Mathematical modeling of 16S ribosomal DNA amplification reveals optimal conditions for the interrogation of complex microbial communities with phylogenetic microarrays

Oleg Paliy\(^1,\)^* and Brent D. Foy\(^2\)

\(^1\)Department of Biochemistry and Molecular Biology and \(^2\)Department of Physics, Wright State University, Dayton, OH 45435, USA

ABSTRACT

Motivation: Many current studies of complex microbial communities rely on the isolation of community genomic DNA, amplification of 16S ribosomal RNA genes (rDNA) and subsequent examination of community structure through interrogation of the amplified 16S rDNA pool by high-throughput sequencing, phylogenetic microarrays or quantitative PCR.

Results: Here we describe the development of a mathematical model aimed to simulate multitemplate amplification of 16S ribosomal DNA sample and subsequent detection of these amplified 16S rDNA species by phylogenetic microarray. Using parameters estimated from the experimental results obtained in the analysis of intestinal microbial communities with Microbiota Array, we show that both species detection and the accuracy of species abundance estimates depended heavily on the number of PCR cycles used to amplify 16S rDNA. Both parameters initially improved with each additional PCR cycle and reached optimum between 15 and 20 cycles of amplification. The use of more than 20 cycles of PCR amplification and/or more than 50 ng of starting genomic DNA template was, however, detrimental to both the fraction of detected community members and the accuracy of abundance estimates. Overall, the outcomes of the model simulations matched well available experimental data. Our simulations also showed that species detection and the accuracy of abundance measurements correlated positively with the higher sample-wide PCR amplification rate, lower template-to-template PCR bias and lower number of species in the interrogated community. The developed model can be easily modified to simulate other multitemplate DNA mixtures as well as other microarray designs and PCR amplification protocols.

Contact: oleg.paliy@wright.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on March 2, 2011; revised on May 2, 2011; accepted on May 22, 2011

1 INTRODUCTION

Owing to the development of sensitive and novel DNA and RNA interrogation technologies, there is a surge of studies in the current literature exploring the populational structure and function of various complex microbial communities (Brodie et al., 2007; Gao et al., 2007; Huber et al., 2007). The ability to isolate and subsequently examine total community DNA and RNA without any need to culture individual microbial species and cells allows analysis of systems that would otherwise be difficult to profile and examine including microbiota of human intestine and other epithelial surfaces and the microbes of soils and ocean waters (Eckburg et al., 2005; Gao et al., 2007; Huber et al., 2007; Kent and Triplett, 2002).

Gene coding for the small ribosomal subunit RNA molecule (16S rRNA in prokaryotes and 18S rRNA in eukaryotes) has been used in the vast majority of such studies due to its ubiquitous presence in all organisms and because of the conservation of its nucleotide sequence (Cannone et al., 2002). In a typical experimental design to profile microbial community structure, total genomic DNA (gDNA) isolated from a sample of interest is subjected to rounds of 16S rRNA gene (rDNA) specific amplification in polymerase chain reaction (PCR) using two ‘universal’ primers complementary to the 16S rRNA gene sequences, any PCR amplification of community DNA is multitemplate. PCR amplification of such gDNA has been described that include (i) unequal denaturation of templates based on GC content of DNA sequences; (ii) higher binding efficiency of GC-rich variants of degenerate amplification primers to the template at the same annealing temperature; and (iii) competitive re-annealing

\(\text{Article in Press} \quad \text{doi:10.1093/bioinformatics/btr326} \quad \text{Advance Access publication June 7, 2011} \)
of abundant templates at later amplification cycles discriminating against major templates (Kurata et al., 2004; Polz and Cavanaugh, 1998; Sachse, 2004; Sipos, et al., 2007). In addition, the stochastic nature of the PCR amplification process during the first few cycles of amplification can lead to a reaction-specific skew in relative distribution of amplified templates producing so-called PCR drift (Polz and Cavanaugh, 1998; Wagner et al., 1994). Finally, genomic DNA isolated from environmental sources or human bodily fluids such as fecal matter is known to often contain PCR inhibitors that are able to significantly impede DNA amplification sample-wide (Lampe et al., 2000). All these mechanisms result in a skew of the product-to-template ratios for most DNA species in complex samples.

