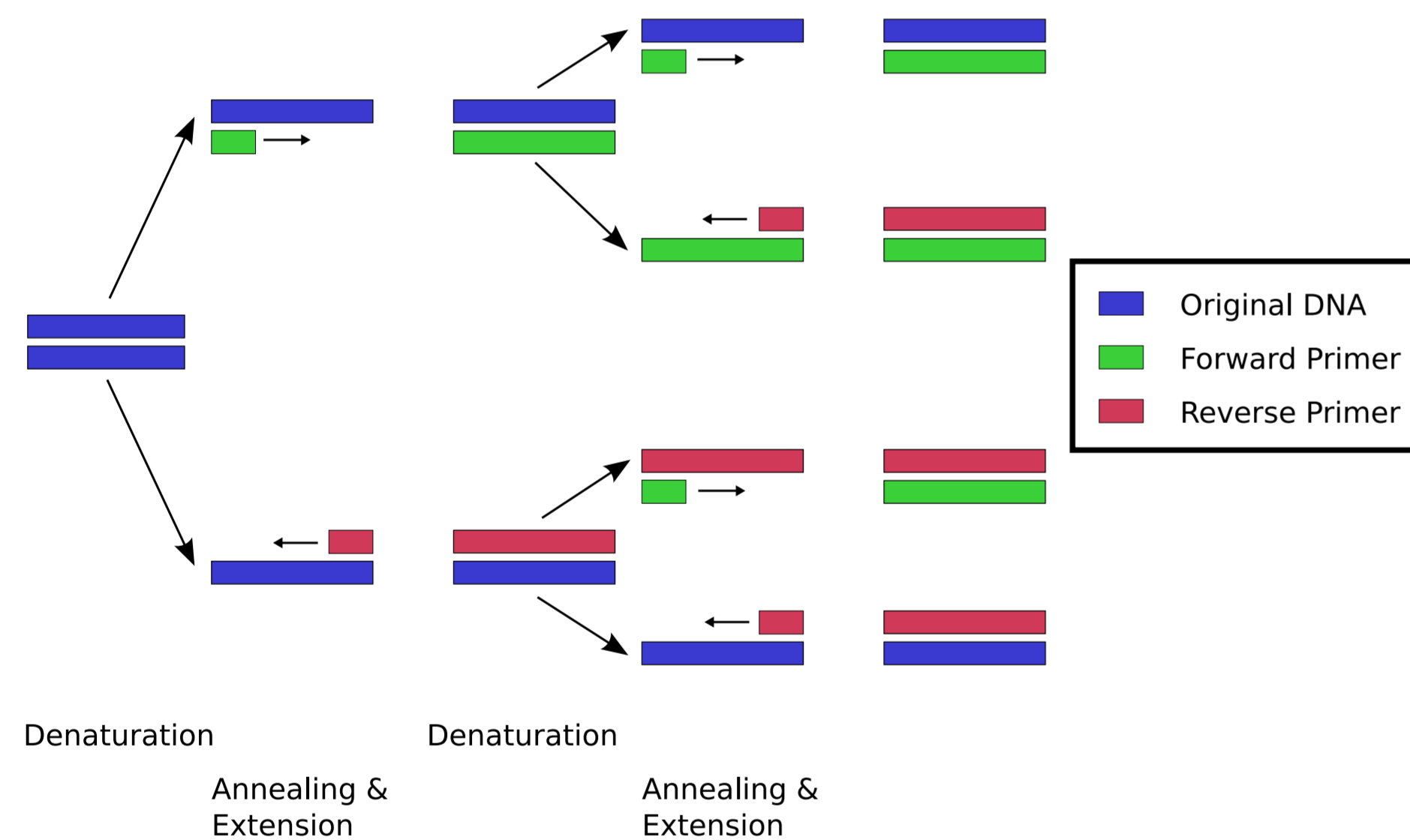


Polymerase chain reaction (PCR) is a technique used to amplify specific regions of DNA and has become a common method in diverse biomolecular applications. Although the primers used in PCR are designed to amplify specific sections of a DNA sequence, it has been shown that amplification of other sections can occur if a close match to the primers occurs around those sections. In most cases this ‘mispriming’ simply causes the main product to be created with reduced efficiency and is not of overriding concern. Recently however, techniques such as quantitative real-time PCR (qRT-PCR) have been developed that can measure the amount of product produced from a PCR reaction. Mispriming in this context could cause a difference in the amount of product produced and hence measured. This work seeks to show a method by which the amount of mispriming in a PCR reaction may be estimated.

