Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Theoretical Biology 234 (2005) 497–509
Journal of Theoretical Biology
www.elsevier.com/locate/yjtbi

A general model of error-prone PCR

Leighton Pritchard[a,*], Dave Corne[b,1], Douglas Kell[a,2], Jem Rowland[c], Mike Winson[a]

[a]*Institute of Biological Sciences, University of Wales, Aberystwyth, Ceredigion, SY23 3DD, Wales, UK*
[b]*School of Systems Engineering, University of Reading, P.O. Box 225, Whiteknights, Reading, Berkshire RG6 6AY, UK*
[c]*Department of Computer Science, University of Wales, Aberystwyth, Ceredigion, SY23 3DB, UK*

Received 16 June 2004; received in revised form 24 November 2004; accepted 7 December 2004
Available online 29 January 2005

Abstract

In this paper, we generalise a previously-described model of the error-prone polymerase chain reaction (PCR) reaction to conditions of arbitrarily variable amplification efficiency and initial population size. Generalisation of the model to these conditions improves the correspondence to observed and expected behaviours of PCR, and restricts the extent to which the model may explore sequence space for a prescribed set of parameters. Error-prone PCR in realistic reaction conditions is predicted to be less effective at generating grossly divergent sequences than the original model. The estimate of mutation rate per cycle by sampling sequences from an in vitro PCR experiment is correspondingly affected by the choice of model and parameters.

Keywords: PCR; Polymerase chain reaction; Error-prone; Model

1. Introduction and background

The polymerase chain reaction (PCR) is an in vitro method capable of producing large amounts of identical copies of a gene or other specified nucleotide sequence from a small amount of DNA (Erlich, 1989; Mullis and Faloona, 1987; Sun, 1995). PCR is essentially a three-stage process that amplifies the region of DNA lying between two short stretches of sequence that complement *primer* sequences (Newton and Graham, 1997). In this process double stranded DNA template is separated into two single strands by heating (*denaturing*), then the reaction is cooled so that primers complementing the short stretches of DNA either side of the region of interest may bind (*annealing*). DNA polymerase then uses the template sequence to extend the bound primer (*extension*). A simple formula for the final number of copies of the target double-stranded sequence under ideal conditions is known for numbers of cycles greater than two: $(2^n - 2n)x$ where n is the number of cycles, $2n$ represents products after the first and second cycles that have undefined length due to having bound only one primer, and x is the number of copies of the original template (Newton and Graham, 1997).

Initially, PCR found use mostly for the accurate amplification of known DNA sequences, but PCR-based gene manipulation has become invaluable for the alteration of genetic information at the molecular level, permitting site-directed mutagenesis of PCR products at

Abbreviations: The following abbreviations are used in this paper: PCR—polymerase chain reaction; DNA—deoxyribose nucleic acid; MDM—mutation data matrix; MMo—Moore and Maranas' original model of error-prone PCR; MMc—Moore and Maranas' model generalised to an arbitrary number of template sequences, but with constant amplification efficiency; MMv—Moore and Maranas' model generalised to an arbitrary number of template sequences, with variable amplification efficiency.

*Corresponding author. Present address: Scottish Crop Research Institute, Invergowrie, Dundee, DD2 5DA, UK. Tel.: +44 1382 562731 ext. 2405; fax: +44 1382 568578.
E-mail addresses: l.pritchard@scri.sari.ac.uk (L. Pritchard), d.w.corne@ex.ac.uk (D. Corne), dbk@umist.ac.uk (D. Kell), jjr@aber.ac.uk (J. Rowland), mkw@aber.ac.uk (M. Winson).
[1]Present address: Department of Computer Science, Harrison Building, University of Exeter, Exeter, EX4 4QF, UK.
[2]Present address: Department of Chemistry, Faraday Building, UMIST, PO Box 88, Sackville Street, Manchester M60 1QD, UK.

0022-5193/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2004.12.005

an arbitrary distance from the ends of the template sequence (Barik, 1998; Tait and Horton, 1998). One such general mechanism for inducing randomised nucleotide sequences via PCR is the use of mutagenic (overhanging) primers. This form of mutagenic PCR also found a crucial role in in vitro selection methods (a protocol in use before PCR itself was used to generate the initial, randomised libraries) (Nieuwlandt, 1998). With in vitro selection a random oligonucleotide library would be expressed and screened; high-performing variants from this library would then be retained and accurately amplified, and the expression/screening cycle would begin again, in analogy with Darwinian selection.

Error-prone PCR introduces random copying errors by imposing imperfect, and thus mutagenic, or 'sloppy', reaction conditions (e.g. by adding Mn^{2+} or Mg^{2+} to the reaction mixture (Cadwell and Joyce, 1991; Leung et al., 1989)). This method has proven useful both for generation of randomised libraries of nucleotide sequences, and also for the introduction of mutations during the expression and screening process in a *mutagenesis step*.

While early directed evolution experiments were largely based on selection-amplification procedures, more recent protocols incorporate gene shuffling as the major mutagenic process. The essential procedure of stepwise screening, where the mutant library is screened for the desired function and either the fittest protein or some recombinant of several of the fittest mutants becomes the parent for the next generation, remains the same (Moore et al., 1997; Voigt et al., 2001; Zhao and Arnold, 1997). Although the emphasis of more recent mutagenesis methods is on shuffling, point mutations remain likely to occur even when not purposefully induced (Sun, 1995). In early studies of PCR, substitution, addition and deletion of bases by DNA polymerase was noted and several models of the substitution process have been proposed to describe this. We briefly summarise some of these approaches below.

The theory of Galton–Watson processes has been employed to determine the theoretical proportion of DNA sequences with no mutations after n PCR cycles, assuming perfect amplification efficiency, and a constant mutation rate μ (Krawczak et al., 1989). This and some other models, e.g. Hayashi (1990) however only report the fraction of perfect copies, and give no detailed information about any mutations that occur, such as the nature of substitutions or their locations.

PCR is represented as a bifurcating tree in the model of Weiss and von Haeseler (1995), and the distribution of duplication events between random pairs as a function of amplification efficiency, number of templates, mutation rate and number of PCR cycles is found. An advance seen in this model is its examination of the influence of variable amplification efficiency. The substitution rate was assumed to be uniform and base-independent, however.

In Sun (1995) the author also represents PCR as a branching process, studying the number of mutations and the distribution of the Hamming distance between randomly-selected pairs of product sequences after n PCR cycles for arbitrary mutation rate and amplification efficiency. This work is a generalisation of the models proposed by Hayashi (1990), Krawczak et al. (1989) and Maruyama (1990). However, none of these models take base-dependent rates of substitution into account.

We take as our starting point in this paper the model of error-prone PCR proposed in Moore and Maranas (2000), with the aim of extending it to incorporate arbitrary template population size and amplification efficiency. This model appears to have been constructed with the particular aim of calculating the probability that a given nucleotide sequence will be generated in a single error-prone PCR experiment, and has a different structure to the models described above, being built around the recursive application of a mutation data matrix of base substitution probabilities, thereby modelling a base-dependent mutation rate. This base-dependence is an improvement over, e.g. the models described in Eckert and Kunkel (1991) and Sun (1995). The original results of Moore and Maranas (2000) were only generalised for a total number N of PCR cycles, but not for either variable amplification efficiency or variable template population. Both of these may be important factors for mutagenesis by 'sloppy' PCR, as they would be expected to affect the overall, observed mutation rate for the experiment.

One disadvantage of the recursive MDM approach in comparison with models based on branching processes is that it is not directional in terms of sequence space. Each successive sequence that is sampled from the virtual PCR pot is a random sampling from the entire sequence space potentially reachable by the experiment, rather than a sampling of the sequence space described by a single genealogy of sequences rooted at the template. The models described in Moore and Maranas (2000) and extensions in this paper therefore cannot provide useful information on the sequence space covered by a single PCR experiment, in terms of Hamming distance between randomly-selected pairs of sequences from one PCR pot, unlike the model of Sun (1995). The models do however provide a useful estimate of the probability, taken over many experiments, of obtaining a specified sequence given a particular template sequence.

