmid library was transformed into the yeast strain Y2HGold, available commercially from Clontech/Takara. The maintenance of the library complexity was determined by the individual colony count for a fraction of a transformation mix as described above for the bacterial transformation. The number of individual yeast transformants for entire library was estimated to be ~10^6. After transformation, the whole-library cell culture was transferred into the –trp synthetic yeast growth medium containing 200 µg/ml of aureobasidin and grown for four days with daily 1/100 dilution to maintain the culture in the mid-log phase. Cell culture samples were taken at 0, 1, 2, 3, and 4 days. DNA was isolated using a Thermo Scientific Pierce Yeast DNA Extraction Reagent Kit. The library component was

isolated by PCR using the Invitrogen AccuPrime SuperMix I kit with primers containing Illumina adapters and barcodes unique for each DNA sample. DNA samples were controlled for purity, repeatedly quantitated for DNA content, and sequenced at the NovoGene commercial facility.

The semirandom WD5 library, containing sequences encoding peptides with five Ws and five Ds separated by random amino acids, was constructed from oligonucleotides synthesized at the IDT commercial facility according to the target sequence: ATCTCAGAGGAGGACCTGCATATGGGATGGNNNGATNNNTGGNNNGATNNNTGGNNNGATNNNTGGNNNGATNNNTGGNNNGATNNNTAGGTAGCTATGCGACCTGCAGCGGCCGCATAACTAGCATA where Ns are random nucleotides forming a triplet for a random amino acid.

At the GenScript commercial facility, the oligo was converted into the double-stranded form, digested with *NcoI* and *SalI*, and cloned into the corresponding restriction sites of the parental library vector, as described for the design library. The library complexity was estimated as described above by individual colony counts after transformation and assessed to be ~10^6. The insertions and in-frame fusions with Gal4 DBD were confirmed by PCR and Sanger DNA sequencing for 40 randomly chosen individual plasmid isolations. After the bacterial cloning and verification, the entire plasmid library was isolated for the following yeast transformation.

The isolated WD5 library was transformed into yeast Y2HGold strain, the maintenance of the library complexity was confirmed by the individual colony count for the fraction of the transformation mix, as described above. Screening procedure, sample preparation, and NGS sequencing were also the same as for the design library.

Sequence read processing

The reads from the semirandom WD5 library were processed similarly to those of the random library in [5]. All processing steps aside from Illumina adapter sequence removal were completed using VSEARCH [22]. Forward and reverse read pairs were merged, allowing a maximum of one expected error when considering the quality scores per base. Adapter sequences were removed using cutadapt [23]. Sequences were then deduplicated across all samples, counting the number of times each unique sequence appeared. The sequences were then filtered to include only those with a length of at least 60 bases and appearing at least twice across the library. The deduplicated sequences were clustered between those with a minimum sequence similarity of 90%, in an attempt to prevent two sequences with minor differences from being considered distinct sequences [5]. These sequence clusters were then considered centroids, to which the original reads (merged, without adapters) were mapped and counted. For this step, the sequence identity parameter was set to 80%.

Reads from the design library were mapped using Kallisto[24], thanks to the improved alignment rate offered by the 20-nucleotide barcode present in each sequence of the library. Kallisto performs pseudoalignment, a probabilistic method, to map reads to the design library sequences and their barcodes, thereby providing the abundances of each sequence in the library. For each sample, cutadapt was used to remove adapters and demultiplex by the biorep barcode [23]. Reads were then pseudoaligned to the design library (tAD sequence plus individual barcode) using Kallisto, with a default kmer size of 31.

Estimation of sequence growth rates

Correlations among the read counts of biological replicates were calculated to ensure reasonable consistency, and sequences with at least five counts in at least two of the five biological replicates at baseline (identified by the five unique BioRep barcodes that were appended during PCR amplifications) were retained for subsequent analysis. Sequence counts were then averaged across biological replicates, resulting in one value for each sequence in each sample. These were then normalized within each sample (read counts divided by total reads in sample) to control for overall quantification differences between samples, and normalized to the baseline to quantify cell growth (by calculating the log2 fold change of each sequence at each time point versus its counts at time 0). The result of this step was a set of baseline-centered read counts for each sample at each of the five time points, which could then be plotted to determine whether abundance increases or decreases over time. Robust linear regression (implemented in the MASS package in R) was used to estimate the slope of each sequence over time; this was our final estimate for the functionality of each sequence [25]. Regression of sequence counts vs. day was conducted from day 0 to 4 for most sequences, forcing a y intercept of 0 (because counts were normalized against day 0). Sequences for which the read count dropped and stayed below 3 were regressed from day 0 through the first day at which their read count was below 3. To define a strict binary cutoff for defining functional versus nonfunctional sequences, the 5 highest growth slopes out of 50 total stop codon sequences (the unique sequences that started with a stop codon) were averaged and used for individual sequence classification. For data visualization, all growth slopes were recentered to this cutoff slope so that the cutoff slope became zero.

