2019 Library Design

Rationale for WD(G) library design:

The information that we already gained using ML approaches to analyze nonbiased random library was critical to formulate the transformative nucleosome detergents model and to compare it to the conventional direct recruitment model  (PMID: 29759983, PMID: 30297207). The analysis of TAD features and the refinement of the mechanistic models can be greatly extended. One of the current limitations is the relatively small size of the usable for ML dataset fraction. In case of the random library, only 1% of the pool represented functional sequences suitable for ML analysis, together with another 1% used for equilibration of the functional dataset, while the rest 98% being dimmed largely as a ballast. Another problem with the actual experimental random library is that it represents a tiny sampling of the total combinatorial space (2020), which together with the small size of the experimental dataset greatly reduces precision of ML models and the resolution between features comparing their ML gain, which is critical for determination of molecular mechanisms.

To overcome these drawbacks we will create a new library using the information that we already gained from our preliminary studies. Since we know that combination of aromatic and negatively charged residues is important for functionality, G is neutral, we will use W and D amino acids for creation of functional combinations, G as a spacer.

Sequences within the 2019 WD Library

1. library\_2019$combinatorial
   1. we want to investigate all combinations of W (aromatic) and D (acidic) for 1 to 11 spaces
      1. W, D (21 sequences)
      2. WW, DD, WD, DW (22 sequences)
      3. 211 sequences or 2,048 sequences
      4. 4094 sequences total
   2. Glink­: we used G as a filler to link the activity module to DNA binding domain
      1. Glycine is proposed to be neutral to functionality
   3. We want to use this portion of the library for broad analysis and machine learning
   4. Is.supp: this indicates if a sequence is covered within the combinatorial set or if it is part of the supplemental set.
2. library\_2019$W\_first and D\_first
   1. we want to compare aromatics followed by acidics and vise versa and the effects on functionality.
   2. 10 sequences each for a total of 20 sequences
      1. GGGGGGGGGGGGGWWWWWDDDDDD
      2. GGGGGGGGGGGGGDDDDDDDWWWW
3. library\_2019$set\_g\_spacing
   1. we start with one aromatic and one acidic and progressively space them out with G
      1. GGGGGGGGGGGGGWD
      2. GGGGGGGGGGGGWGD
      3. Ect all the way out to 9 G spaces
   2. we did the same thing for 2 aromatic and 2 acidic ect all the way out to 5 aromatics and acids
   3. we are trying to understand optimal distance between aromatic and acidic residues to create functional tADs
   4. 50 sequences in total
4. Library\_2019$g\_in and g\_out
   1. We are investigating if tADs perform better with one large functional region or two smaller functional regions.
      1. 40 sequences total
   2. G\_out
      1. WDWDWDWDGDWDWDWDWD
      2. WDWDWDWDGGWDWDWDWD
      3. WDWDWDWGGGWDWDWDWD
   3. G\_in
      1. GWDWDWDWDWDWDWDWDWG
      2. GGDWDWDWDWDWDWDWDGG
      3. GGGWDWDWDWDWDWDWGGG
5. Library\_2019$terminal\_wd and internal\_wd
   1. We are trying to see if the position of the activity module is essential for functionality
      1. Is it better to have the activity module close to the DBD or give it more spatial freedom by putting them at the end terminus?
   2. Terminal\_wd
      1. GWDWDWDWDWDWDWDWDWDW
      2. GGWDWDWDWDWDWDWDWDWD
      3. NOTE: these sequences double as the sequences for length
   3. Length: we are trying to find the essential length for tADs and we are extending the active module out to 20 postions
   4. Internal\_wd
      1. WDWDWDWDWDWDWDWDWDWG
      2. WDWDWDWDWDWDWDWDWDGG
      3. WDWDWDWDWDWDWDWDWGGG
6. Library\_2019$oneR, twoR, threeR
   1. We know basics are detrimental to functionality, but we have several sequences with basics that maintain functionality.
      1. We are trying to better understand the rules for when R is the most detrimental
   2. One way to describe these sets is ‘R scanning’ where we do a flythrough in which R gets substituted at every position (in both the G linker and the activity module)
   3. RGGGGGGGGGWDWDWDWD