In this paper, we generalise the model proposed in Moore and Maranas (2000) to an arbitrary number of initial template sequences (PCR rarely actually begins from only two strands). Additionally, since amplification efficiency for any particular cycle may be dependent on the population size at the time of that cycle, or some other factor, we generalise the model to account for this

also. That the population size in any cycle of a PCR experiment is dependent on the initial population size, and may affect the progress of error-prone PCR was suggested in the original paper (Moore and Maranas, 2000), and is described more formally herein. We consider that some plausible causes of variation in amplification efficiency with population size/number of cycles may be:

- As population size increases, the number of free nucleotides relative to the number of sequences falls. The probability of a complete extension of the sequence in later cycles may be reduced, as a result
- As the experiment proceeds, the proportion of active polymerase enzyme falls, relative to the number of sequences present, potentially reducing the efficiency of later cycles
- As the experiment proceeds, the accumulated degradative effect of thermal cycling may reduce the efficiency of the *Taq* polymerase.

The estimate of per-cycle mutation rate in an in vitro PCR experiment is often made by random sampling of sequences from the final 'pot'. The distribution of the mutation frequency per sequence is expected to vary under the effects listed above, and this in turn affects the estimate of per-cycle mutation rate. We describe the effects of model and parameter choice on this estimate.

2. Model

The protocol of the model proposed by Moore and Maranas (2000) may be summarised as follows:

1. An initial mutation data matrix is defined representing nucleotide-dependent substitutions for a single cycle of the PCR experiment:

$$M = \begin{pmatrix} M_{AA} & M_{AC} & M_{AG} & M_{AT} \\ M_{CA} & M_{CC} & M_{CG} & M_{CT} \\ M_{GA} & M_{GC} & M_{GG} & M_{GT} \\ M_{TA} & M_{TC} & M_{TG} & M_{TT} \end{pmatrix}. \quad (1)$$

   Here, M_{ij} represents the probability that nucleotide i will be substituted by nucleotide j in a single PCR cycle.
2. Mutation data matrices for successive cycles of the PCR experiment are generated (this is assumed to be independent of other factors, such as the uneven depletion of nucleotides that may have a large effect in real PCR).

$$C_{ij}^{n} = \begin{cases} \delta_{ij}, & n = 0, \\ M_{ij}, & n = 1, \\ \sum_{k=A,C,G,T} C_{ik}^{n-1} \cdot M_{kj}, & n \geqslant 2, \end{cases} \quad (2)$$

   where δ_{ij} is the Kronecker delta.
3. The number of sequences produced at each cycle n of a PCR experiment that runs for a total of N cycles is calculated as $Z_{N,n}$. These values will usually be dependent on the amplification efficiency, which itself may be variable and dependent on the initial population size. For an arbitrary but constant amplification efficiency λ, and arbitrary initial number of template sequences S_0, after N total PCR cycles the number of sequences present that are the result of n extension steps is given by

$$Z_{N,n} = \binom{N}{n} S_0 \lambda^n \quad (3)$$

   as suggested in Moore and Maranas (2000) (proof in Appendix B). The total number of sequences S_N at the end of the experiment is given by (see Appendix A)

$$S_N = S_0(1+\lambda)^N. \quad (4)$$

   For an arbitrary amplification efficiency that is also dependent on the current PCR cycle, we performed the recursive procedure explicitly. The analytical solution of the total number of sequences at the end of the experiment S_N is given by (Appendix C)

$$S_N = S_0 \prod_{k=0}^{N} (1+\lambda_k). \quad (5)$$

   The total number of sequences generated by an in silico experiment as described here, and the number of sequences generated after n cycles of such an experiment, are determined by these equations. The only influence of stochasticity on these models is derived entirely from the substitution process. The assumption is made throughout these models that mutations at different locations along the sequence are independent. This is probably true if, as is the case with in vitro selection, there is no selective pressure between rounds of screening. As such, a sum of the product of the mutation data matrix for each generation and the proportion of all sequences in the final PCR pot contributed by that generation will produce an aggregate mutation data matrix representing the probabilities of any base having mutated by cycle N, represented by P_{ij}^{N}.

$$P_{ij}^{N} = \frac{1}{S_N} \sum_{n=0}^{N} Z_{N,n} C_{ij}^{n}. \quad (6)$$

4. The application of this aggregate mutation data matrix to each base in a given template sequence

generates a sequence that is equivalent to the random selection of a single sequence from the final PCR pot.

For simulations in which the amplification efficiency is not constant throughout the reaction, the variation may take one of a range of forms. There may be a linear decrease in efficiency with the number of completed cycles (analogous to the progressive denaturing of polymerase) or a linear decrease with total population (analogous to depletion of the relative concentration of nucleotides, polymerase or primer etc. with respect to template). Alternatively, both effects may be seen simultaneously, or there may be some combination of unconsidered nonlinear effects. For the simulation of variable amplification efficiency described in this paper we implement a numerical solution, describing only a linear decrease in amplification efficiency with increasing total population, bounded by the initial template population at the initial amplification efficiency and an upper limit on the population size at amplification efficiency $= 0$. This makes the assumption that the dominant effect on amplification efficiency will be expressed through the depletion of reagents for sequence extension. Selecting only one dominant effect allows for comparison between the original Moore and Maranas (2000) model and generalised variants incorporating terms for variable amplification efficiency without introducing further issues of parameterizing the balance between polymerase degradation and dilution effects.

3. Results and discussion

We implemented the Moore and Maranas (2000) model of PCR, and the variants described above in a computer program written in the Python programming language by one of the authors (LP). The three variants are named, and differ from one another, as described below:

1. *MMo*: The model of PCR as described in Moore and Maranas (2000). The model assumes perfect amplification efficiency ($\lambda = 1$) and initially only two single strands of template sequence.
2. *MMc*: Generalisation of the MMo model to an arbitrary number of initial template sequences, and possibly imperfect but constant amplification efficiency ($\lambda \leqslant 1$)
3. *MMv*: Generalisation of the MMo model to an arbitrary number of initial template sequences, and a variable amplification efficiency that decreases linearly as the total sequence population size grows to a predefined limit.

3.1. Variation of the final number of sequences with models

For in vitro PCR experiments, the sequence population does not increase indefinitely in an exponential manner, but often reaches an upper limit due to constraints of polymerase, primer and nucleotide availability, with secondary effects from polymerase degradation. Figs. 1a–d describe the growth of total population of sequences in the PCR reaction with number of cycles for the three model variants. In the models, these results are entirely deterministic and are found by solving Eqs. (4) and (5) for the appropriate parameters. For in vitro PCR, this idealism is probably not appropriate, and there is likely to be stochastic variation that deviates from these models.

As can be seen from Fig. 1a, the original model described in Moore and Maranas (2000) results in exponential growth of the number of sequences and, as implied by the model, this growth remains unchecked. This is an unrealistic model of population increase in real PCR experiments, as population growth is eventually limited, possibly by the factors detailed above.

Fig. 1b demonstrates that generalisation of the *MMo* model to arbitrary initial population and arbitrary amplification efficiency does not result in a pattern of population growth that is any more representative of a real PCR experiment. As in the *MMo* model and Fig. 1a, population growth is permitted to continue unchecked. The introduction of a constant, imperfect amplification efficiency ($\lambda < 1$) is therefore not sufficient to describe the commonly-observed variation with cycle-number of total population in an in vitro PCR experiment.

In Fig. 1c, population growth for the MMv model with variable amplification efficiency is described. The introduction of variable amplification efficiency to the model causes a variation of the population curve from those of the MMo and MMc models, where S_0 and λ_0 are large enough for the total population size to reach its limit. A relationship in which amplification efficiency increases with cycle number would result in even faster exponential growth than the MMo model. It is presumed that in real PCR experiments, amplification efficiency decreases with successive PCR cycles. To reflect this effect in the current study, amplification efficiency is set to decline linearly to zero at the population limit, with increasing total population. From Fig. 1c, it is clear that, until the total population approaches an upper limit, growth of the sequence population proceeds exponentially as in the MMc model. Only at higher initial amplification efficiencies does the population size approach the limit, causing the plot to differ from that of the MMo model (Fig. 1a).