Sequence feature analysis

Sequence features such as the presence or number of individual amino acids or multiresidue motifs, the balance of aromatic versus acidic residues, and the mixing between amino acid residues were used in machine learning analyses. Balance was defined as the difference between the number of aromatic and acidic amino acids, while mixing was defined as the number of aromatic-acidic dipeptides plus the number of acidic-aromatic dipeptides. Ridge regression was conducted using the caret package in R [26]. Neural network prediction of functionality was performed using the Keras and TensorFlow packages, where each sequence is transformed into a one-hot encoded 3 x 20 matrix (3 amino acids G, W, or D & 20 positions) [27] [28]. The neural network architecture was simple, consisting of a flattened input matrix, two fully connected hidden layers of 60 and 30 nodes, with a dropout rate of 0.2 after each hidden layer, and finally connected to a softmax output layer predicting 1 – functional or 0 – nonfunctional.

Structure prediction and analysis

Secondary structure prediction was performed with SPOT-1D [29]. The predictions were calculated for each candidate 30-aa-long tAD sequence only. The SS3 output of SPOT-1D was then used to assign a helicity percentage to each candidate tAD sequence. The visualization of structure predictions utilized the AlphaFold2 Colab notebook [30]. Sequence structures were predicted for both the candidate tAD and the preceding “linker” (PEFVIRLTIGRAAIMEEQKLISEEDLHMAMG). Visualizations of the candidate tADs were finalized in PyMOL V 1.8; the common “linker” sequence was removed, and key amino acids were colored.

Data and code availability

The datasets generated for this study are available in the Gene Expression Omnibus (GEO) repository, under the SuperSeries GSE200787. These data are currently private and can be accessed by reviewers using the code **crwvssuivzgtvyl**; all data will be made public upon publication. Code used for data analysis and figure generation are provided at https://doi.org/10.5281/zenodo.6461744.

References

1 Luscombe, N. M., Austin, S. E., Berman, H. M. & Thornton, J. M. An overview of the structures of protein-DNA complexes. *Genome Biol.* **1**, REVIEWS001, doi:10.1186/gb-2000-1-1-reviews001 (2000).

2 Hossain, M. A., Barrow, J. J., Shen, Y., Haq, M. I. & Bungert, J. Artificial Zinc Finger DNA Binding Domains: Versatile Tools for Genome Engineering and Modulation of Gene Expression. *J. Cell. Biochem.*, doi:10.1002/jcb.25226 (2015).

3 Ma, J. & Ptashne, M. A new class of yeast transcriptional activators. *Cell* **51**, 113–119 (1987).

4 Abedi, M. *et al.* Transcriptional transactivation by selected short random peptides attached to lexA-GFP fusion proteins. *BMC Mol. Biol.* **2**, 10 (2001).

5 Ravarani, C. N. *et al.* High-throughput discovery of functional disordered regions: investigation of transactivation domains. *Mol. Syst. Biol.* **14**, e8190, doi:10.15252/msb.20188190 (2018).

6 Erijman, A. *et al.* A High-Throughput Screen for Transcription Activation Domains Reveals Their Sequence Features and Permits Prediction by Deep Learning. *Mol. Cell* **78**, 890-902 e896, doi:10.1016/j.molcel.2020.04.020 (2020).

7 Broyles, B. K. *et al.* Activation of gene expression by detergent-like protein domains. *iScience* **24**, 103017, doi:10.1016/j.isci.2021.103017 (2021).

8 Mapp, A. K. & Ansari, A. Z. A TAD further: exogenous control of gene activation. *ACS Chem. Biol.* **2**, 62–75, doi:10.1021/cb600463w (2007).

9 Sanborn, A. L. *et al.* Simple biochemical features underlie transcriptional activation domain diversity and dynamic, fuzzy binding to Mediator. *eLife* **10**, doi:10.7554/eLife.68068 (2021).

10 Tuttle, L. M. *et al.* Mediator subunit Med15 dictates the conserved "fuzzy" binding mechanism of yeast transcription activators Gal4 and Gcn4. *Nature communications* **12**, 2220, doi:10.1038/s41467-021-22441-4 (2021).

11 Erkine, A. M. 'Nonlinear' Biochemistry of Nucleosome Detergents. *Trends Biochem. Sci.* **43**, 951-959, doi:10.1016/j.tibs.2018.09.006 (2018).

12 Wu, Y., Reece, R. J. & Ptashne, M. Quantitation of putative activator-target affinities predicts transcriptional activating potentials. *EMBO J.* **15**, 3951–3963 (1996).

