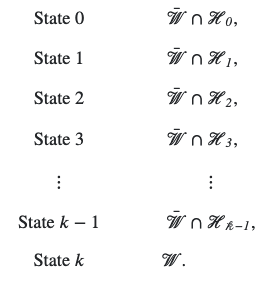
**Answer taken from https://stats.stackexchange.com/questions/26988/probability-of-finding-a-particular-sequence-of-base-pairs**

It is possible to obtain the exact probability of a particular subsequence by using a Markov chain representation of the problem. The specifics of how to construct the chain depend on the particular subsequence of interest, but I will give a couple of examples of how to do this.

**Exact probability via Markov chain:** Consider a discrete sequence of outcomes of *𝐴*,*𝑇*,*𝐶*,*𝐺*

where the outcomes in the sequence are exchangeable, and suppose we are interested in some substring of length *𝑘*. For any given value of *𝑛*, let *𝒲* be the event that the substring of interest occurs, and let *ℋ𝑎* be the event that the last *𝑎* outcomes are the first *𝑎*<*𝑘* characters in the substring of interest (but no more than this). We use these events to give the following partition of *𝑘*+1 possible states of interest:



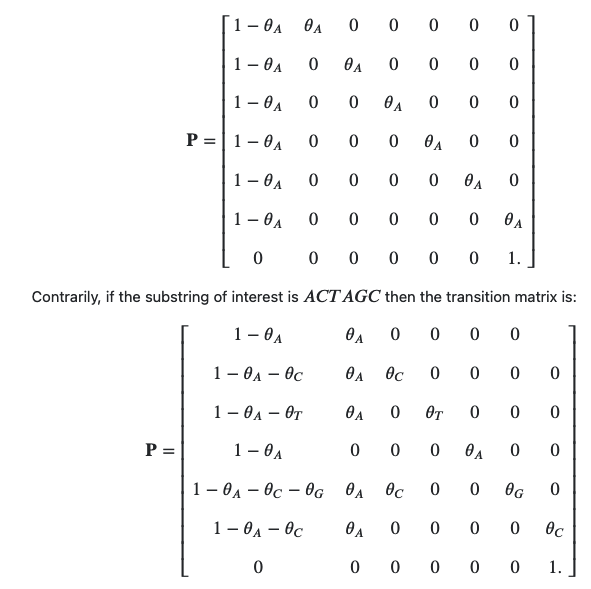
Since the sequence of outcomes is assumed to be exchangeable we have independent outcomes conditional on their respective probabilities

*𝜃𝐴*+*𝜃𝑇*+*𝜃𝐶*+*𝜃𝐺* = 1.

Your process of interest can be represented as a discrete-time Markov chains that begins in State 0 at *𝑛* = 0 and transitions according to a probability matrix that depends on the particular substring of interest. The transition matrix will always be a (*𝑘*+1)×(*𝑘*+1) matrix representing the probabilities of transition using the above states. If the substring of interest has not been reached then each transition can either bring you one step closer to the substring or it can set you back to a previous state that depends on the particular substring. Once the substring is reached, this is an absorbing state of the chain, representing the fact that the event of interest has occurred.

For example, if the substring of interest is *𝐴𝐴𝐴𝐴𝐴𝐴 (or indeed numbers 1..4 referring to quadrants based on intergenic distance)*

then the transition matrix is:



As can be seen above, constructing the transition matrix requires attention to the particular substring. An incorrect outcome sets you back to a previous state in the string that depends on the particular substring of interest. Once the transition matrix is constructed, for a given value of *𝑛* the probability of having the substring in the chain is ℙ(*𝒲*|*𝑛*)={**𝐏***𝑛*}0,*𝑘*. (This probability is zero for all *𝑛*<*𝑘*.)

**Programming this in R:** You can program this as a function in R by creating a function that generates the transition matrix for the Markov chain and an array of its powers up to some desired number of trials. You can then read the appropriate transition probability for the value of *𝑛*

that is of interest. Here is an example of some code to do this:

#Create function to give n-step transition matrix for n = 1...N

#We will use the example of the substring of interest "AAAAAA"

#a is the probability of A

#t is the probability of T

#c is the probability of C

#g is the probability of G

#N is the last value of n

PROB <- function(N,a,t,c,g) { TOT <- a+t+c+g;

a <- a/TOT;

t <- t/TOT;

c <- c/TOT;

g <- g/TOT;

P <- matrix(c(1-a, a, 0, 0, 0, 0, 0,

1-a, 0, a, 0, 0, 0, 0,

1-a, 0, 0, a, 0, 0, 0,

1-a, 0, 0, 0, a, 0, 0,

1-a, 0, 0, 0, 0, a, 0,

1-a, 0, 0, 0, 0, 0, a,

0, 0, 0, 0, 0, 0, 1),

nrow = 7, ncol = 7,

byrow = TRUE);

PPP <- array(0, dim = c(7,7,N));

PPP[,,1] <- P;

for (n in 2:N) { PPP[,,n] <- PPP[,,n-1] %\*% P; }

PPP }

#Calculate probability for N = 100 for equiprobable outcomes

N <- 100;

a <- 1/4;

t <- 1/4;

c <- 1/4;

g <- 1/4;

PROB(N,a,t,c,g)[1,7,N];

[1] 0.01732435