Simulation scheme

1 Sequence data simulation

We generate n sequences $s_1, \ldots, s_n$ of length N based on real frequences of amino acids on full alphabet.

A	L	A	V	P	H	G	K	T	F
S	L	Q	W	E	P	V	L	D	T
R	I	F	N	N	V	Q	G	A	A
G	C	S	D	G	Y	D	Q	T	R
Y	L	R	R	S	R	P	D	A	V
N	V	S	M	M	T	R	G	D	I

Table 1: Example sequences of length 10

2 Motif generation

We generate a set of m motifs $(m_1, \ldots, m_m)$ with the following parameters:

- n_m: denoting maximum number of letters in motif
- d_m: denoting maximum number of gaps in motif

A	_	_	B			
Y	L	_	G	_	_	D
N	_	_	_	_	M	
A	B	_	T			

Table 2: Example motifs

3 Motif injection

We inject motif by replacing a randomly selected part of a sequence with this motif. For example:
We inject from 0 to k motifs to a single sequence according to the following procedure:

	A	L	A	V	P	H	G	K	T	F
	S	L	Q	W	E	P	V	L	D	T
	R	I	F	N	N	V	Q	G	A	A
green!25	G	A	B	D	T	Y	D	Q	T	R
green!25	Y	L	R	R	S	R	A	B	A	T
green!25	N	V	A	B	M	T	R	G	D	I

Table 3: Example sequences of length 10 with addition of motif AB_T

1. Set the ratio r_m of sequences with at least one motif.
2. For each sequence s_i sample a number of motifs to inject as follows:

$$k_i \in \begin{cases} \{1, 2, \ldots, k\}, & 1 < i \leq \lfloor n \cdot r_m \rfloor \\ \{0\}, & \lfloor n \cdot r_m \rfloor < i \leq n \end{cases}$$

3. For s_i sample k_i motifs from set $\{m_1, \ldots, m_m\}$

[width=1]figs/kmers.drawio.pdf

Figure 1: Scheme of sequence data simulation.

4 Target variable sampling

Let's define a random variable X_K on a set of sequences $s_1, \ldots, s_n$ which describes whether a sequence K is a subsequence of s_i. Namely, for any sequence s_i:

$$X_K(s_i) = \begin{cases} 1, & K \subseteq s_i \\ 0, & otherwise \end{cases}$$

where $K \subseteq s_2$ means that K is a subsequence of s_2. For example

$$X_{\texttt{AB}}(\texttt{ABCD}) = 1,$$
$$X_{\texttt{A\_B}}(\texttt{ABCD}) = 0.$$

[width=0.9]figs/models.pdf

Figure 2: Probability of success for the fearure sampling.

4.1 Logistic regression

In this case we consider a standard logistic regression model where the joint effect of all motifs is the sum of their individual effects. Let $w_1, \ldots, w_m$ be weights related to motifs $m_1, \ldots, m_m$. Let w_0 be an effect for the sequences without motifs. Then, we can define an additive logistic model as follows:

$$g(EY) = w_0 + w_1 X_{m_1} + w_2 X_{m_2} + \ldots + w_m X_{m_m}$$

We assume some particular values of $w_1, \ldots, w_m$ and calculate vector of probabilities p as follows:

$$p = g^{-1}(w_0 + w_1 X_{m_1} + w_2 X_{m_2} + \ldots + w_m X_{m_m}).$$

Having p we simulate y_i from binomial distribution $B(1, p_i)$.

4.2 Logistic regression with interactions

Another approach is based on interactions indicating that the effect of one predictor depends on the value of another predictor. Let's define maximum number of motifs per sequence $k = \max\{k_i, i = 1, \ldots, n\}$. Let $w_1, \ldots, w_k$ denote weights of single effects. Namely:

$$g(EY) = w_0 + \sum_{i=1}^{k} w_i X_{m_i} + \left(\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} w_{ij} X_{m_i} X_{m_j} \right) + \ldots + w_{1\ldots k} X_{m_1} \ldots X_{m_k}$$

4.3 Logic regression

Here, we consider new variables, $L_1, \ldots, L_l$ where each of them is a logic expression based on a subset of motifs $m_1, \ldots, m_m$. For example,

$$L_1(m_1, m_2, m_3) = (X_{m_1} \wedge X_{m_2}) \vee X_{m_3}.$$

Each variable L_i obtains its own weight in the model. Our model is following:

$$g(EY) = w_0 + \sum_{i=1}^{l} w_i L_i.$$