13 Piskacek, S. *et al.* Nine-amino-acid transactivation domain: establishment and prediction utilities. *Genomics* **89**, 756–768, doi:10.1016/j.ygeno.2007.02.003 (2007).

14 Cochran, A. G., Skelton, N. J. & Starovasnik, M. A. Tryptophan zippers: stable, monomeric beta -hairpins. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 5578-5583, doi:10.1073/pnas.091100898 (2001).

15 Boija, A. *et al.* Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains. *Cell* **175**, 1842-1855 e1816, doi:10.1016/j.cell.2018.10.042 (2018).

16 Buhrlage, S. J. *et al.* Amphipathic small molecules mimic the binding mode and function of endogenous transcription factors. *ACS Chem. Biol.* **4**, 335–344, doi:10.1021/cb900028j (2009).
17 Saha, S., Ansari, A. Z., Jarrell, K. A. & Ptashne, M. RNA sequences that work as transcriptional activating regions. *Nucleic Acids Res.* **31**, 1565–1570 (2003).
18 Mitchell, P. J. & Tjian, R. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science* **245**, 371–378 (1989).
19 Ferreira, M. E. *et al.* Mechanism of transcription factor recruitment by acidic activators. *J. Biol. Chem.* **280**, 21779–21784, doi:10.1074/jbc.M502627200 (2005).
20 Oldfield, C. J. & Dunker, A. K. Intrinsically disordered proteins and intrinsically disordered protein regions. *Annu. Rev. Biochem.* **83**, 553–584, doi:10.1146/annurev-biochem-072711-164947 (2014).
21 Bondos, S. E., Dunker, A. K. & Uversky, V. N. On the roles of intrinsically disordered proteins and regions in cell communication and signaling. *Cell communication and signaling : CCS* **19**, 88, doi:10.1186/s12964-021-00774-3 (2021).
22 Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahe, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584, doi:10.7717/peerj.2584 (2016).
23 Martin, M. Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads. *EMBnet.Journal* **17**, 10-12, doi:https://doi.org/10.14806/ej.17.1.200. (2011).
24 Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525-527, doi:10.1038/nbt.3519 (2016).
25 Venables, W. N., Ripley, B. D. & Venables, W. N. *Modern applied statistics with S*. 4th edn, (Springer, 2002).
26 Kuhn, M., Jed Wing, Steve Weston, Andre Williams , Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan, Tyler Hunt. caret: Classification and Regression Training. R package version 6.0-86. (2020).
27 Chollet, F. *Keras*, <https://github.com/fchollet/keras> (2015).
28 Abadi, M. *et al.* TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. (2015).
29 Singh, J. *et al.* SPOT-1D-Single: Improving the Single-Sequence-Based Prediction of Protein Secondary Structure, Backbone Angles, Solvent Accessibility and Half-Sphere Exposures using a Large Training Set and Ensembled Deep Learning. *Bioinformatics*, doi:10.1093/bioinformatics/btab316 (2021).
30 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589, doi:10.1038/s41586-021-03819-2 (2021).

Acknowledgements

We thank Marcos Oliveira and Daisuke Kihara for discussion of results and constructive suggestions, and Charles N.J. Ravarani for initial library design discussions. This research was supported by grants to A.M.E. and Daisuke Kihara from the National Science Foundation (MCB 1925646, MCB 1925643).

Contributions

A.M.E. conceptualized the study. A.M.E., B.K.B, and A.T.G. designed experiments. T.Y.E. performed wet lab experiments. A.M.E., B.K.B, C.A.C. T.P.M., D.A.C., T.M.W. analyzed data. B.K.B., and C.A.C. produced visualizations of the results and made figures. A.M.E wrote the manuscript.

Ethics declarations

The authors declare no competing interests.