Introduction to PCR

A PCR reaction amplifies the amount of a section of DNA by the following method. Two sets of oligonucleotides, known as *primers*, are chemically designed so that they match regions of DNA that bound the section of interest. These primers, purified polymerase, the four deoxyribonucleoside triphosphates and the DNA to be amplified are placed together in a reaction. This mixture is repeatedly heated and cooled for multiple cycles. At each cycle, the heat causes the DNA to separate into two strands. The cooling means that primers can then bind to their respective regions and DNA synthesis occurs. This in effect doubles the amount of DNA. The result is an exponential increase in the amount of the DNA of interest over multiple cycles.



Estimating the Amount of Mispriming

In order to estimate the amount of mispriming in a PCR reaction, two quantities are needed: the rate at which a section of DNA will misprime and the starting amount of that section.

Determining the Mispriming Rate

Let $\rho_{f,r,s,fp,rp}$ be the rate at which primers fp and rp misprime on the sequence of DNA s at locations f and r . To model this probabilistically let

$$\rho_{f,r,s,fp,rp} = 1 + P(MP = (f, r) | SEQ = s, FP = fp, RP = rp, F = f, R = r)$$

This can be expanded in terms of binding events FB and RB , which can take on True/False values. The following probabilities are all conditional on s, fp, rp, f and r

$$P(MP = (f, r)) = P((f, r) | fb, rb) P(fb, rb) \\ = P((f, r) | fb, rb) P(fb) P(rb)$$

This says that the probability of mispriming at locations f and r is determined by three probabilities; the probability of mispriming given that the primers have bound, the probability that the forward primer will bind and the probability that the reverse primer will bind.

Determining the Starting Amount of DNA

Let $a_{f,r,s}$ be the number of pieces of DNA that contain locations f and r on sequence s . This can be calculated as follows. Firstly calculate the probability that f and r are on the same fragment of DNA given a single sequence s . This is given as $p_{f,r,s}$. In the following equation $frag_{f,r}$ means the event that a fragment of DNA sequence s contains locations f and r .

$$p_{f,r,s} = P(frag_{f,r} | f, r, s) \\ = \int_{r-f+1}^s P(frag_{f,r} | f, r, s, l) P(l | f, r, s) dl \\ = \int_{r-f+1}^s P(frag_{f,r} | f, r, s, l) P(l | s) dl$$

$P(l | s)$ is the prior probability that a fragment of sequence s is of length l . $P(frag_{f,r} | f, r, l, s)$ is the probability that a fragment of length l on sequence s contains positions f and r . This can be given as

$$P(frag_{f,r} | f, r, l, s) = \begin{cases} \frac{f+|s|-r-|l-r|-|f+l-|s|-l|}{2(|s|+1-l)} & , \text{if } f < r, r - f + 1 < l < |s| \\ 0 & , \text{otherwise} \end{cases}$$

If we integrate against the uniform distribution on $(0, |s|)$ we get

$$p_{f,r,s} = \frac{1}{|s|} ((|s| + f - r + 1) \log(|s| + f - r + 1) \\ - (|s| + 1 - r) \log(|s| + 1 - r) \\ - f \log f)$$

Given a sequence of n_s DNA sequences, each of which has a probability $p_{f,r,s}$ of containing f and r on the same fragment, we get a binomial distribution for the number of fragments $N_{f,r,s}$ containing f and r . To estimate the number n_s of a particular DNA sequence, we can use the ratio of the number of bases to weight ζ_s . Therefore with w_s of extra material in the mixture there are $w_s \zeta_s$ bases. We then divide by the length of s to get

$$n_s = \frac{w_s \zeta_s}{|s|}$$

of each sequence present in the reaction. With $N_{f,r,s} \sim B(n, p_{f,r,s})$, $a_{f,r,s}$ can be estimated as the expected value of this distribution, i.e. $a_{f,r,s} = n_s p_{f,r,s}$. With the amount of starting material $a_{f,r,s}$ and the mispriming rate $\rho_{f,r,s,fp,rp}$, the amount of material misprimed at cycle x is given by $a_{f,r,s} \rho_{f,r,s,fp,rp}^x$. Because this could happen at multiple locations, we need to sum over all possible primers, all possible sequences and all locations on those sequences. The total amount of misprimed sequences at PCR cycle x is

$$m_x = \sum_{fp, rp \in PRIM} \sum_{s \in SEQ} \sum_{f, r \in s} a_{f,r,s} \rho_{f,r,s,fp,rp}^x$$

where $PRIM$ is the set of primers and other oligos in our reaction.