As is expected, increasing the initial population in the MMv model causes the population of the PCR

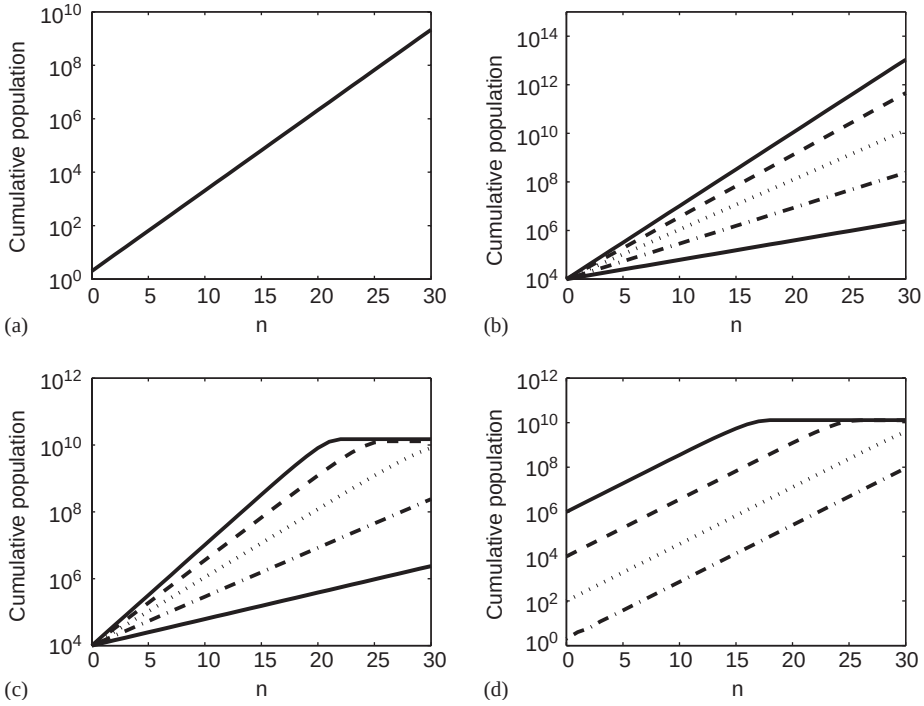

Fig. 1. Total population size S_n against cycle number n for (a) MMo, the PCR model described in Moore and Maranas (2000); (b) the MMc model with amplification efficiencies $\lambda = 1$ (solid line), $\lambda = 0.8$ (dashed line), $\lambda = 0.6$ (dotted line), $\lambda = 0.4$ (dash-dot line) and $\lambda = 0.2$ (thick line), and an initial template population size $S_0 = 10^4$; (c) the MMv model with variable initial amplification efficiencies of $\lambda_0 = 1$ (solid line), $\lambda_0 = 0.8$ (dashed line), $\lambda_0 = 0.6$ (dotted line), $\lambda_0 = 0.4$ (dash-dot line) and $\lambda_0 = 0.2$ (thick line), an initial template population size of $S_0 = 10^4$, and population size limit of $S_{limit} = 10^{10}$, and (d) the MMv model described above with initial amplification efficiency $\lambda_0 = 0.8$, total population size limit $S_{limit} = 10^{10}$ and initial population size $S_0 = 1e6$ (solid line), $S_0 = 1e4$ (dashed line), $S_0 = 1e2$ (dotted line), and $S_0 = 2$ (dash-dot line).

experiment to reach the imposed upper limit after fewer cycles, and this effect can be seen in Fig. 1d, where for an initial amplification efficiency of 0.8, only the smaller ($<10^4$) initial populations fail to reach the upper limit of population. On a linear scale, the population profile is approximately sigmoidal, as routinely observed in RT-PCR experiments.

3.2. Variation of $Z_{N,n}$ by model

Error-prone PCR may be used deliberately to generate random mutations, but the potential for generating random mutations is always present in any PCR experiment. In either case it is useful to be able to predict and understand the distribution and frequency of occurrence of random mutations. When deliberately generating random mutations to explore sequence space it is helpful to know the size of sequence space that may potentially be explored by the generated mutations; estimates of mutation rate per cycle and an understanding of how to maximize overall mutation rate per sequence are important to this aim. In the case of amplifying or proof-reading PCR, where mutations are undesirable, it is useful to know how many random mutations are likely to occur, and how best to minimize them. Again, an estimate of the mutation rate per cycle and an understanding of how to minimise overall mutation rate will be important. If the mutation rate is constant for each generation, it is expected that the number of mutations observed in a single sequence extracted from the final pool will be proportional to the number of extension steps of which it is a product. What is important is how many sequences from each generation are present in the final pool. This is represented by $Z_{N,n}$, which denotes how many sequences present in the pot after a total of N cycles are the result of n extension steps. Control of the modal number of extension steps may therefore be an important mechanism for determining overall mutation rate.

For example, we might consider two hypothetical PCR experiments with the same total number of cycles, MDM and the same final total number of sequences. In one (perhaps unlikely) experiment, amplification efficiency begins at a low value and increases towards the end of the experiment; $Z_{N,n}$ is then greatest at larger n ('late-peaking'). In the other experiment, amplification efficiency is initially high, but drops very rapidly; $Z_{N,n}$ is greatest at smaller n ('early-peaking').The greatest contribution to the final sequence set in the latter, 'early-peaking' experiment, comes from sequences that have not been through many cycles and so do not carry many mutations. The 'late-peaking' experiment by

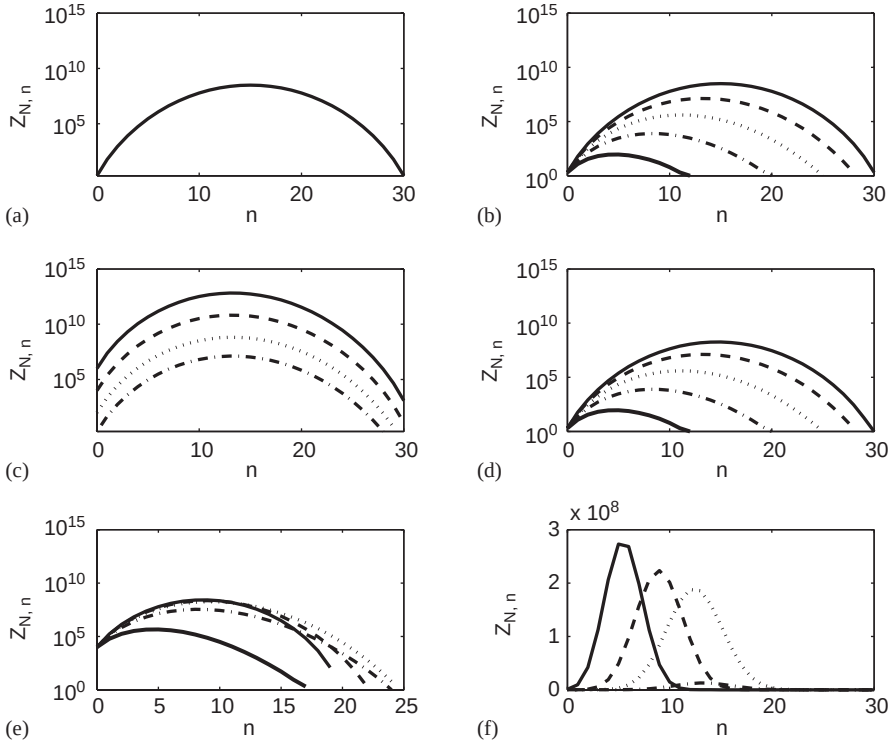

Fig. 2. Number of sequences $Z_{N,n}$ resulting from n extension steps after $N = 30$ cycles, against number of extension steps n, for (a) the MMo model; (b) the MMc model with initial population size $S_0 = 2$, and amplification efficiencies $\lambda = 1$ (solid line), $\lambda = 0.8$ (dashed line), $\lambda = 0.6$ (dotted line), $\lambda = 0.4$ (dash-dot line) and $\lambda = 0.2$ (thick line); (c) the MMc model, with initial amplification efficiency $\lambda = 0.8$, and initial population sizes $S_0 = 10^6$ (solid line), $S_0 = 10^4$ (dashed line), $S_0 = 10^2$ (dotted line) and $S_0 = 2$ (dash-dot line); (d) the MMv model with an upper population limit $S_{limit} = 10^9$, initial population $S_0 = 2$, and initial amplification efficiencies $\lambda_0 = 1$ (solid line), $\lambda_0 = 0.8$ (dashed line), $\lambda_0 = 0.6$, (dotted line), $\lambda_0 = 0.4$ (dash-dot line) and $\lambda_0 = 0.2$ (thick line); (e) the MMv model with an upper population limit $S_{limit} = 10^9$, initial population $S_0 = 10^4$, and initial amplification efficiencies of $\lambda_0 = 1$ (solid line), $\lambda_0 = 0.8$ (dashed line), $\lambda_0 = 0.6$, (dotted line), $\lambda_0 = 0.4$ (dash-dot line) and $\lambda_0 = 0.2$ (thick line), and (f) the MMv model with an upper population limit of $S_{limit} = 10^9$, initial amplification efficiency $\lambda_0 = 0.8$ and initial population sizes $S_0 = 10^6$ (solid line), $S_0 = 10^4$ (dashed line), $S_0 = 10^2$ (dotted line) and $S_0 = 2$ (dash-dot line).