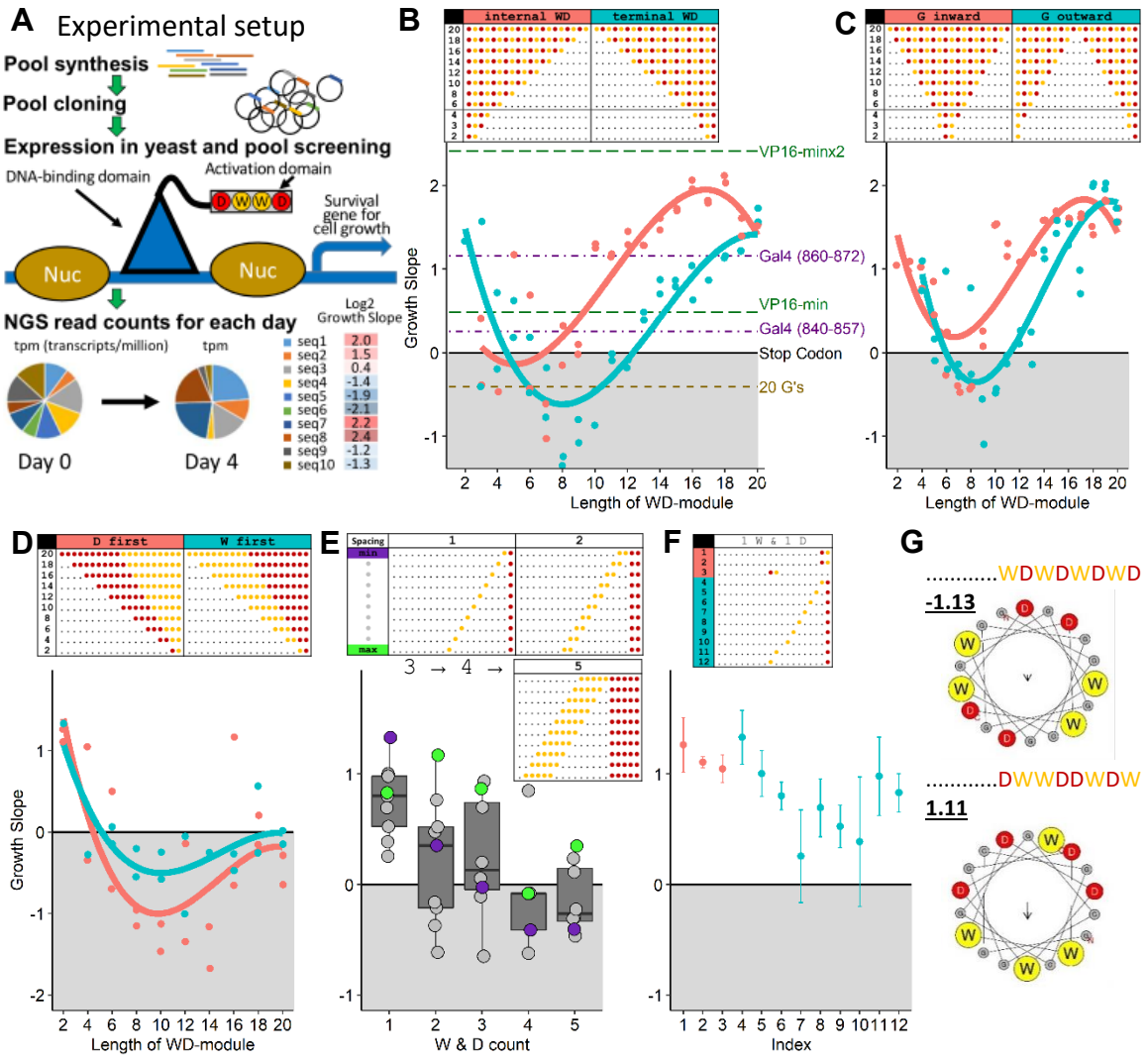
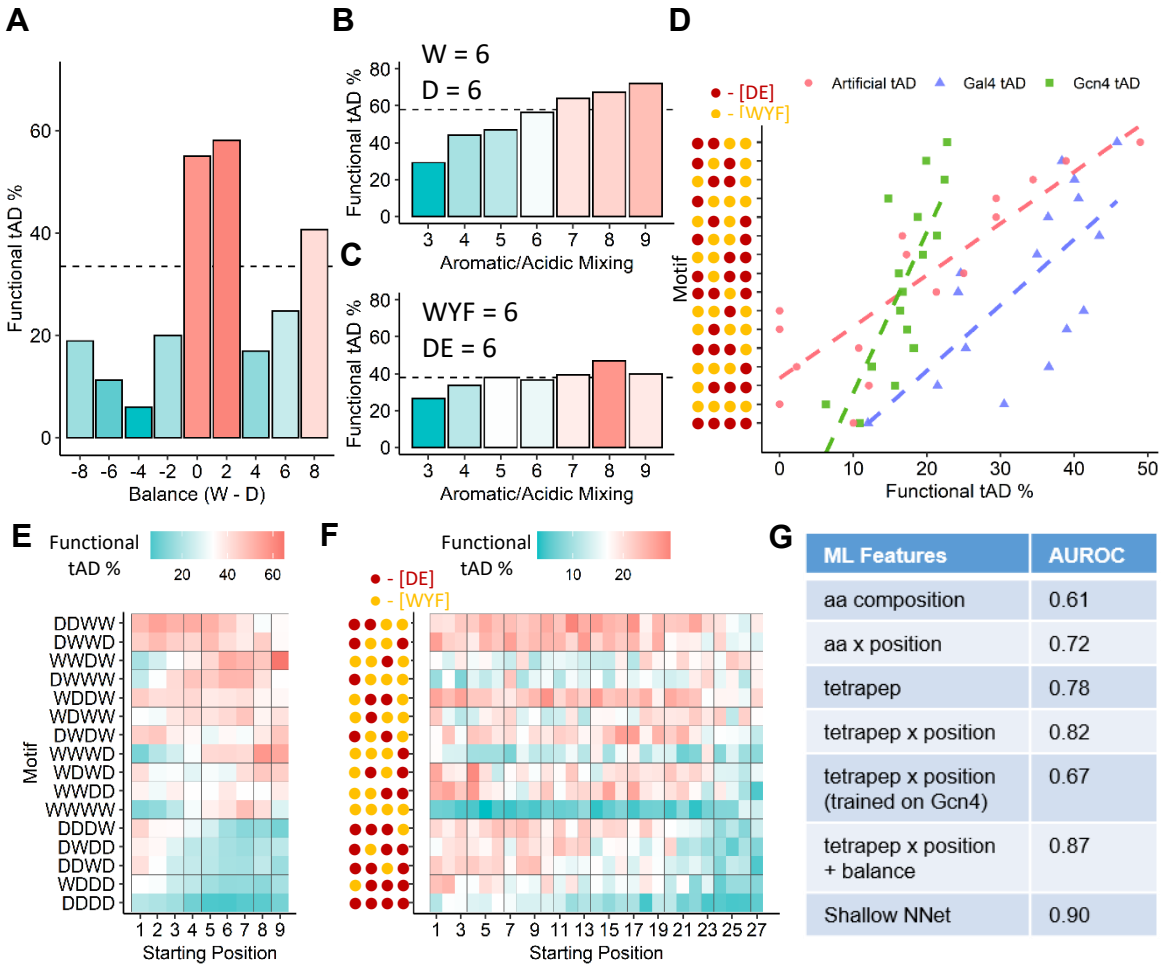