   4. GRGGGGGGGGWDWDWDWD
   5. GGRGGGGGGGWDWDWDWD
   6. Ect
7. Natural tADs
   1. This set acts as proof of concept for this new screen, demonstrating that naturally occurring tADS are also functional within this screen. (Pulled from Charlee’s GitHub)
      1. AH, EBNA2, EKLF TAD1 (19-37), ESX, Gal4 (840-857), Gal4 (860-872), Gln3, hHSF1 AD1 (401-420), KLF4, LexA scr top#1, Msn2 (235-246), Msn2 (motif b), Oaf1 (1035-1047), Pho4, plantHSFA2, STAT6 (1OJ5), TAT(1-13), VP16 (437-447), VP16 21aa, VP16 min, vP16 min x 2, Pdr1, p53-AD1
8. Random\_WD20
   1. We know WDx10 (length 20) works but we haven’t investigated how different combinations of WD with a length 20 functions.
      1. Random subsampling up to length 20
         1. 50 sequences total
9. Dipeptide Repeats
   1. It has been shown that a dipeptide sequences can be highly functional (WDx10) but we do not know how other dipeptides work. This set investigates each possible dipeptide combination.
   2. We are also testing sequences which are purely one amino acid (all W ect)
   3. 800 sequences total
      1. Each sequences has length 20
10. Equal R&D and equal K&D
    1. These sets have an equal number of negative and positive amino acids for a net zero charge. This also tests the killing power of K and R to compare if they are both as detrimental for functionality. It also tests to see having multiple positive charged amino acids can be counteracted with the negative amino acids.
       1. 1,024 sequences
11. Dr. Kihara nucleosome binding sequences
    1. Pending
    2. 1000 sequences
12. Stop codon:
    1. 10 sequences
    2. NT code for 10 stop codon sequences so we have a threshold for binary stop
13. Umayr Set: submitted sequences from our journal club
    1. 10 sequences
14. Histone Binding Set: these sequences are known to interact with histones and this set was derived from the APLF motif and the DEFY motif
    1. NOTE: several of the APLF motifs were 24 amino acids long therefore 4 amino acids were removed from the beginning or the end of the sequences
       1. If the name has ‘\_a’ then the amino acids were removed from the end of the sequences (to achieve length 20)
       2. If the name has ‘\_b’ then the amino acids were removed from the front of the sequences (to achieve length 20)
15. Motifs
    1. This set is designed to test the different tAD motifs which are found within the literature (Staby et al). We will test a random sample of 50 sequences which match each motif’s regex.
    2. Hypothesis: there are numerous instances of these motifs within the random library, the functionality of the motifs ranged from between 1-3% functional. We believe that we will find similar functionality within these motifs in this screen.
    3. There will be an additional set where we double the motifs. (WxxLF-> WxxLFWxxLF) an additional
       1. 100 total sequences per motif
16. WxxLF
    1. 400
17. P53: [ALVIMWYF]..[ALVIMWYF][ALVIMWYF]
    1. 204,000
18. RelA\_2: [ALVIMWYF]..[ALVIMWYF][ALVIMWYF]
    1. 655,360,000
19. CREBZF: D[VILM][VILM][RKDEQNHSTYC][RKDEQNHSTYC][VILM][VILFWYM]
    1. 54,208
20. AR: F..LF
    1. 400
21. ANAC013: [DE].{1,2}[YF].{1,4}[DE]L
    1. 565,891,200
22. EKLF: [RKDEQNHSTYC][ALVIMWYF][ALVIMWYF]..[ALVIMWYF]..[RKDEQNHSTYC][RKDEQNHSTYC]
    1. 109,035,520,000

Sequences from the Daisuke Kihara Lab at Purdue University

The lab generated a large library of 3,000 novel peptide sequences with three smaller 1,000 sequence sub-libraries within it. The first library was based on a set of 16 sequences sent from the Erkine lab, the Daisuke lab determined the amino acid distribution of this set and then generated 1,000 new sequences based on the determined amino acid distribution. The second sub-library was an additional 1,000 sequences which were based off of the amino acid distribution of proteins from a *protein data base (which one?).* The final 1,000 sequences were based off a combination of the amino acid distributions of the first set of 16 sequences as well as the protein data base. This last set of sequences were each 15 amino acids in length. They randomly selected 8 positions which were given amino acids based off of the distribution form the set of 16 sequences. The remaining 7 positions were filled with amino acids based off the distributions calculated from the protein data base.