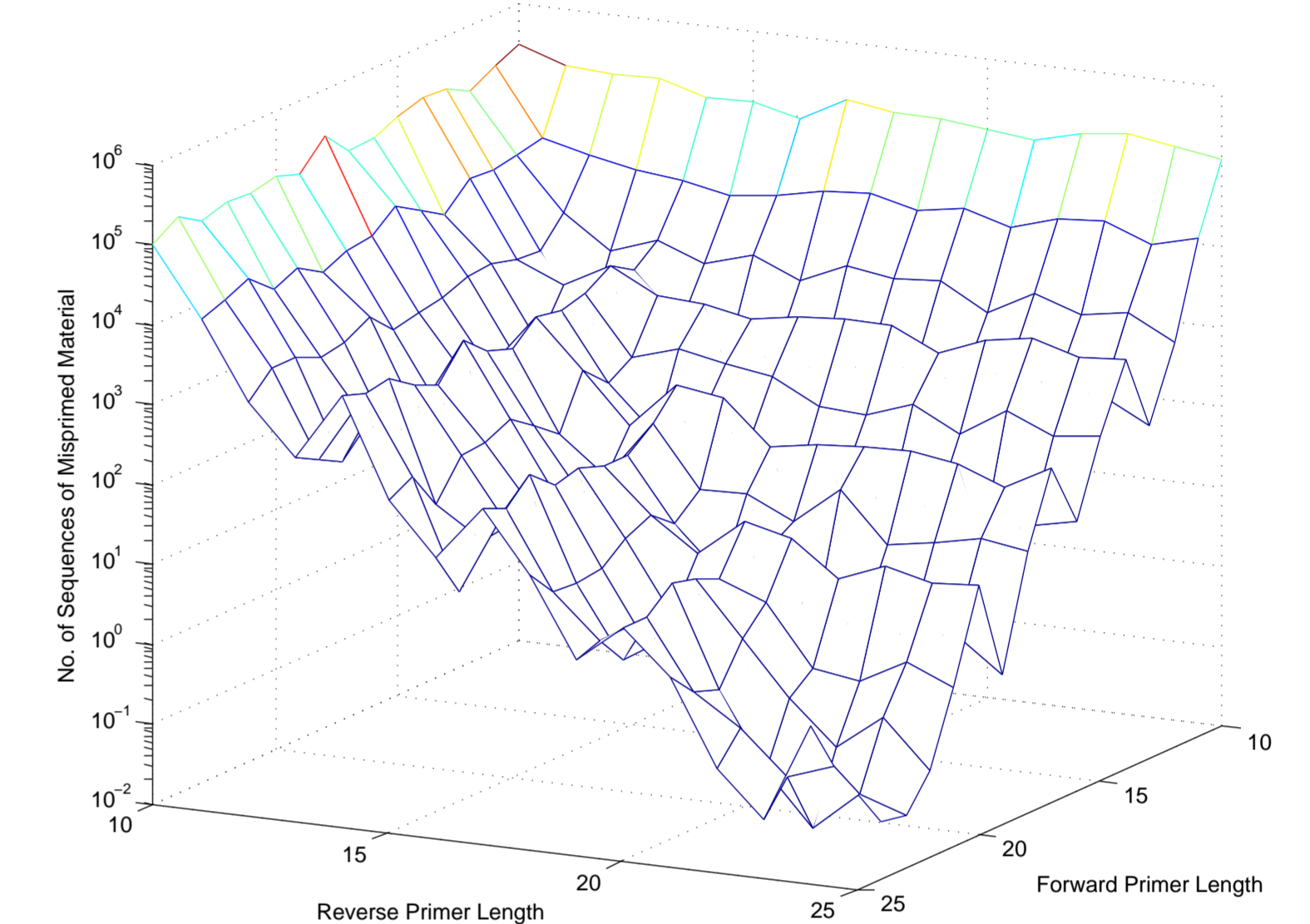
Experiments Using Proposed Model

For these experiments, random primers were created to see how they misprimed in a typical PCR reaction. Certain characteristics of the primers were modified in order to see the effect this would have on the amount of mispriming. The mispriming base was taken as the mRNA transcripts of *Equus Caballus*. The probabilities from the model were fully specified as follows. The probability of a binding event ($P(fb)$ or $P(rb)$) came from a heuristic rule often used by biologists. This probability was given as 1 if at least 70% of the bases matched and at least 3 of the

last 5 bases at the 3' end matched and 0 otherwise. The probability of mispriming given binding ($P((f, r) | fb, rb)$) was given as 1 if f and r were up to 3000 bases apart and 0 otherwise. To estimate the amount of starting material, $a_{f,r,s}$ the weight of each sequence was assumed to be uniformly distributed with the same base/weight ratio ζ_s of $330 \times 10^{27} \text{b} \cdot \text{kg}^{-1}$. The results are given in terms of an arbitrary weight of $w = \frac{10^{-27}}{330} \text{kg}$ and are calculated at PCR cycle $x = 10$.