Several studies have provided theoretical framework aimed to describe the individual components of the PCR or to simulate a particular PCR bias. Initial efforts focused on the modeling of individual biochemical reactions including template–enzyme kinetics (Schnell and Mendoza, 1997), DNA renaturation (Weimur and Davidson, 1968) and DNA re-annealing (Suzuki and Giovannoni, 1996). Two recent studies described detailed simulations of kinetic processes of PCR based on the calculations of mass preservation and equilibrium coefficients (Lee et al., 2006; Mehra and Hu, 2005). Both studies focused on the behavior of single-template amplification process, and in addition provided considerations of biochemical constants for each equilibrium reaction. In this work, we extend these previous PCR simulations with a quantitative model of multitemplate PCR amplification of complex 16S ribosomal DNA mixture derived from the simulated genome pool of 400 different microbial species. The multitemplate PCR amplification is computationally assessed through the detection of amplified 16S ribosomal DNA by phylogenetic microarray.

### 2 METHODS

#### 2.1 General model description

A typical experimental procedure of microbial community analysis consists of selective amplification of 16S rDNA from the total genomic DNA, applying the pool of amplified and total gDNA onto a microarray, and then measuring hybridization of this mixture to different microarray probes. Detection of bacterial species of intestinal microbiota with Microbiota Array (Paly et al., 2009) was used as a model of this method. In the in silico simulation of the interrogation process, developed in Matlab (version 7.3: The Mathworks, Inc.), the starting gDNA sample was subjected to successive rounds of PCR amplification producing a mixed sample containing original gDNA and amplified 16S rDNA molecules. After each PCR amplification round, detection of 16S rDNA species in the mixture was assessed by the microarray detection algorithm.

#### 2.2 Model of the multitemplate PCR

The process of PCR amplification was simulated for each bacterial species in the mixture individually, and the accumulation of amplified 16S rDNA was modeled for each species as:

\[
P_N = P_0 \times \prod_{i=1}^{N} (1 + \Delta E_i)
\]

(1)

where \(P_N\) is the total amplified 16S rDNA after \(N\) cycles of PCR amplification, \(P_0\) is the starting amount of 16S rDNA and \(\Delta E_i\) is amplification efficiency at cycle \(i\) for each species 16S rDNA (Schnell and Mendoza, 1997).

#### 2.3 Simulation of microarray detection

Microarray analysis of species 16S rDNA in the amplified sample was simulated by comparing the amount of each species rDNA versus the microarray detection limit (see below). If the species-specific 16S rDNA amount was above detection limit, the corresponding species was treated as detected, and its gDNA abundance was equal to its 16S (DNA amount in the amplified mixture. Otherwise, the species was treated as undetected, and its measured gDNA abundance was assumed to be 0.

#### 2.4 Parameter estimation

The following modeling parameters and coefficients were estimated (Table 1):

1. Yield of PCR amplification as a function of the amount of starting material: PCR amplification yield was estimated from semi-quantitative PCR tests that utilized between 50 and 500 ng of gDNA template and between 15 and 30 cycles of amplification (Supplementary Fig. S1). The yield rate (defined as the ratio of the final amplified amount divided by the starting DNA amount)