contrast will contain mostly sequences that are the accumulated result of several mutation cycles, and would be expected to carry a greater mutational load. As a result, the extent of sequence space sampled by several 'early-peaking' experiments may be expected to be less than that sampled by several relatively 'late-peaking' experiments.

In this section, we examine distributions of $Z_{N,n}$ for the three models, MMo, MMc and MMv, where the total number of PCR cycles is 30. In all three model types, $Z_{N,n}$ is deterministic, following from Eqs. (3) and (5), which were used to generate the plots in Fig. 2. We again note that the distribution of $Z_{N,n}$ may be influenced by stochastic effects in the case of in vitro PCR.

$Z_{N,n}$ for the MMo model as defined in Moore and Maranas (2000), where $N = 30$, describes a symmetrical distribution, with most sequences in the final pot being the result of 15 or 16 extension steps (Fig. 2a).

The MMc model (Fig. 2b) introduces an arbitrary, but constant, amplification efficiency. As amplification efficiency falls from $\lambda = 1$, the total number of sequences produced also declines, and so does the modal number of extension steps (and therefore overall mutation rate per sequence) for each sequence in the final pot.

As there is no upper limit on the total number of sequences in the MMc model, it would be expected that if the initial template population was to be increased for constant λ, no shift in the modal value of n for the sequences would be seen, but that the total number of sequences that are the result of n extension steps would increase. The overall observed mutation rate would therefore be expected to be more dependent on λ than on S_0. This is demonstrated in the plot of $Z_{N,n}$ against n in Fig. 2c, and in Fig. 4b.

The gross effect of varying initial amplification efficiency for the MMv model (Fig. 2d) is much the same as that seen with the MMc model (Fig. 2b) for an initial population size $S_0 = 2$. The effect of varying initial population size, because it causes the rate at which λ varies to change in turn, is significant and marked (Figs. 2e and f). Fig. 2e has a similar profile of $Z_{N,n}$ scores by cycle to 2d. With an initial population size $S_0 = 10^4$, varying the initial amplification efficiency (as shown in Fig. 2e) exerts a similar effect by reducing the modal number of extension steps for each of the final

sequences as in the MMc model (Fig. 2d). The modal number of extensions (and to an extent, overall mutation rate) however, does not begin to fall until λ_0 is greatly reduced. Hence, in contrast to the unconstrained MMc model, the MMv model is seen here to be practically insensitive to changes in λ_0 once the upper population limit is reached (this upper limit is reached for $\lambda_0 \geqslant 0.6$).

When the initial population size is increased from $S_0 = 2$ to $S_0 = 100$ (Fig. 2f) for $\lambda_0 = 0.8$, the upper population limit $S_{limit} = 10^9$ is not reached, the modal number of extension steps per sequence in the final pool varies little, and the overall number of sequences that are the result of a given number of extension steps rises much less rapidly than for the MMc model. However, as the initial template population increases more significantly, the population size limit is reached at a smaller number of cycles, so λ_n falls to zero, and the modal number of extension steps (and implicitly mutation rate) falls (Fig. 2f), in contrast to the MMc model. The strategy of varying mutation rate by varying initial template population size has been used in commercial mutagenic PCR kits (e.g. Arezi et al., 2002). These models (Figs. 2c and f) imply that these methods are viable because there is a cycle-wise reduction in amplification efficiency and/or some constraint on total population size. It thus appears that the generalised MMc and MMv models imply a less broad search of sequence space for in vitro PCR than does the MMo model.

3.3. Variation in probability of generating a given sequence by model

As seen above, the inclusion of variable amplification efficiencies and initial template population sizes implies a notable effect of these factors on the observed number of substitutions per sequence (overall mutation rate) compared to the original Moore and Maranas (2000) model. In the MMo model, the probability of assembling a sequence S through successive single point mutations on an original sequence S^0 after N PCR cycles is given by

$$\Pi^N_{S^0,S} = \frac{1}{S_N}\sum_{n=0}^{N} Z_{N,n}\prod_{j=1}^{B}[P^n]_{s^0_j,s_j}, \tag{7}$$

where B is the length of the two sequences and s^0_j and s_j are the nucleotides at position j for sequences S^0 and S. This provides a quantitative means of estimating a priori the fraction of sequences in the final pool that conform to a target sequence S given the mutation data matrix M and initial sequence S^0. For the generation of functional nucleotide sequences this is adequate, but for the generation of functional proteins issues of codon redundancy must be taken into account with a set of target sequences $S \in \{T_1, T_2, \ldots\}$ where T_n are nucleotide sequences that encode a desired protein target.

The above description also does not account precisely for functional targeting. While one may have a good idea of which residues in a protein sequence one would wish to adapt for the development of improved or novel function, this function may also be achieved despite (or even because of) substitutions at unexpected locations (e.g. Zaccolo and Gherardi, 1999; Oue et al., 1999). Also, the final pool of sequences generated by an in vitro PCR reaction is strongly dependent on the history of that experiment. A single PCR experiment will typically sample a *hyper-cone* of the theoretically accessible sequence space, whereas the equation describing the probability of assembling a prescribed target sequence by the model described herein implies a random sampling of the whole, theoretically accessible space.

Nevertheless, this probability is a useful metric for estimating of the extent of sequence space that can be explored over many experiments given initial parameters of the model (assuming that each individual experiment generates a 'branch' of sequences independent of the other experiments in the set). As the model assumes no branched evolution-like directionality in sequence evolution, its theoretically-accessible sequence space will approximate a B-dimensional 'Hamming sphere', or *hypercube*. Each point within this 'sphere' represents a possible target sequence, and the probabilities of the model generating the prescribed target sequence at any particular point within this space are not uniformly distributed, due to the base-dependent substitution probabilities of the MDM. Thus, the 'Hamming sphere' is distorted and, in particular, not all sequences of equal Hamming distance from the original template S^0 are equally likely after a total of N PCR cycles. The uneven coverage of this space is expected to form a shell-like hypercube of high probability sequences and a tailing off of probabilities towards the template sequence (centre of the space) and towards sequences with B substitutions, reflecting, in section, the frequency profiles of Figs. 2a–f.

In order to estimate the coverage of this sequence space, we used the MMo, MMc and MMv models described above to estimate the mean probability of generating 31 target nucleotide sequences containing from 0 to 30 random base substitutions relative to a template sequence. One hundred template sequences were chosen randomly from the complete set of coding sequences in the fully-sequenced bacterial genomes available at October 2004 from ftp.ncbi.nih.gov. Ten in silico PCR experiments were performed for each template using the MMo model, the MMc model ($\lambda_0 = 0.8, S_0 = 2$), and the MMv model with three sets of parameters: ($\lambda_0 = 0.8, S_0 = 2, S_{limit} = 10^8$), ($\lambda_0 = 0.8, S_0 = 10^6, S_{limit} = 10^8$) and ($\lambda_0 = 0.2, S_0 = 10^6$, $S_{limit} = 10^8$). The mean probability of obtaining each

target is plotted for each of the models in Fig. 3. Probabilities were estimated using an artificial mutation data matrix (MDM) M defined below:

$$M = \begin{pmatrix} 0.98707 & 0.00122 & 0.00559 & 0.00612 \\ 0.00122 & 0.99493 & 0.00017 & 0.00367 \\ 0.00367 & 0.00017 & 0.99493 & 0.00122 \\ 0.00612 & 0.00559 & 0.00122 & 0.98707 \end{pmatrix}. \quad (8)$$

This MDM was also used as a basis for all of the following model comparisons, (with modifications to obtain the desired mutation rate per cycle).