Figure 1. A single W and a single D are sufficient for the functionality of AD. A – Experimental setup: oligo pool synthesis, followed by cloning in bacteria, then isolation of plasmid library and transformation in yeast, followed by screening for growth phenotype determined by expression of the reporter gene regulated by the activator with a specific AD, then isolation of DNA pool, NGS sequencing, and data analysis (for more details see Methods section and Supplementary Figure 1). **B** – Growth of sequences with different numbers of WD repeats. X axis: individual sequences indicated in the inset table, where black dots represent glycine, yellow dots represent tryptophan, and red dots represent aspartic acid residues. Y axis: Log2 growth slope. Axes are same in C, D, E, F. **C** – Sequences with different numbers of WD repeats, either surrounded or interrupted by repeated glycines. **D** – Sequences with non-alternating clusters of Ws and Ds. **E** – Sequences with non-alternating clusters of Ws and Ds, separated by varying numbers of Gs. **F** – Sequences with a single W and D, separated by varying numbers of Gs. Error bars show growth slope +/- root-mean-square-deviation (RMSD) of the fit of the growth slope. **G** – Log2 growth slopes and images of the α-helix frontal view for two different sequences with four W's and four D's, showing that despite identical composition, only one is functional.



ML Features	AUROC
aa composition	0.61
aa x position	0.72
tetrapep	0.78
tetrapep x position	0.82
tetrapep x position (trained on Gcn4)	0.67
tetrapep x position + balance	0.87
Shallow NNet	0.90

Figure 2. The balance and intermixing of acidic and aromatic residues is beneficial for AD function. A – X axis: balance score, calculated as Balance=n(W)-n(D); Y axis: % of functional sequences in the sub-library of sequences containing all combinations of W and D for 12 positions (WD12 library, 3968 sequences quantified). **B** – X axis: mixing score, calculated as Mixing=n(WD)+n(DW); Y axis: % of functional sequences in the set of sequences containing all combinations of 6 W and 6 D (906 sequences quantified). **C** – X axis: mixing score, calculated as Mixing=n([WYF][DE]) + n([DE][WYF]) for the previously published AD dataset screened within the Gcn4 context [6]; Y axis: % of functional sequences in the set of sequences from the Gcn4 random peptide library with 6 [WYF] and 6 [DE] (3018 sequences total). **D** – X axis: % functionality of sequences that contain the specified tetrapeptide motif; Y axis: tetrapeptide motifs. Regression lines are provided to demonstrate concordance between the three libraries, and motifs were ordered based on average % functionality between the three libraries, with the most functional on top. **E** – X axis: Starting amino acid position of tetrapeptide in tAD module for the WD12 library; Y axis: 16 sequence combinations for tetrapeptides containing D and W, Tile fill: % functionality of sequences that contain the specified tetrapeptide motif at the indicated position. Motifs were ordered by overall % functionality. **F** – X axis: Starting amino acid position of the tetrapeptide in the tAD module for the Gcn4 library [6]; Y axis: 16 sequence combinations for tetrapeptides containing [DE] and [WYF], Tile fill: % functionality of sequences that contain the specified tetrapeptide motif at the indicated position. Motifs order is the same as in panel E. **G** – ML accuracy on the reserved testing set (20% of WD12 library) and trained on 80% of WD12 library unless noted otherwise, measured as area under the receiver operating characteristic (AUROC).