### Table 1. Definition of parameter values used in simulation experiments

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(N_{sp})</td>
<td>Number of species in sample</td>
<td>400*</td>
</tr>
<tr>
<td>16Strac</td>
<td>Fraction of 16S rDNA per genome</td>
<td>0.0025</td>
</tr>
<tr>
<td>(AE_{\text{MAX}})</td>
<td>Maximum amplification rate, (1 + AE_{\text{MAX}})</td>
<td>1.75</td>
</tr>
<tr>
<td>VOlPCR</td>
<td>PCR reaction volume</td>
<td>50 (\mu)L</td>
</tr>
<tr>
<td>CONCorgx</td>
<td>Concentration of total nucleotides</td>
<td>(8.0 \times 10^{-3}) M</td>
</tr>
<tr>
<td>CONCENZ</td>
<td>Concentration of DNA polymerase</td>
<td>(2.7 \times 10^{-3}) M</td>
</tr>
<tr>
<td>CONCCPR</td>
<td>Concentration of each primer</td>
<td>(2.0 \times 10^{-3}) M</td>
</tr>
<tr>
<td>(K_T0)</td>
<td>Template-template self-annealing</td>
<td>(1.0 \times 10^{10}) M^{-1}</td>
</tr>
<tr>
<td>(K_T1)</td>
<td>Template–primer annealing</td>
<td>(1.1 \times 10^{11}) M^{-1}</td>
</tr>
<tr>
<td>(K_{TI})</td>
<td>Template–primer–enzyme–nucleotide annealing</td>
<td>(1.0 \times 10^{11}) M^{-1}</td>
</tr>
<tr>
<td>(K_{TN})</td>
<td>Template–primer–enzyme–nucleotide annealing</td>
<td>(1.8 \times 10^{9}) M^{-1}</td>
</tr>
</tbody>
</table>

*Based on data from (Rogacheva et al., 2011) and (Paly et al., 2009).

*Based on available mixture composition of standard Taq PCR kits.

The overall species-specific PCR amplification rate \(AR_i = 1 + AE_{\text{MAX}}\) at cycle \(i\) was defined as:

\[
AR_i = 1 + AE_{\text{MAX}} \times \frac{\text{CONC}_i}{\text{CONC}_i + \text{CONC}_{\text{PR} - \text{ENZ} - \text{NUCL}}}
\]

(2)

where \(AE_{\text{MAX}}\) is maximum amplification efficiency, \(\text{CONC}_i\) is the concentration of the template–enzyme–nucleotide complex (growing complementary nucleotide chain) at equilibrium. The last term in the equation models the adjustment of amplification efficiency based on biochemical reaction equilibria and takes into account template self-annealing (Mehra and Hu, 2005), template-specific PCR bias due to variability of the 16S rDNA nucleotide sequences complementary to universal primers among different species (Polz and Cavanaugh, 1998) and sample-wide reduction in amplification efficiency due to enzyme and nucleotide resource limitation (Lee et al., 2006; Mehra and Hu, 2005; Sachse, 2004). A detailed description of the mathematical model is available in Supplementary Material.

#### 2.4 Parameter estimation

The following modeling parameters and coefficients were estimated (Table 1):

1. Yield of PCR amplification as a function of the amount of starting material: PCR amplification yield was estimated from semi-quantitative PCR tests that utilized between 50 and 500 ng of gDNA template and between 15 and 30 cycles of amplification (Supplementary Fig. S1). The yield rate (defined as the ratio of the final amplified amount divided by the starting DNA amount)
where SPFrac\textsubscript{TRUE} is the true fraction of 16S DNA represented by species \( i \) in the initial sample, SPFrac\textsuperscript{EXP} is the fraction of 16S DNA represented by species \( i \) estimated by microarray interrogation of PCR-amplified sample and \( N_{\text{SPECIES}} \) is the total number of species in the sample. SPFrac\textsuperscript{EXP} for all undetected species was set to 0. In words, this accuracy measure is the average of the absolute fractional error for each species fraction estimate. A lower value represents greater accuracy, with a zero value being perfect accuracy.

2.6 Model limitations
Several parameters known to influence the behavior of standard PCRs were omitted from the model for simplicity because they were not found to be modified often in phylogenetic studies examined. These included \( \text{Mg}^{2+} \) concentration, varying annealing temperature and temperature ramp rate. Because DNA polymerases currently used for PCRs are very thermostable, denaturation of the enzyme during PCR was considered to be insignificant. Since our model was focused on the determination of amplification efficiency of each template at each PCR cycle, it did not include detailed simulation of the biochemistry of melting and extension reactions that were considered in detail previously (Lee et al., 2006; Mehra and Hu, 2005). Nevertheless, the behavior of our PCR simulation was similar to those reported elsewhere (Lee et al., 2006; Mehra and Hu, 2005). Finally, our algorithm did not model complexities of microarray hybridizations; rather it assumed a perfect measurement of each species abundance as long as it was above the microarray detection limit.