Table 1 lists approximate probabilities for obtaining a specified sequence with each number of substitutions. From a comparison of the results of models MMo, MMc and MMv here and in Fig. 3, it can be seen that a small reduction in λ_{init} from 1 to 0.8 in both the MMc and MMv models, with no change in S_0, results in no great effect on the probability of obtaining a prescribed sequence regardless of the number of nucleotide substitutions. The introduction of variable amplification efficiency in the form of a limit on population size in the MMv model ($S_0 = 10^6, S_{limit} = 10^8$) drastically increases the probability of obtaining a prescribed sequence with fewer than around 15 nucleotide substitutions (by three orders of magnitude in some cases), and shows a commensurate decrease in the probabilities of obtaining a prescribed sequence with more than around 15 substitutions.

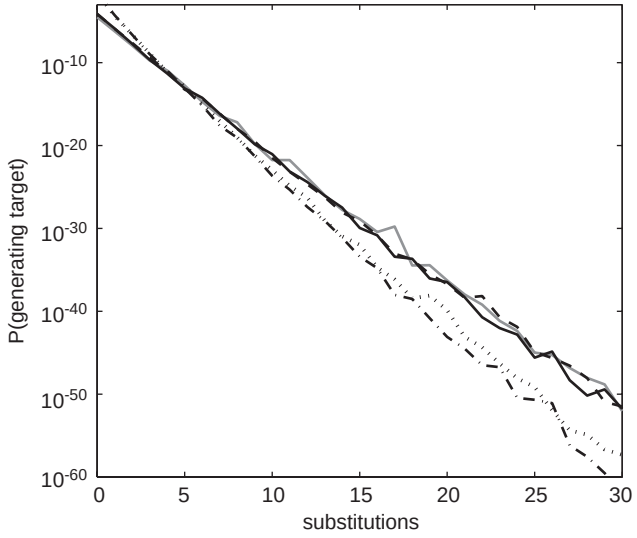

Fig. 3. The probability of selecting a given sequence with a specified number of nucleotide substitutions from the final pool of the simulation, for the MMo model (grey line), MMc model, $\lambda = 0.8$, $S_0 = 2$ (solid line); MMv model, $\lambda_0 = 0.8$, $S_0 = 2$, $S_{limit} = 10^8$ (dashed line); MMv model, $\lambda_0 = 0.8$, $S_0 = 10^6$, $S_{limit} = 10^8$ (dotted line); MMv model, $\lambda_0 = 0.2$, $S_0 = 10^6$, $S_{limit} = 10^8$ (dash-dot line).

This reinforces the earlier suggestion that generalisation of the Moore and Maranas (2000) model described in this paper describes systems that favour exploration of sequence space lying closer to the template, and that the MMo model may overestimate the ability of experimental error-prone PCR to explore regions of sequence space distant from the template. The corollary of this is that use of the MMo model to estimate mutation rate per cycle from such sampling of a PCR experiment may underestimate the true per-cycle mutation rate of the experiment.

3.4. Variation in mean Hamming distance of generated sequences from template by model

The measure of mean Hamming distance is a proxy for overall mean mutation rate, and is dependent on template length, hence all measurements are normalised to template length. We examined the effect on the mean Hamming distance from the initial template sequence of 100 sequences randomly selected from the final pool of a PCR experiment (emulating the experimental estimation of mutation rate per cycle) for 100 templates randomly chosen from a large set of bacterial coding sequences, as described above, using the MMo, MMc and MMv models under a broad range of parameter settings. In the MMo model, only mutation rate per cycle may be varied ($0.002 \leqslant \mu \leqslant 0.010$); for the MMc model, amplification efficiency ($0.2 \leqslant \lambda \leqslant 1.0$) and the initial template

Table 1
Approximate probabilities for obtaining the stated number of specified substitutions in a sequence using the specified models and parameters

Substitutions	MMo	MMc	MMv[a]	MMv[b]	MMv[c]
0	2.94E−005	7.04E−005	7.64E−005	2.11E−002	2.61E−002
5	1.75E−013	7.83E−014	5.63E−014	8.18E−014	1.33E−013
10	1.73E−022	9.00E−022	3.22E−022	2.18E−024	1.07E−023
15	1.36E−029	1.11E−030	7.61E−030	3.89E−034	1.02E−032
20	5.27E−037	3.34E−037	1.90E−037	8.02E−044	1.38E−040
25	1.00E−045	2.25E−046	1.16E−045	2.08E−051	7.08E−050
30	9.17E−053	1.69E−052	3.19E−052	1.67E−062	4.68E−058

MMc: $S_0 = 2$, $\lambda = 0.8$. MMv:
[a] $S_0 = 2$ $S_{limit} = 10^8$ $\lambda_0 = 0.8$;
[b] $S_0 = 10^6$ $S_{limit} = 10^8$ $\lambda_0 = 0.8$;
[c] $S_0 = 10^6$ $S_{limit} = 10^8$ $\lambda_0 = 0.2$.

population size were additionally varied ($2 \leqslant S_0 \leqslant 10^6$). In addition to these the population limit was also varied ($10^8 \leqslant S_{limit} \leqslant 10^{12}$) for the MMv model.

The settings tabulated in Table 2 were employed in all combinations available for each model (five run types for MMo, 100 for MMc and 300 for MMv). One hundred sequences were sampled from the final pool in each case, and the mean Hamming distance between sampled sequences and the original template sequence calculated.

Table 2
Parameter settings for runs of the MMo, MMc and MMv models

μ	λ_{init}	S_0	S_{limit}
0.0002	0.2	2	10^8
0.0004	0.4	10^2	10^{10}
0.0006	0.6	10^3	10^{12}
0.0006	0.8	10^4	
0.0010	1.0		

MMo models varied only the mutation rate, MMc models the mutation rate μ, initial amplification efficiency λ_0 and initial population S_0, MMv models varied μ, λ_0, S_0 and population limit S_{limit}.

Fig. 4a describes the relationship between mean Hamming distance (normalised to sequence length) of sequences in the final pool from their template sequence for the unaltered Moore and Maranas (2000) model. There is a monotonic increase in mean Hamming distance with increasing mutation rate. The indicated standard deviations suggest that estimation of mutation rate per cycle from this sampling scheme will be imprecise.

Reducing the amplification efficiency in the MMc model (Fig. 4b) results in a decline of mean Hamming distance from the template, as a proportion of template length, of sequences in the final pool. The effect is not strongly dependent on variations in S_0, as might be expected intuitively. This is explained as a low amplification efficiency reduces the number of amplification, and so mutation, events for the overall reaction.