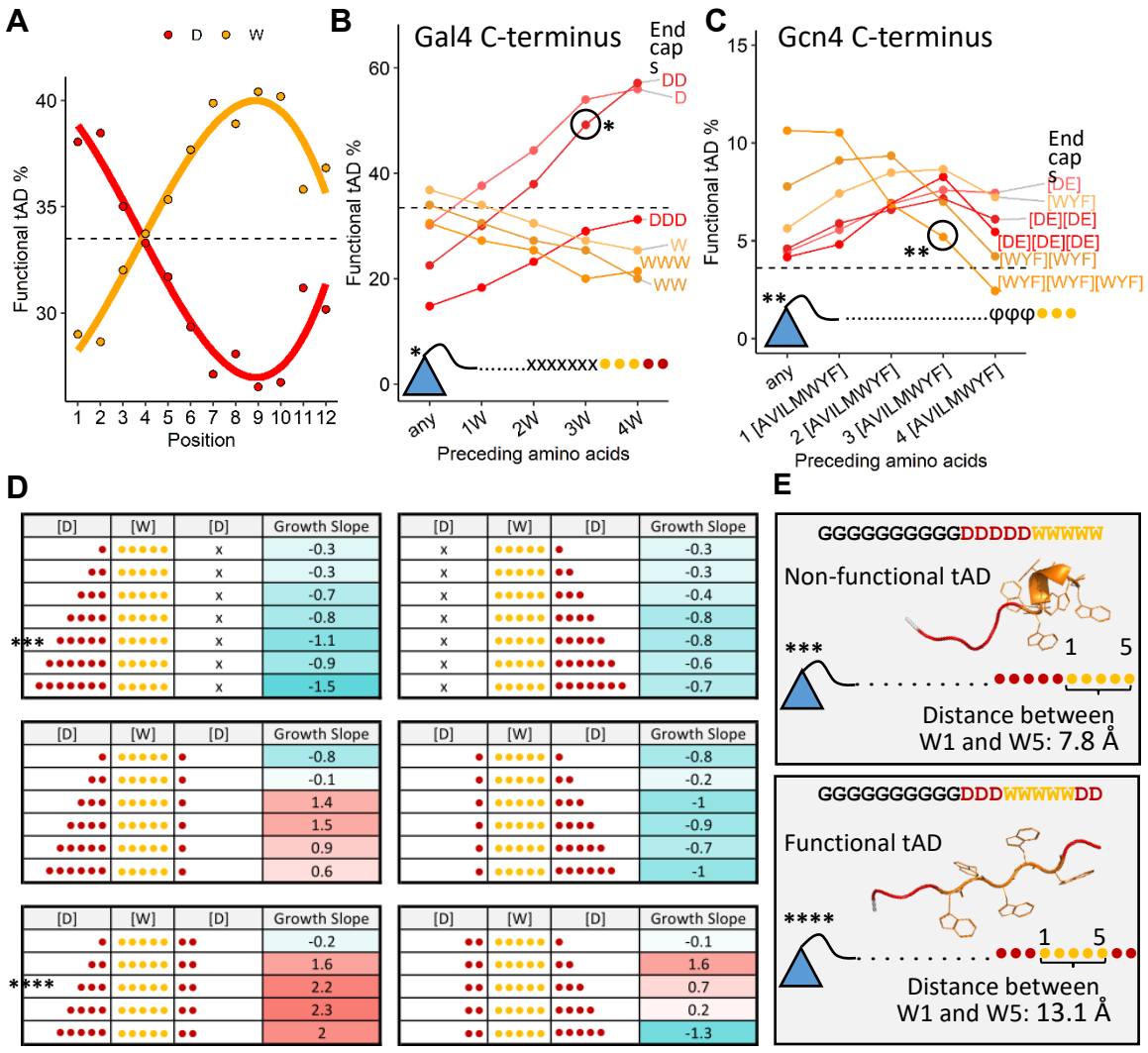


Figure 3. Ws are generally beneficial at the end of the molecule, while Ds – internally, with the exception when Ds rescue the AD functionality by flanking adjacent Ws in the sequence. A – X axis: position of D (red) or W (yellow) within the sequence; Y axis: % functionality of sequences in the sub-library of sequences containing all combinations of W and D for 12 positions (3968 sequences quantified). **B** –X axis: size of W cluster preceding indicated end cap for sequences representing each line; Y axis: same as in A. (*) sequence construct shows tAD constructs of indicated sequence where "." = G, "x" = [DW], ● = W, and ● = D. **C** – same as in B, calculated for the Gcn4 library [6] using [WYF] instead of just W and [DE] instead of just D for end caps. (**) sequence construct shows tAD constructs of indicated sequence where "." = any AA, ϕ = [AVILMWYF], and ● = [WYF]. **D** – Growth slopes of sequences with different numbers of Ds flanking a stretch of 5 Ws. ● = W, and ● = D. **E** – sequence constructs of (***) and (****) sequences from panel D, "." = G, ● = W, and ● = D. AlphaFold2 predicted structures shown for tAD region. Distance between α-carbon of first tryptophan (W1) and last tryptophan (W5) was measured from predicted structure.