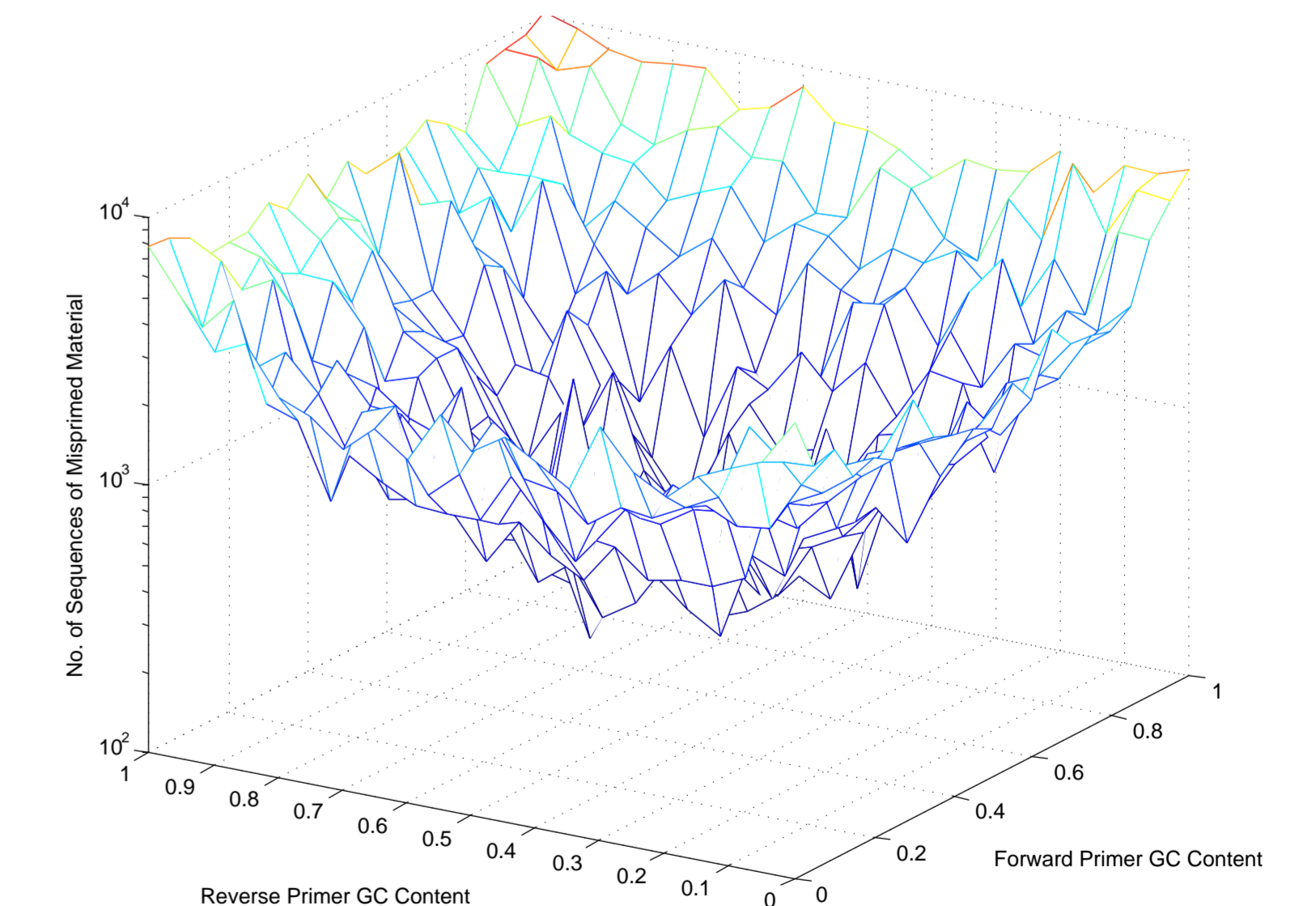
Experiment 1

This shows the mispriming effect as the lengths of the primers are modified between 10 and 25.



Experiment 2

This shows the mispriming effect as the GC content of the primers of length 15 are modified between 0% and 100%



Acknowledgements

This work is supported by an EPSRC grant for the ‘Molecular Nose’ project (EP/E032745/1). Much thanks goes to Drs Meesbah Jiwaji of the Integrative and Systems Biology Research Theme and Gráinne Barkess of the Department of Pathology and Gene Regulation at the University of Glasgow.