The effect of decreasing amplification efficiency with a (reachable) limit on total experimental population size is

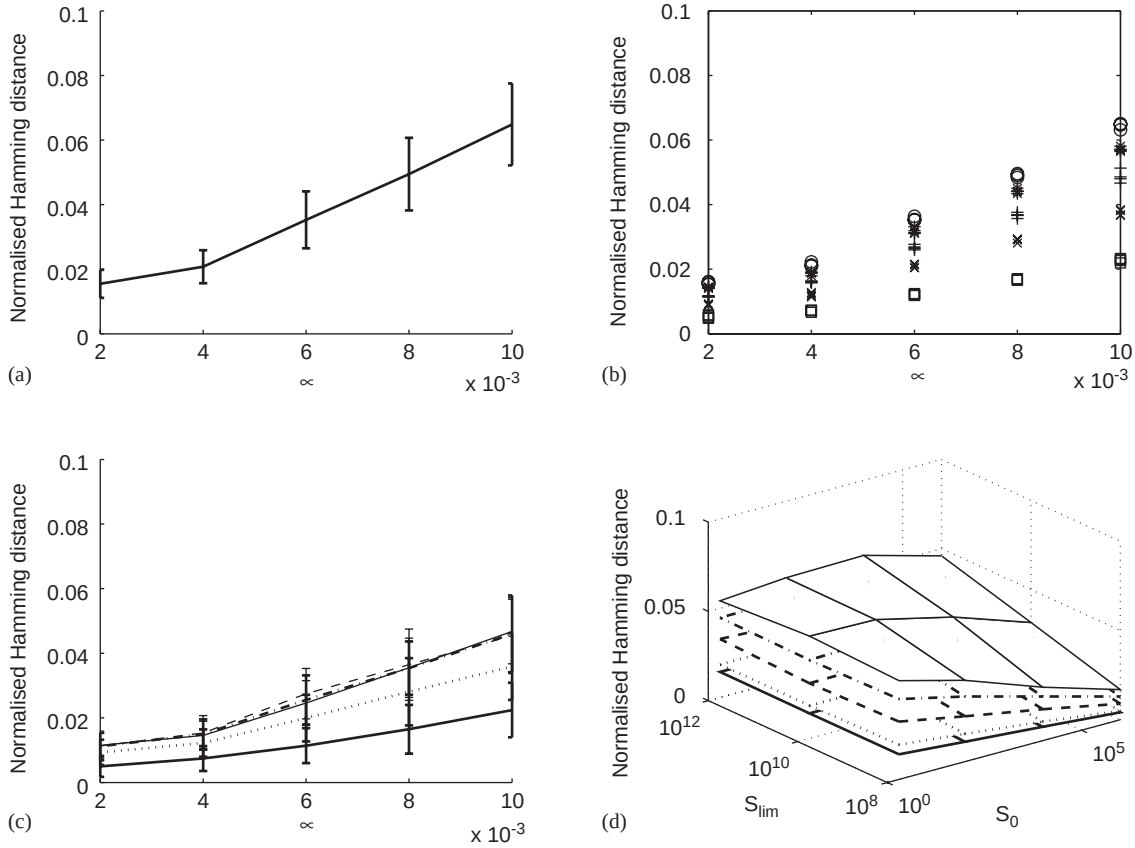

Fig. 4. Plot of mean Hamming distance from the template against mutation rate per cycle for 100 randomly-sampled sequences from the final pool of a PCR experiment, for 100 randomly-sampled template sequences using (a) MMo model; (b) MMc model with $\lambda = 1.0$ (circles), $\lambda = 0.8$ (star), $\lambda = 0.6$ (plus), $\lambda = 0.4$ (cross) and $\lambda = 0.2$ (square). Multiple points for each mutation rate and λ indicate results for $S_0 \in \{2, 10^2, 10^4, 10^6\}$; (c) MMv model with $S_0 = 10^6$, $S_{limit} = 10^{12}$ and $\lambda_0 = 1.0$ (solid line), $\lambda_0 = 0.8$ (dash-dot line), $\lambda_0 = 0.6$ (dashed line), $\lambda_0 = 0.4$ (dotted line) and $\lambda_0 = 0.2$ (thick solid line), and (d) a plot of mean Hamming distance from template against initial template population S_0 and population limit S_{limit} for the MMv model with $\lambda_0 = 0.8$, $\mu = 0.010$ (solid line), $\mu = 0.008$ (dash-dot line), $\mu = 0.006$ (dashed line), $\mu = 0.004$ (dotted line) and $\mu = 0.002$ (thick solid line).

seen in Fig. 4c for the MMv model. Here, S_0 is set to 10^6, and at this value moderate reductions of λ_0 do not have a significant effect on mean Hamming distance as a proportion of sequence length. As S_0 is close to S_{limit}, the total number of generated sequences is similar for large values of λ_0 due to the restriction of total population size that results from a rapid, early reduction in amplification efficiency. At lower values of S_0, this early total population size 'capping' effect is not seen, and a reduction in mean Hamming distance is seen for each value of μ, the same effect as seen for the MMc model. The upper limit on population size thus has a similarly limiting effect on the region of sequence space explored by a PCR experiment, where the initial template population is large enough. For all λ_0, increasing μ results in an increase in mean Hamming distance. Variations in λ during the experiment need not be controlled only by population size, and this suggests an alternative mechanism for maximising the exploration of sequence space in a single error-prone PCR experiment.

Fig. 4d shows the effect of varying S_0 and S_{limit} for a range of mutation rates at $\lambda = 0.8$. There is a general tendency for the mean Hamming distance, normalised to template length, from the template of sequences in the final pool to decline with increasing S_0, due to the 'capping' effect described above. Population limit S_{limit} has no effect on mean Hamming distance for low S_0, but increasing S_{limit} results, not surprisingly, in an increased mean Hamming distance for large S_0. For any combination of S_0 and S_{limit}, increasing μ results in a larger mean Hamming distance from the template for sequences selected randomly from the final pool.

3.5. Stochastic variation in the results of a model experiment

In the discussion above, the overall observed mutation rate has been expressed as a single measure over several experiments in terms of Hamming distance. The stochastic nature of individual substitutions has the potential to generate significant variation in observed Hamming distance from the template of sequences drawn from the final experimental pool, which is not described by this single measure. Fig. 5a illustrates the variation of this measure in a single MMv model ($\lambda_0 = 0.8$, $S_0 = 10^4$, $S_{lim} = 10^{10}$, $\mu = 0.010$) as a boxplot of this distance, normalised to template sequence length, for a randomly-selected bacterial coding sequence. The interquartile range of this measure is compressed, with only a single outlier seen. The variation in the mean value of this measure is shown in Fig. 5b for 10 runs of the MMv model with the same parameters, for 10 further randomly-selected bacterial coding sequences. The mean values for all of these sequences lie within the interquartile ranges seen for the single experiment in