This library of 3,000 sequences was then ran though the ‘Otodoc’ (need to find the actual name) software which quantifies peptide/receptor docking. In this case they measured the interaction between the peptide library and the acidic patch on the nucleosome. Each sequence was tested 50 times based off of 50 different confirmations of the protein, and they collected the binding energy.

NOTE: We are currently at 13,065 sequences

* 665 sequences too many
  + We need to discuss which sequences to remove

Questions to Investigate

1. Is the length of tAD sequence critical for functionality?
   1. Hypothesis: Length and functionality are correlated to an extent, as length increases so does percent live until we reach the minimal length for a detergent motif. Then the functionality is expected to plateau off.
   2. Sets to answer this question:
      1. Combinatorial space: for length 1-11
      2. Terminal\_WD: for length 12-20
2. Redundancy: how does repeating motif’s impact functionality?
   1. Hypothesis: As redundancy increases so does functionality
      1. Analysis: repeat of tripeptides (each combination of WD for 3 positions; 23)
         1. WWW, WWD, WDD, DDD, WDD, WWD, WDW, DWD
   2. Sets to answer this question:
      1. Combinatorial space: we plan to create a feature ‘redundancy’ and apply it to the entire library
3. How do highly acidic or aromatic regions impact functionality, is it beneficial to have a high concentration of acidic or aromatic at the beginning, middle or end of the tAD?
   1. Hypothesis: highly acidic or aromatic regions are not beneficial to functionality
   2. Sets to answer this question:
      1. Combinatorial space
4. Homogeneous vs Heterogenous: is it beneficial to have a more homogenous sequence (GGGG[WW][DDDD][WWW][DD]) or a heterogenous sequence (GGGG[D][W][D][W][D][W][D][W][D][W][D])?
   1. Hypothesis: more heterogenous sequences will perform better.
   2. Sets to answer this question:
      1. Combinatorial space
5. WD ratio: is there a beneficial ratio found within functional sequences?
   1. Hypothesis: an even spread between W and D will improve functionality. Too much acidic or aromatic will be detrimental to functionality
   2. Sets to answer this question:
      1. Combinatorial space
6. Internal vs Terminal activity module (GGGGGGGGGWDWDWDW) vs (WDWDWDWDGGGGGG)
   1. Hypothesis: Having the activity module in the terminal position will allow special freedom and thus increase activity
   2. Sets to answer this question:
      1. Internal\_WD (20 sequences)
      2. Terminal\_WD (20 sequences)
7. Is it better to have acidic amino acids before aromatic amino acids and vise versa? (WWWWDDDD) vs (DDDDDDWWWWW)
   1. Hypothesis: W at the end of the sequence will improve functionality (based off preliminary data in the WAD grid)
   2. Sets to answer this question
      1. Combinatorial space
      2. W\_first
      3. D\_first
8. How will spacing of W and D extremities effect functionality? (WWWGGGGGGGDDD)
   1. Hypothesis: We predict that having a spacer between 2-5 will be detrimental, but longer than 5 may be functional because the sequence will be able to fold upon itself allowing the W and D to interact.
   2. Sets to answer this question:
      1. Set\_g\_spacing
9. Is it better to have two small activity modules (WDWDWDGGGGGDWDW) or one large activity module (GGGWDWDWDWDWDWDWDWDGG)?
   1. Hypothesis: sequences with two small activity modules will be more functional because G is flexible and will allow the two modules to interact
   2. Sets to answer this question:
      1. G\_out
      2. G\_in
10. How do basic amino acids affect functionality?
    1. Hypothesis: we believe that having several W’s and D’s will outweigh the detrimental effect of R in some sequences.
       1. Because we are doing a scan, we can determine which positions R is detrimental
          1. Beginning, middle or end of the sequence
    2. Sets to answer this question:
       1. OneR
       2. TwoR
       3. ThreeR