3 RESULTS
3.1 Fraction of bacterial detection
Figure 1A and B displays detection fraction of microbial species in the sample as a factor of the number of PCR amplification cycles and the amount of starting material. As expected, bacterial detection increased with an initial increase in the number of PCR amplification cycles. The curves of species detection fraction were similar when starting material was kept <50 ng. After initial increase, the detection fraction started to plateau out around 15th PCR cycle and reached maximum level around cycle 20 (91–93% of bacterial species detected for samples with <50 ng starting gDNA). The detection slowly decreased with further PCR amplification to ~90%.

The detection fraction curve was different for samples with higher starting amounts of gDNA. When increasing initial gDNA template from 100 ng to 1000 ng, the rate of detection became progressively lower. At 15 cycles of PCR amplification, 86, 78, 71 and 63% of all species in the sample were detected for samples containing 100, 300, 600 and 1000 ng of starting genomic DNA, respectively. The lower rate of detection is explained in part by an increased amount of PCR inhibitors in large gDNA samples reducing the rate of template amplification. The initial sample-wide amplification rate dropped from 1.75 for reactions with 1–50 ng starting material to 1.50 for samples with 1000 ng of starting gDNA (Supplementary Fig. S2). Elevated template self-annealing in large samples also contributed to a lower amplification rate. In addition, because of higher initial concentration of DNA templates, PCR reached resource limitation stage faster for larger gDNA samples. A drop in the amplification rate occurred at cycle 20 for PCR with 1 ng starting gDNA versus cycle 10 for reactions with 1000 ng starting gDNA (Supplementary Fig. S2). This lowered the 16S rDNA enrichment factor of the amplified samples (Supplementary Fig. S3). Whereas 16S rDNA constituted 86% of total sample after 15 cycles of PCR amplification for sample containing 1 ng of starting gDNA (99% after 20 cycles), this fraction was 73% for 100 ng starting gDNA sample (83% after 20 cycles, 93% after 40 cycles) and only 22% for 1000 ng starting gDNA sample (30% after 20 cycles, 51% after 40 cycles).
fraction of 16S rDNA in genomic DNA samples was 0.25%. The relative fractions of detection at 15, 20, 25 and 30 cycles of PCR amplification observed in our simulations were in good accordance with our previous experimental analyses of human fecal samples using Microbiota Array (Rigsbee et al., 2011).

Carrying out PCR amplification past 20 cycles actually led to a slight reduction of species detection in samples with small starting amount of gDNA. This was connected to the existence of the simulated PCR bias among different species, since species detection did not drop when PCR bias was turned off (Fig. 2C). Our simulation results are consistent with previous observations that the use of large number of PCR amplification cycles to enrich genomic DNA samples with 16S rDNA results in reduced estimates of microbial community diversity (Bonnet et al., 2002).

3.2 Accuracy of 16S rDNA abundance measurements

Figure 1C and D displays the accuracy of microarray estimation of species relative abundance in a sample. Independent of the amount of starting material used, the accuracy of community composition estimates improved with an initial increase in the number of PCR amplification cycles a sample was subjected to. The optimum accuracy was inversely proportional to the starting gDNA amount: it was similar for samples with 1–10ng of starting gDNA, but was progressively worse for samples with higher starting gDNA amounts. The best accuracy (minimum value on Figure 1C and D graphs) was observed at cycle number 17, 15, 12 and 11 for samples with 1, 10, 100 and 1000ng of starting gDNA, respectively. In all cases, the accuracy then started to decline sharply as the number of PCR amplification cycles continued to grow.

Several different effects influenced the observed accuracy of 16S rDNA abundance measurements. Since many bacterial species are not detected in samples subjected to a low number of PCR amplification cycles, this leads to a poor concordance of relative abundance values between initial (true) and measured sample compositions. However, because each additional PCR amplification cycle skews the original distribution of species abundances due to unequal amplification efficiency of 16S rDNA of different species, use of many PCR amplification cycles leads to a significant overall deviation of measured signals from the initial species distribution (Fig. 2D). In addition, template self-annealing introduced another amplification deviation in samples with higher starting gDNA that influenced the position of optimum accuracy (see Supplementary Fig. S4). Thus, even though a progressively higher fraction of community members is detected until at least the 20th PCR amplification cycle, accumulation of PCR amplification biases leads to optimum accuracy being reached at an earlier cycle.