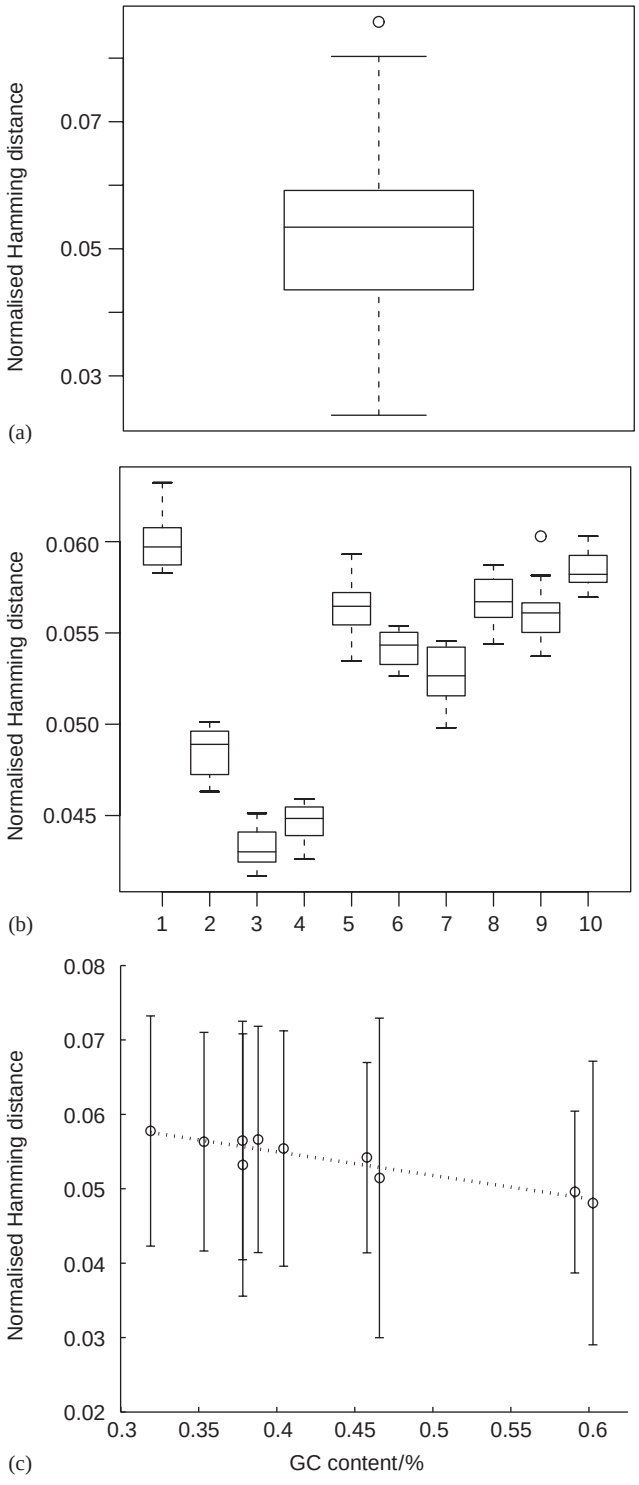

Fig. 5. Plots indicating the influence of the stochastic substitution process, plotting for the MMv model with initial amplification efficiency $\lambda_0 = 0.8$, initial population size $S_0 = 10^4$, a limit on population size $S_{lim} = 10^10$, and mutation rate per cycle $\mu = 0.010$: (a) Hamming distance (normalised to sequence length) for 100 randomly-selected sequences from the final pool of an individual experiment; (b) mean Hamming distance (normalised to sequence length) for 10 runs of the model for 10 randomly-selected bacterial coding sequences, and (c) mean Hamming distance (normalised to sequence length) for the 10 runs in subfigure (b) against the %GC content of the template sequence (linear regression line: $y = 0.068 - 0.032x$, $P \leqslant 0.0001$).

Fig. 5a, although the means show significant variation between templates. These differences are explicable in terms of the variation in %GC content of the template sequences, as the MDM used (equation 8) favours conservation of G and C nucleotides, and this relationship is demonstrated in Fig. 5c to be linear ($y = 0.068 - 0.032x$, $P \leqslant 0.0001$), mean Hamming distance from the template (normalised to template length) decreasing with increasing %GC content.

4. Conclusions

Generalisation of the Moore and Maranas (2000) (MMo) model to include terms for variable amplification efficiency and initial template population size results in a system that tends to generate sequences that are more similar to the template (in terms of Hamming distance), by comparison with the MMo model (Fig. 3).

The introduction of a term for constant amplification efficiency (MMc model) implies that a reduction in amplification efficiency reduces the mean Hamming distance from the template of sequences generated in the experiment, independently of initial template population size (Fig. 4b). This extension is not, however, sufficient to reproduce qualitatively the population size profile per cycle for an in vitro PCR experiment, nor to provide the intuitively-expected behaviour resulting from variation of initial template population size.

Introducing a term for amplification efficiency that varies as the experiment proceeds, such that amplification efficiency falls as an upper limit on the sequence population is approached (MMv model) suggests that the mean Hamming distance of the final population from the template is dependent on mutation rate, initial amplification efficiency, the initial number of template sequences and the limit on the size of the sequence population (Figs. 4c and d). Generalisation of the MMo model to the MMv model also results in the expected sigmoidal profile of sequence population size, qualitatively similar to those generated by RT-PCR, and recreates the intuitively expected behaviour of varying initial template population size.

Directly altering the limit on population size is generally outside the scope of practical experiment, while increasing mutation rate (e.g. by the addition of Mn^{2+} or mutagenic polymerase), altering dNTP balance, and reducing the number of template sequences are typically used to increase the sequence space explored in practical experiments. The possibility of manipulating amplification efficiency directly at defined stages in an error-prone PCR experiment is identified as a potential mechanism for more precisely controlling the overall observed mutation rate per sequence.

Acknowledgements

The authors would like to thank Dr Hywel Griffiths, Clive Evans and Dr Bronwen Presswell for their advice and useful discussions about practical PCR, the Biotechnology and Biological Sciences Research Council for funding the work, and the anonymous reviewers for their helpful comments and improvements to the manuscript.

Appendices

The formal proofs in these appendices are structured and notated to facilitate comparison with the proofs in Moore and Maranas (2000).

Appendix A. Proof of the total number of sequences after N cycles (constant amplification efficiency)

Let the initial number of template sequences be S_0. After 0 cycles ($N = 0$), no amplification has been performed, and the total number of sequences is S_0.

$$S_0 = S_0. \tag{A.1}$$

After the first cycle of amplification, the initial template sequences are still present and will continue to be present for all N. The number of sequences generated by this first cycle is equal to the number of template sequences multiplied by the amplification efficiency λ.

$$S_1 = S_0 + S_0\lambda = S_0(1 + \lambda). \tag{A.2}$$

After the second and subsequent cycles of amplification, the initial template sequences and the sequences generated by earlier amplification cycles are still present and will remain for all following N.

$$\begin{aligned} S_2 &= S_1 + S_1\lambda = S_1(1 + \lambda) = S_0(1 + \lambda)(1 + \lambda) \\ &= S_0(1 + \lambda)^2. \end{aligned} \tag{A.3}$$

The general form of equation (A.3) is

$$\begin{aligned} S_N &= S_{N-1} + S_{N-1}\lambda = S_{N-1}(1 + \lambda) = S_{N-2}(1 + \lambda)^2 \\ &= S_{N-N}(1 + \lambda)^N \\ &= S_0(1 + \lambda)^N. \end{aligned} \tag{A.4}$$

Appendix B. Calculation of $Z_{N,n}$(constant amplification efficiency)

We represent the number of strands that are the product of n extension steps after N total PCR cycles

by $Z_{N,n}$ and amplification efficiency by λ. Initially,

$$Z_{0,0} = S_0 \tag{B.1}$$

but since these template sequences are not altered in the experiment, and are the only sequences not to be the product of any PCR amplification,

$$Z_{N,0} = S_0. \tag{B.2}$$

Also, after N cycles, no sequences can be the result of more than N extension steps, so

$$Z_{N,n} = 0, \qquad n > N. \tag{B.3}$$

After N PCR cycles, a sequence that is the result of n extension steps must have been amplified from a sequence that was itself the product of $n-1$ extension steps. The sequence must either have been produced in the current, Nth cycle (i.e. was amplified from one of $Z_{N-1,n-1}$ sequences), or was already present in the reaction mixture (i.e. is one of $Z_{N-1,n}$ sequences). So

$$Z_{N,n} = Z_{N-1,n-1}\lambda + Z_{N-1,n}. \tag{B.4}$$

As in Moore and Maranas (2000), we construct a proof by induction of

$$Z_{N,n} = \binom{N}{n} S_0 \lambda^n \tag{B.5}$$

by first demonstrating that it is valid for $n = 0$, 1 and 2.

For $n = 0$,

$$Z_{N,0} = S_0 = \binom{N}{0} S_0 \lambda^0. \tag{B.6}$$

For $n = 1$, from (B.4)

$$\begin{aligned} Z_{N,1} &= Z_{N-1,1} + Z_{N-1,0}\lambda \\ &= Z_{N-1,1} + S_0\lambda \\ &= Z_{N-2,1} + Z_{N-2,0}\lambda + S_0\lambda \\ &= Z_{N-2,1} + 2S_0\lambda \\ &= Z_{N-3,1} + 3S_0\lambda \\ &= Z_{N-k,1} + kS_0\lambda, \qquad \forall 0 \leqslant k \leqslant N. \end{aligned} \tag{B.7}$$

At the limit of recursion, $k = N$ and

$$Z_{N-k,1} = Z_{N-N,1} = Z_{0,1} = 0 \tag{B.8}$$

$$\Rightarrow Z_{N,1} = Z_{0,1} + NS_0\lambda = \binom{N}{1} S_0 \lambda^1. \tag{B.9}$$

For $n = 2$, from (B.4)

$$\begin{aligned} Z_{N,2} &= Z_{N-1,2} + Z_{N-1,1}\lambda \\ &= Z_{N-2,2} + Z_{N-2,1}\lambda + Z_{N-1,1}\lambda \\ &= Z_{1,1}\lambda + Z_{2,1}\lambda + \cdots + Z_{N-2,1}\lambda + Z_{N-1,1}\lambda \\ &= \lambda \sum_{k=1}^{N-1} Z_{k,1} \\ &= \lambda \sum_{k=1}^{N-1} kS_0\lambda \\ &= \binom{N}{2} S_0 \lambda^2. \end{aligned} \tag{B.10}$$

