Analyzing Peptide Libraries with peptider

Eric Riemer Hare

January 6, 2014

1 Introduction

Libraries of peptides, or amino acid sequences, have a number of applications in the Biological sciences, from studying protein interactions, to vaccine research. Despite their importance, little analysis has been done to assess the statistical properties of different peptide libraries.

Peptider is a newly-released R package which helps to evaluate many important statistical properties of these libraries. It supports a number of built-in library schemes, including NNN, NNB, NNK, NNS, and trimer schemes. It also allows for easy analysis of user-created custom library schemes. Peptider makes use of the R package *discreteRV*, which allows for manipulation and analysis of discrete random variables. By treating each amino acid in a peptide as a realization of an independent draw from the pool of all possible amino acids, probabilities for the occurrence of peptides can easily be formulated.

This paper will focus on two distinct functional areas of peptider. The first is Library Diversity, or statistical measures of the quality of the library itself. The second is Peptide Coverage, or how likely the library is to include particularly desired peptides, or peptides that are most similar to desired peptides. Before proceeding to discuss these measures, we will first discuss the built-in library schemes, and how to define custom schemes.

1.1 Library Schemes

Peptider has several built-in library schemes. The first is the NNN scheme, in which all four bases (Adenine, Guanine, Cytosine, and Thymine) can occur at all three positions in a particular codon, and hence there are 64 possible nucleotides. The second is the NNB scheme, where the first two positions are unrestricted, but the third position can only be three bases, yielding 48 nucleotides. Both NNK and NNS have identical statistical properties in this analysis, with the third position restricted to two bases for a total of 32 nucleotides. Finally, there are trimer-based libraries in which the codons are pre-defined. Each of these scheme definitions can be accessed with the *scheme* function.

```
scheme("NNN")
```

```
##   class    aacids c
## 1     A       SLR 6
## 2     B     AGPTV 4
## 3     C         I 3
## 4     D DEFHKNQYC 2
## 5     E        MW 1
## 6     Z         * 3
```

To build a library of an appropriate scheme, the libscheme function is used. By default, peptides of length one amino acid (k = 1) will be used, but this can be specified.

```
nnk6 <- libscheme("NNK", k = 6)
```

libscheme returns a list containing two elements. The first, data, describes the probability of occurrence of each possible peptide class. The second, info, describes the number of nucleotides, the number of valid nucleotides, and the scheme definition used.

We can also create a custom library scheme by building a data frame of the same format as in Code Example 1. This code creates a custom trimer-based library with peptides of length six.

```
custom <- data.frame(class = c("A", "Z"), aacids = c("SLRAGPTVIDEFHKNQYMW",
    "*"), c = c(1, 0))
custom6 <- libscheme(custom, k = 6)
```

Having created the library of interest, we now turn our attention to assessment of these libraries.

2 Library Diversity

In this section, we introduce a number of properties which can be used to determine the quality of a given peptide library, and which are computable using peptider.

2.1 Functional Diversity

The functional diversity of a library is the overall number of different peptides in the library. Analyzing the peptide sequences directly is complex, so a useful approach is to partition the library into amino acid classes, wherein each amino acid belonging to a particular class has the same number of codon representations as all other amino acids in that class. Letting v represent the number of valid amino acid classes, k represent the number of amino acids in each peptide, b_i represent the number of different peptides in class i, N represent the total size of the peptide library in number of peptides, and p_i represent the probability of peptide class i, then the functional diversity is:

$$D(N,k) = \sum_{i=1}^{v^k} b_i(1 - e^{-Np_i/b_i})$$

To compute this diversity measure in peptider, the *makowski* function can be used.

```
makowski(6, "NNK")
```

```
## [1] 0.2918
```