Bio 131 Final Project

August Staubus

Background:

Restriction Enzymes are a valuable molecular tool used in virtually all instances where DNA must be manipulated. Their utility arises from their unique ability to cut DNA at specific sites rapidly, efficiently, and with high fidelity. While over 100,000 restriction enzymes are known, only those originating from closely related bacteria or those that recognize the same restriction site show a significant degree of sequence homology [1]. Historically, the low homology between these enzymes was believed to indicate that these enzymes did not share a common evolutionary origin [2]. However, this hypothesis began to be reversed as structural motifs began to be observed among many restriction enzymes. Among the most significant of these motifs is a proline residue followed by an aspartate residue, then a glutamate or aspartate residue followed by any residue followed by a lysine residue, with the two motifs separated by 10-30 amino acids. This motif, abbreviated as PD…D/EXK, is widely conserved, and comprises residues directly involved in catalysis and is therefore referred to as a catalytic motif.

This motif was originally discovered not by computational analysis, but by crystallographic inspections of structure and mutational analysis of function [3, 4]. Here, I attempted to develop an algorithm that could identify this motif in the sequences of restriction enzymes known to harbor the PD…D/EXK motif and could therefore potentially identify the motif in other, unanalyzed protein sequences.

Data:

To develop this algorithm, I downloaded the amino acid sequences of eight restriction enzymes known to have the PD…D/EXK motif (BamHI, BglI, BsoBI, EcoRI, EcoRV, MunI, PvuII and RsrI) in FASTA format from the national center for biotechnology information (NCBI). These files were then read into the python code and the first line of each file (containing the file name and protein name, among other comments) was removed. The following lines were then concatenated, resulting in eight continuous strings each comprising the full amino acid sequence of each protein. These strings were then compiled into a list. A second list was created containing the names of the proteins in the same order as their respective sequences.

High Level Algorithm Steps:

In essence, the algorithm I developed to find the PD…D/EXK motif is a greedy algorithm with two random elements. To begin with, the order of the list of amino acid sequences is shuffled. Then, a random k-mer is selected from the first sequence in the shuffled list and added to an empty list of putative motifs. A profile is then generated from that k-mer, wherein the counts of each amino acid in each position is calculated, i.e. the number of alanines in the first position, the number of glycines in the first position, etc. then the number alanines in the second position, the number of glycines in the second position, etc. During the first iteration, the counts of all amino acids but one in every position are 0, and the count of one amino acid (the one that appears the selected k-mer) is 1. A pseudocount of 1 is added to each count, and then the counts are converted into a frequency by dividing each count by the sum of all counts for that position. Using this frequency profile, the most probable k-mer of the second sequence in the shuffled list is found by calculating the product of the frequency of each amino acid in a given position in each possible k-mer in the sequence. The profile-most-probable k-mer from the second string is then added to the list of putative motifs. The counts are again calculated for the two k-mers now in the list of putative motifs, and this process is iterated through all sequences in the shuffled list of sequences. Once the iterations are complete, an consensus motif is calculated from the list of putative motifs by finding the most frequent amino acid at each position of the k-mer. The list of motifs is then given a score equal to the sum of the hamming distance between each putative motif and the consensus motif. Then, a new random k-mer is selected from the first sequence in the shuffled list, and again the process is iterated through all sequences in the shuffled list, an consensus motif determined, and a score calculated. If the score is greater than any score observed so far, the new list of putative motifs is saved as the current "best" motifs. The process of random k-mer selection and subsequent motif generation and scoring is repeated n times. Currently, n is coded to be proportional to the length of the first sequence in the shuffled list of sequences, but can be easily changed to another integer value. After repeating n times, the list of restriction enzyme sequences is again shuffled. The shuffling process and subsequent steps are repeated i times, where i is an integer, and the motif ensemble with the highest score from all i runs is reported, along with the consensus motif and score. The algorithm is summarized in figure 1.

This algorithm also outputs a plot of the highest motif score as a function of the number of times the algorithm is run (i). This plot may be used as a diagnostic to determine whether i and n are sufficiently large. If the highest score does not increase over the course of many runs, it is likely that the optimal motif has been discovered.

```
1.  Shuffle List Of Sequences
2.      Select a random k-mer from the first item in Shuffled List of Sequences
3.      Add k-mer to List of Putative Motifs
4.      Create a frequency profile for the List Of Putative Motifs
5.          Find the Profile-Most-Probable k-mer from the next sequence in the
            Shuffled List of Sequences
6.          Add the Profile-Most-Probable k-mer to the List of Putative Motifs
7.          Repeat steps 4-6 for all sequences in Shuffled List of Sequences
8.      Determine Consensus Motif for List of Putative Motifs
9.      Calculate Score of List of Putative Motifs
10.     If Score is greater than current Best Score, save Score as Best Score
11. Repeat steps 2-10 n times
12. Repeat steps 1-11 i times
13. Return List of Putative Motifs with the highest score from all i runs as well as the
    Consensus motif and Best Score
```

Figure 1: The high-level steps of the algorithm for finding motifs in amino acid sequences.