To complete the proof by induction, we now assume (B.5) to be true and use it to demonstrate that

$$Z_{N,n+1} = \binom{N}{n+1} S_0 \lambda^{n+1} \tag{B.11}$$

as follows:

$$\begin{aligned} Z_{N,n+1} &= Z_{N-1,n+1} + Z_{N-1,n}\lambda \\ &= Z_{N-2,n+1} + Z_{N-2,n}\lambda + Z_{N-1,n}\lambda \\ &= Z_{N-3,n+1} + Z_{N-3,n}\lambda + Z_{N-2,n}\lambda + Z_{N-1,n}\lambda \\ &= \lambda(Z_{n,n} + Z_{n+1,n} + \cdots + Z_{N-2,n} + Z_{N-1,n}) \\ &= \lambda \sum_{k=n}^{N-1} Z_{k,n}. \end{aligned} \tag{B.12}$$

Substituting from (B.5),

$$\Rightarrow Z_{N,n+1} = \lambda \sum_{k=n}^{N-1} \binom{k}{n} S_0 \lambda^n. \tag{B.13}$$

Letting $s = k - n$,

$$\begin{aligned} Z_{N,n+1} &= S_0\lambda^{n+1} \sum_{s=0}^{(N-n)-1} \binom{n+s}{n} \\ &= \binom{N}{n+1} S_0 \lambda^{n+1}. \end{aligned} \tag{B.14}$$

Appendix C. Proof of total number of sequences after N cycles (arbitrary, variable amplification efficiency)

Let the initial number of template sequences be S_0 and the amplification efficiency for the Nth cycle be λ_N. That is, λ_N is the amplification efficiency that prevails at amplification cycle N, resulting in the production of $(S_N - S_{N-1})$ sequences, which contribute to the S_N sequences present at the conclusion of cycle N. For the 0th cycle, $\lambda_0 = 0$, so for $N = 0$,

$$S_N = S_0(1 + \lambda_0) = S_0. \tag{C.1}$$

After the first cycle of amplification, the initial template sequences are still present and will continue to be present for all subsequent amplification cycles. The number of sequences generated by an amplification cycle depends on the number of sequences present at the beginning of the cycle and also the amplification efficiency for the current cycle, here λ_1. The total number of sequences present at the end of the first cycle is thus ($N = 1$)

$$\begin{aligned} S_1 &= S_0 + S_0\lambda_1 = S_0(1+\lambda_1) \\ &= S_0(1+\lambda_0)(1+\lambda_1) \\ &= S_0 \prod_{k=0}^{1}(1+\lambda_k). \end{aligned} \tag{C.2}$$

The total number of sequences present after two cycles is dependent on the number of sequences in the reaction mixture after the first amplification cycle (i.e. at the beginning of the second amplification cycle), and the amplification efficiency for the second amplification cycle λ_2, so for $N = 2$

$$\begin{aligned} S_2 &= S_1 + S_1\lambda_2 \\ &= S_1(1+\lambda_2) = S_0(1+\lambda_1)(1+\lambda_2) \\ &= S_0(1+\lambda_0)(1+\lambda_1)(1+\lambda_2) \\ &= S_0 \prod_{k=0}^{2}(1+\lambda_k). \end{aligned} \tag{C.3}$$

Generally, for any N

$$\begin{aligned} S_N &= S_{N-1}(1+\lambda_N) \\ &= S_{N-2}(1+\lambda_{N-1})(1+\lambda_N) \\ &= S_{N-3}(1+\lambda_{N-2})(1+\lambda_{N-1})(1+\lambda_N) \\ &= S_{N-k}(1+\lambda_{N-k+1})(1+\lambda_{N-k+2}) \times \cdots \\ &\quad \times (1+\lambda_{N-1})(1+\lambda_N). \end{aligned} \tag{C.4}$$

At the limit of recursion, $k = N$ and

$$\begin{aligned} S_N &= S_0(1+\lambda_1)(1+\lambda_2) \times \cdots \times (1+\lambda_{N-1})(1+\lambda_N) \\ &= S_0 \prod_{k=0}^{N}(1+\lambda_k). \end{aligned} \tag{C.5}$$

Setting $\lambda_N = \lambda$ (constant amplification efficiency) gives

$$S_N = S_0 \prod_{k=0}^{N}(1+\lambda_k) = S_0(1+\lambda)^N, \tag{C.6}$$

which is equivalent to equation (A.4). Setting $\lambda = 1$ recovers the equivalent equation from Moore and Maranas (2000)

$$S_N = S_0(1+1)^N = S_0 \cdot 2^N. \tag{C.7}$$

References

Arezi, B., Hansen, C.J., Hogrefe, H.H., 2002. Efficient and high fidelity incorporation of dye-terminator by a novel archaeal DNA polymerase mutant. J. Mol. Biol. 322, 719–729.

Barik, S., 1998. Mutagenesis by megaprimer PCR. In: Horton, R.M., Tait, R.C. (Eds.), Genetic Engineering With PCR, Horizon Scientific Press, pp. 25–38.

Cadwell, R., Joyce, G., 1991. Randomization of genes by PCR mutagenesis. PCR Meth. Appl. 2, 28–33.

Eckert, K.A., Kunkel, T., 1991. DNA polymerase fidelity and the polymerase chain reaction. PCR Meth. Appl. 2, 17–24.

Erlich, H.A. (Ed.), 1989. PCR Technology: Principles and Applications for DNA Amplification, Stockton Press, pp. 1–5.

Hayashi, K., 1990. Mutations induced during the polymerase chain reaction. Technique 2, 216–217.

Krawczak, M., Reiss, J., Schmidtke, J., Rosler, U., 1989. Polymerase chain reaction: replication errors and reliability of gene diagnosis. Nucl. Acids Res. 17, 2197–2201.

Leung, D., Chen, E., Goeddel, D., 1989. A method for random mutagenesis of a defined DNA segment using a modified polymerase chain reaction. Technique 1, 11–15.

Maruyama, I.N., 1990. Estimation of errors in the polymerase chain reaction. Technique 2, 302–303.

Moore, G.L., Maranas, C.D., 2000. Modeling DNA mutation and recombination for directed evolution experiments. J. Theor. Biol. 205, 483–503.

Moore, J., Jin, H., Kuchner, O., Arnold, F.H., 1997. Strategies for the in vitro evolution of protein function: enzyme evolution by random recombination of improved sequences. J. Mol. Biol. 272, 336–347.

Mullis, K.B., Faloona, F., 1987. Specific synthesis of DNA in vitro via a polymerase catalyzed chain reaction. Methods Enzymol. 155, 335–351.

Newton, C.R., Graham, A., 1997. PCR. Introduction to Biotechniques, second ed., BIOS Scientific.

Nieuwlandt, D., 1998. In vitro selection of functional nucleic acid sequences. In: Horton, R.M., Tait, R.C. (Eds.), Genetic Engineering With PCR. Horizon Scientific Press, pp. 117–132.

Oue, S., Olamoto, A., Yano, T., Kagamiyama, H., 1999. Redesigning the substrate specificity of an enzyme by cumulative effects of the mutations of non-active site residues. J. Mol. Biol. 274, 2344–2349.

Sun, F., 1995. The polymerase chain reaction and branching processes. J. Comput. Biol. 2, 63–86.

Tait, R.C., Horton, R.M., 1998. An introduction to genetic engineering with PCR. In: Horton, R.M., Tait, R.C. (Eds.), Genetic Engineering With PCR. Horizon Scientific Press, pp. 117–132.

Voigt, C.A., Kauffman, S., Wang, Z.-G., 2001. Rational evolutionary design: the theory of in vitro protein evolution. Adv. Prot. Chem. 55, 79–160.

Weiss, G., von Haeseler, A., 1995. Modeling the polymerase chain reaction. J. Comput. Biol. 2, 49–61.

Zaccolo, M., Gherardi, E., 1999. The effect of high-frequency random mutagenesis on in vitro protein evolution: a study on TEM-1 β-lactamase. J. Mol. Biol. 285, 775–783.

Zhao, H., Arnold, F.H., 1997. Combinatorial protein design: strategies for screening protein libraries. Curr. Opin. Struct. Biol. 7, 480–485.