3.3 Sensitivity analysis of estimated model parameters

To test the dependence of model predictions on values chosen for key parameters, sensitivity analyses were carried out where each parameter was systematically permuted while keeping all other
values unchanged. In all cases, the starting gDNA amount was set to 10 ng.

Maximum amplification rate: the rate of gDNA amplification can be influenced by the amount of inhibitors in the sample, by the complementarity of amplification primers used and by specific PCR conditions employed. Figure 2A and B shows how different maximum PCR amplification rates affect species detection and accuracy of abundance estimates. As expected, the slope of detection
fraction was generally proportional to the PCR amplification rate. The 90% detection point was reached at PCR cycle 12, 14, 17, and 21 for reactions with amplification rates of 2.00, 1.75, 1.60, and 1.45, respectively. The highest detection fraction was the same for all simulations and thus was independent from the starting amplification rate. The optimum accuracy was generally observed for the PCR cycle that corresponded to detection fraction reaching 90% level.

**Template PCR amplification bias:** due to the differences in the primer annealing and in sequence replication rates, different 16S rDNA species display slightly different rates of PCR amplification, which in a complex sample can lead to deviations of the post-PCR template ratios from the original sample composition. A number of approaches to experimentally affect PCR bias level are known and can be used (Kurata et al., 2004; Sipos et al., 2007). For this sensitivity analysis, the SD of the template–primer equilibrium $K_{EQ}$ constant was systematically permuted to determine the influence of PCR bias level on both the detection and accuracy performance of microarray measurements (Fig. 2C and D). Maximum detection fraction was inversely proportional to the bias level. With a 0.15% median per-cycle amplification rate bias, 95% of species were detected in our simulations; however, this fraction dropped to 87% for amplifications with 5.9% median per-cycle bias (Fig. 2C).

An even larger negative effect of PCR amplification bias was observed for the accuracy of abundance measurements (Fig. 2D). This was expected since each PCR cycle introduces deviations into relative species abundances. With PCR bias close to 0, accuracy stayed at the same level after reaching an optimum.

**Concentration of amplification primers:** Figure 2E and F shows simulation results of varying the initial concentration of amplification primers. Because in standard PCR primers are always in abundance compared with the amount of template ($2.6 \times 10^{-12}$ M of 16S rDNA template in 50 ng of total genomic DNA versus $2.0 \times 10^{-13}$ M of each primer), primer concentrations are often assumed to be invariant during PCR. Somewhat unexpectedly, both of the main assessed parameters showed primer concentration dependence. Though the optimum accuracy and the highest detection fraction were generally not affected, the decline in both parameters with higher number of amplification cycles was inversely proportional to the overall primer concentration in the reaction. Increasing each primer concentration beyond $350 - 400$ nM had negligible additional effect. Our finding can be explained by the fact that because phylogenetically conserved primers with several degenerate positions are often used in 16S rDNA amplification studies (Paly et al., 2009), the actual effective concentrations of each forward and reverse primer for each specific 16S rDNA template are likely significantly lower. This effect was included in our model. A lower effective primer concentration will lead to a lower overall template amplification rate through shifting the reaction equilibrium. This reduction will be particularly noticeable when concentration of the templates is comparable with primer amounts at the later stages of PCR amplification.

**Number of species in the community sample:** based on the previous analysis of intestinal microbial communities (Rigsbee et al., 2011), our simulations were run with 400 species members. However, community diversity is expected to be vastly different for diverse ecological environments and microbial communities. Figure 2G and H shows detection and accuracy results for simulations where community complexity was varied between 160 and 640 members. As expected, overall fraction of species detection decreased with an increase in the number of community members, since the same amount of genomic DNA was distributed among larger number of members. We detected all members of the 160-member community but only ~4% of the 640-member community. A similar relationship was also observed for accuracy estimates—microarray results were generally more accurate for a low-member community due to the higher fraction of detected members.