Results and Discussion:

Because the algorithm requires a length of k-mer to look for as an input, this algorithm is not particularly adept at identifying motifs of variable length. Unfortunately, the length of the PD…D/EXK motif *is* widely variable, with the number of residues separating the PD and D/EXK regions varying from under 10 to over 40 [1]. To overcome this difficulty, I tried searching for 2-mers (in an effort to identify the PD motif) and 3-mers (to find the EXK motif) independently. However, even after setting i to 100, the algorithm returned motifs of variable score and sequence for both 2-mers and 3-mers (Figure 2) and only rarely would the motifs include the desired sequences (Figure 2). Many motifs have equivalent scores to either PD or D/EXK. It would appear that this motif is too poorly conserved to be identified using this computational method.

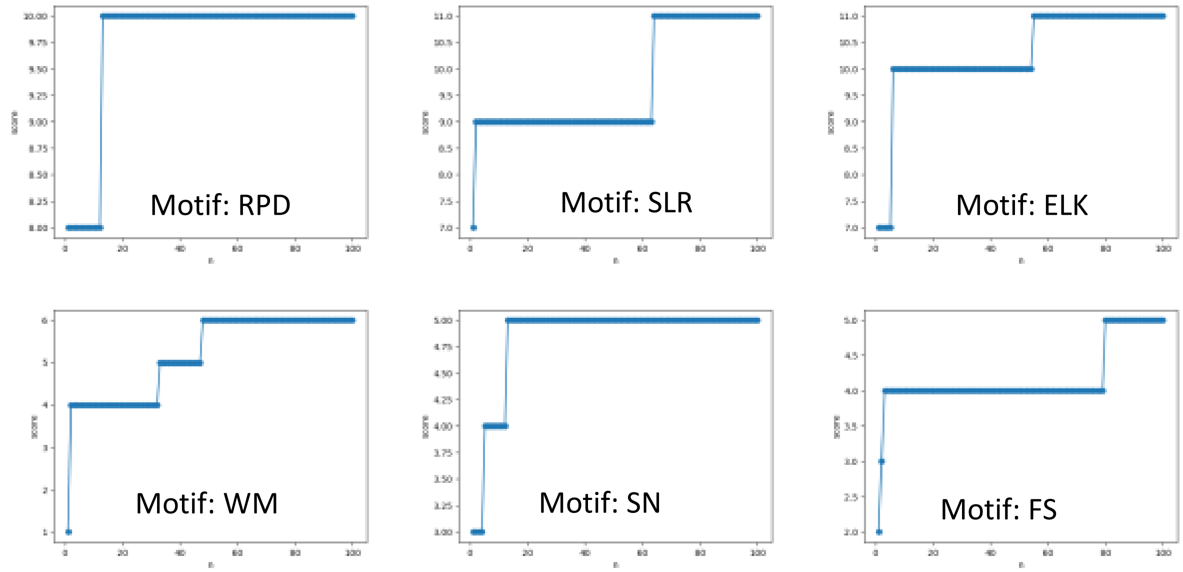
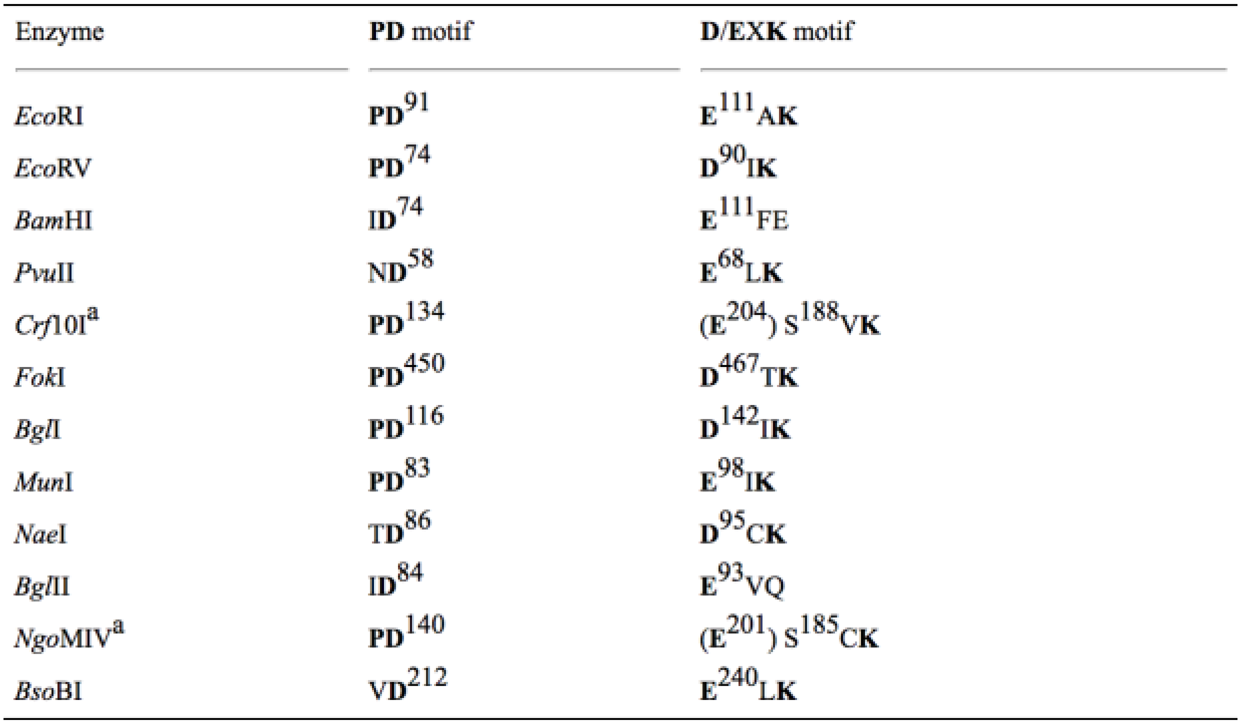