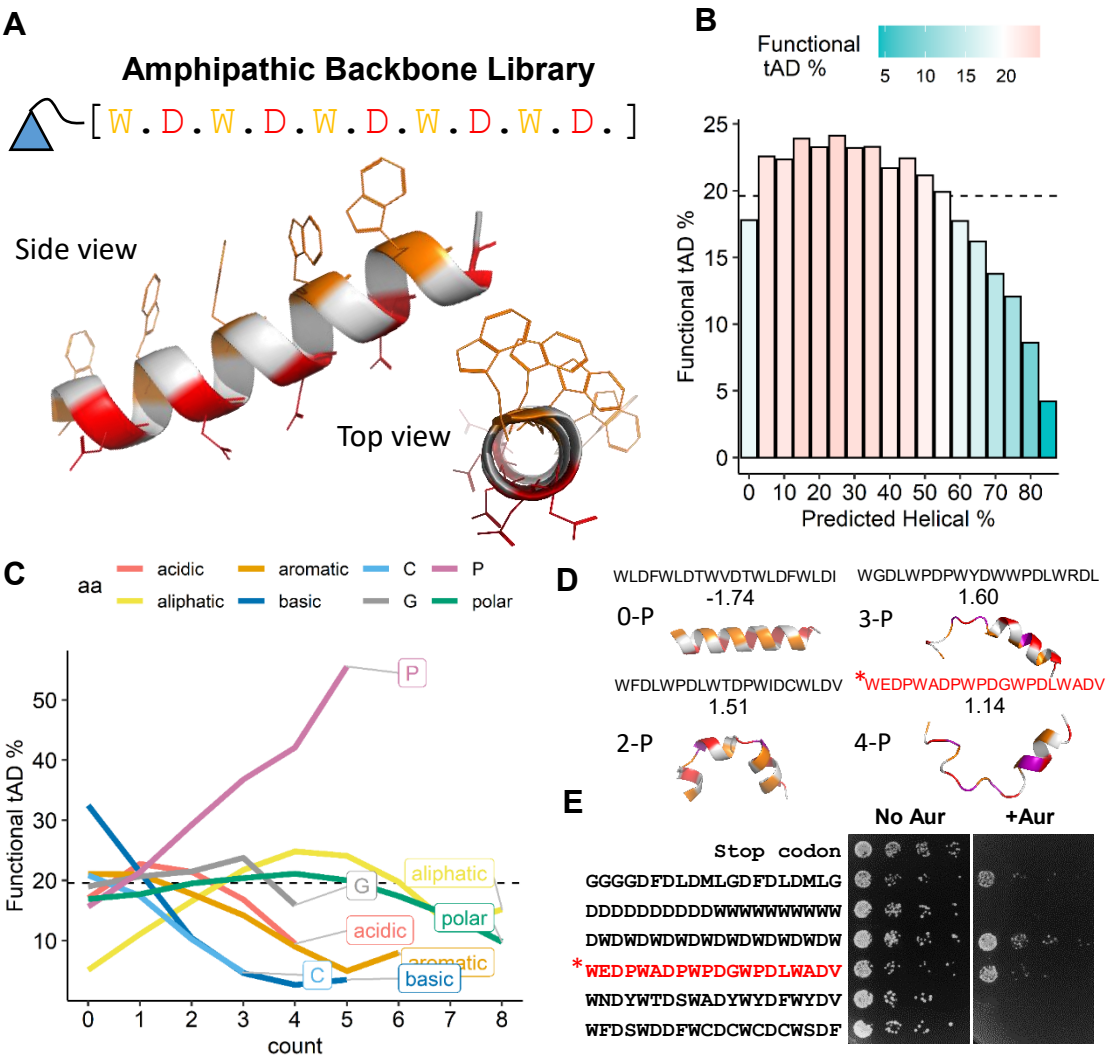
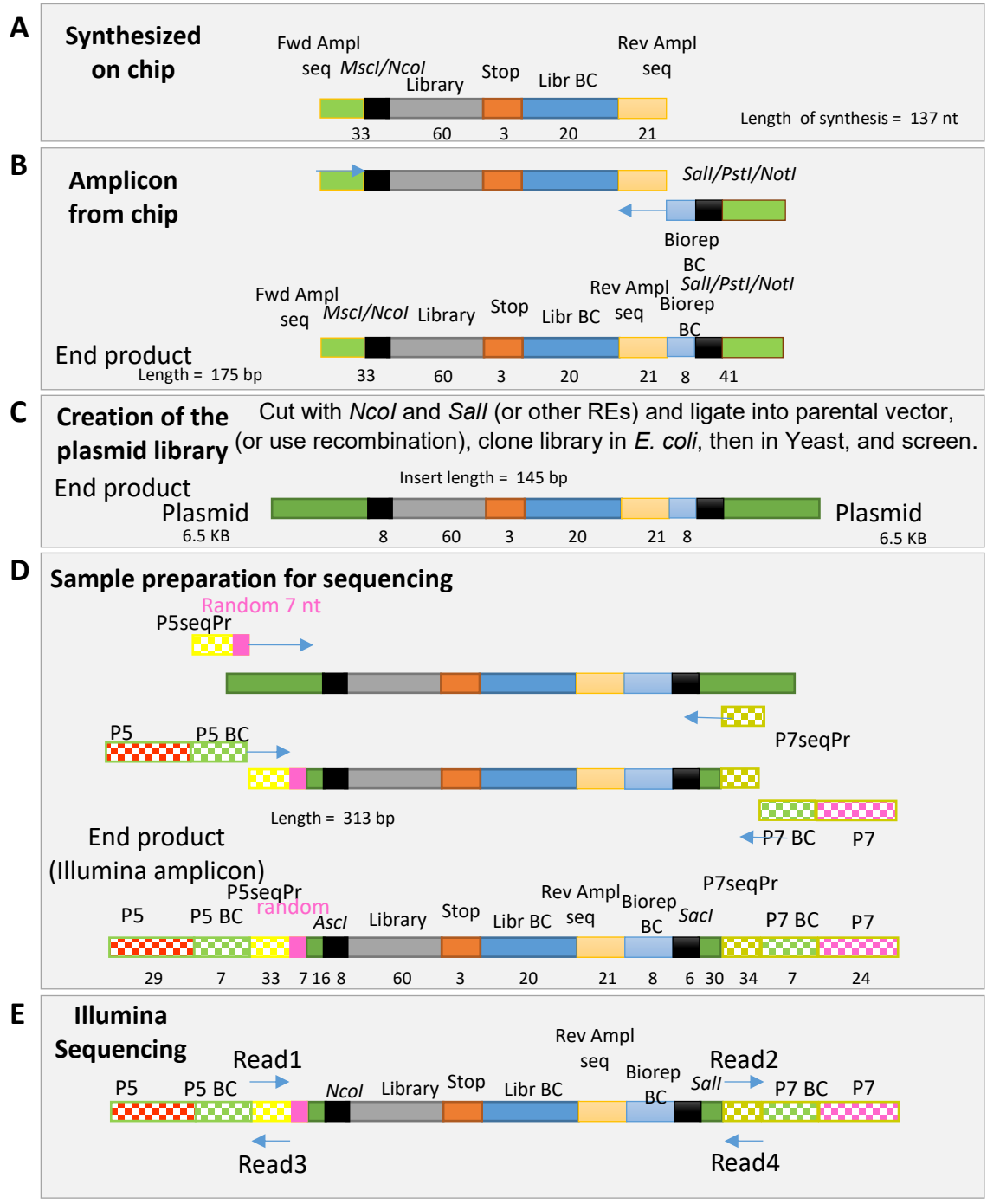


Figure 4. Formation of the amphipathic α-helix is generally detrimental for AD functionality, and insertion of proline is beneficial. A – graphical representation of sequences for the WD5 library (107975 sequences quantified): each member of which has five Ws, five Ds, and ten random amino acids represented by black dots/circles between Ws and Ds. **B** – X axis: % α-helix predicted by the SPOT-1D algorithm for each set of sequences from the WD5 library; Y axis: % functionality of the set of corresponding sequences in the WD5 library. **C** – X axis: count of corresponding amino acid residues between set Ws and Ds of WD5 library; Y axis: % functionality of the set of corresponding sequences. Amino acid groups: acidic [DE], aliphatic [AVILM], aromatic [WYF], basic [RHK], special [CGP] not grouped, polar [STNQ]. **D** – Growth slopes with 3D structures of sequences with varying numbers of proline residues predicted by AlphaFold2. **E** – Growth phenotype on media with and without aureobasidin for cells expressing the indicated representative sequences. Spots are conglomerates of yeast colonies representing threefold serial dilutions of corresponding cell cultures.



Supplementary Figure 1. Schematic representation of wet lab steps: **A** – massively 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.