### 4 DISCUSSION AND CONCLUSIONS

In this work, we have developed the first detailed model of PCR amplification and microarray detection for the quantitative analysis of microbial community structure. Model performance matched well available experimental data including total number of detected species, total PCR reaction yield and species detection differences for 15, 20, 25 and 30 cycles of PCR amplification (see Supplementary Table S1 and (Rigsbee et al., 2011)). Model simulations led to the following conclusions:

1. Species detection showed a positive relationship with the number of PCR amplification cycles and lower template-to-template amplification bias. On the other hand, it was inversely proportional to the number of species comprising the interrogated community. A detection fraction plateau was reached in each case due to a significant and quick drop in the PCR amplification rate once the template concentration became high enough to affect chemical reaction equilibria.

2. Accuracy of 16S rDNA abundance measurements had a non-monotonic relationship with most tested parameters. The accuracy initially improved with each successive PCR amplification cycle due to a significant increase in the fraction of detected species. However, once reaching an optimum, the accuracy started to decline due to the rising PCR amplification bias with each additional amplification cycle. Overall, PCR amplification bias had the most profound effect on the detection and accuracy estimates among all parameters tested in our model.

3. A caution has to be exercised when comparing species composition and abundance among microbial communities with significantly different numbers of community members. With the same experimental protocol used, a community with many members will have a lower detection fraction than a community with fewer species. Thus, the actual differences among such communities would be underestimated.

4. Not surprisingly, our experiments indicated that the total yield of PCR amplification correlated positively with the amount of starting material as well as with the number of amplification cycles employed. The yield neared a plateau in each case (Supplementary Fig. S5) which coincided with a drop in the PCR amplification rate (Supplementary Fig. S2). Previous studies also showed that efficiency of DNA duplex denaturation is decreased at higher concentrations, which leads to poor PCR performance when too much sample DNA is added to the mixture (Sachse, 2004). While this indicates that carrying out many PCRs with smaller amounts of starting material will produce the best overall yield from the same amount of available gDNA, the use of many PCRs carries significant financial and labor costs that also must...
be considered. The same amount of material (1500 ng) was ‘loaded’ onto the microarray in our simulations; therefore, for reactions with small starting amounts and low number of PCR amplification cycles, the simulation assumed that many PCRs were combined to achieve the needed load. This is not likely to be practical for most experiments: assuming a 50% loss of PCR products during reaction purification, 27 separate PCR amplification reactions would have to be combined for samples with 10 ng of starting gDNA subjected to 15 cycles of PCR amplification (see Supplementary Fig. 56).

Overall, when assuming maximum sample-wide PCR amplification rate of 1.75, the optimal conditions for PCR amplification of complex 16S rDNA samples included a combination of smaller amount of starting gDNA (<50 ng) and moderate number of PCR amplification cycles (15–20 cycles). The use of a higher number of PCR amplification cycles should be avoided (Kanagawa, 2003). The particular optimal values for starting gDNA amount and the number of amplification cycles will be affected by the specific microbial community under investigation and the details of the experimental protocol. The developed model is well suited to provide these optimal condition estimates for different experimental setups. Our model predicts that combination of 50 ng starting material and 18–20 cycles of PCR amplification will require pooling of between 5 and 8 independent PCRs (assuming 1500 ng microarray load), which is of reasonable cost and labor-wise. An additional advantage of pooling multiple independent PCR amplifications of the same sample is the mitigation of potential measurement biases associated with PCR drift (Polz and Cavanaugh, 1998; Wagner et al., 1994). Our experimental PCR amplification and purification tests (data not shown and Palíy et al. (2009); Riggsbee et al. (2011)) are consistent with these predictions.

In summary, this work develops a mathematical model which can assist with establishing the optimal experimental conditions for the selective amplification and microarray-based measurements of community 16S rDNA. The developed model can be easily modified to simulate the interrogation of other microbial communities as well as other microarray designs or PCR amplification protocols. For example, to simulate different microbial communities the number of species and their expected abundance distribution in the sample can be modified (Fig. 2G). On the other hand, microarray detection such as cross-hybridization behavior among different probes and 16S rDNA fragments (Rigsbee et al., 2011).

Funding: National Institutes of Health grant (AT003423) to O.P. Conflict of Interest: none declared.

REFERENCES