Figure 2: The algorithm returns variable motifs. The above plots show the score of the highest score motif observed after i runs of the algorithm. Top row: Three representative examples of the randomized algorithm searching for 3-meric motifs where i = 100. Note that each returns a different motif, that the rightmost two have a score of 11 while the left has a score of 10, and that only the right most returned a motif containing the desired sequence: EXK. Bottom row: Three representative examples of the randomized algorithm searching for 2-meric motifs where i = 100. Again, each returns a different motif, the rightmost two have a score of 5 while the left has a score of 6, and none return the desired motif (PD).

Indeed, further investigation revealed that the PD…D/EXK motif is more weakly conserved than I was lead to believe (Figure 3). The PD motif is not strictly conserve; often the P is absent. Additionally, the EXK is often a EXXK motif, further frustrating my efforts to identify the motif.

There is a rich history of those more qualified than I developing and refining algorithms meant to perform virtually the exact task I endeavored to perform [5-7]. The most successful of these techniques utilize structural models such as hidden Markov models in addition to sequence homology analysis. Thus, it would appear that inspection of the primary amino acid sequence alone is insufficient to identify this motif. Rather, a combination of analyses of the primary and tertiary structure is required.

Table 1. The catalytic motif of several restriction enzymes. Note how weakly conserved the motif is.



Enzyme	**PD** motif	**D/EXK** motif
*Eco*RI	**PD**^{91}	**E**^{111}**AK**
*Eco*RV	**PD**^{74}	**D**^{90}**IK**
*Bam*HI	**ID**^{74}	**E**^{111}FE
*Pvu*II	ND^{58}	**E**^{68}L**K**
*Crf*10I[a]	**PD**^{134}	(**E**^{204}) S^{188}V**K**
*Fok*I	**PD**^{450}	**D**^{467}T**K**
*Bgl*I	**PD**^{116}	**D**^{142}I**K**
*Mun*I	**PD**^{83}	**E**^{98}I**K**
*Nae*I	TD^{86}	**D**^{95}C**K**
*Bgl*II	ID^{84}	**E**^{93}VQ
*Ngo*MIV[a]	**PD**^{140}	(**E**^{201}) S^{185}C**K**
*Bso*BI	VD^{212}	**E**^{240}L**K**

1. Pingoud, A., et al., *Type II restriction endonucleases: structure and mechanism.* Cellular and Molecular Life Science, 2005. **62**: p. 685-707.
2. Wilson, G.G. and N.E. Murray, *Restriction and modification systems.* Annual Review of Genetics, 1991. **25**(585-627): p. 603.
3. Anderson, J., *Restriction endonucleases and modification methylases.* Curr. Opin. Struct. Biol., 1993. **3**(1): p. 24-30.
4. Thielking, V., et al., *Site-directed mutagenesis studies with EcoRV restriction endonuclease to identify regions involved in recognition and catalysis.* Biochemistry, 1991. **30**(26): p. 6416-22.
5. Kosinski, J., M. Feder, and J.M. Bujnicki, *The PD-(D/E)XK superfamily revisited: identification of new members among proteins involved in DNA metabolism and functional predictions for domains of (hitherto) unknown function.* BMC Bioinformatics, 2005. **6**(1): p. 172.
6. Laganeckas, M., M. Margelevičius, and Č. Venclovas, *Identification of new homologs of PD-(D/E) XK nucleases by support vector machines trained on data derived from profile–profile alignments.* Nucleic acids research, 2010: p. gkq958.
7. Steczkiewicz, K., et al., *Sequence, structure and functional diversity of PD-(D/E)XK phosphodiesterase superfamily.* Nucleic Acids Research, 2012. **40**(15): p. 7016-7045.