LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-TR-221031

Virtual PCR

S. N. Gardner, D. S. Clague, J. A. Vandersall, G. Hon, P. L. Williams

May 2, 2006

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

This work was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

Final Report

Virtual PCR

Investigators:

Shea N. Gardner (Computations)
Email: gardner26@llnl.gov
Tel: 2-4317

David S. Clague (Engineering)
Email: clague1@llnl.gov
Tel: 4-9770

Jennifer Vandersall (Chemistry and Materials Science)
Email: vandersall2@llnl.gov
Tel: 2-3673

Gary Hon (Computations)
Email: hon1@llnl.gov
Tel: 3-6053

Peter L. Williams (Computations)
Email: williams95@llnl.gov
Tel: 2-3832

Tracking Code: 04-ERD-030
Formerly "Electronic PCR"

FY'04 funding: Began in February 2004. $90K BBRP, $50K Engineering
FY'05 funding: $187K NAI, $90K Bioscience, $50K Engineering; total $227K
Project timeline: 1.5 year

Project Overview

The polymerase chain reaction (PCR) stands among the keystone technologies for analysis of biological sequence data. PCR is used to amplify DNA, to generate many copies from as little as a single template. This is essential, for example, in processing forensic DNA samples, pathogen detection in clinical or biothreat surveillance applications, and medical genotyping for diagnosis and treatment of disease. It is used in virtually every laboratory doing molecular, cellular, genetic, ecologic, forensic, or medical research. Despite its ubiquity, we lack the precise predictive capability that would enable detailed optimization of PCR reaction dynamics. In this LDRD, we proposed to develop Virtual PCR (VPCR) software, a computational method to model the kinetic, thermodynamic, and biological processes of PCR reactions. Given a successful completion, these tools will allow us to predict both the sequences and concentrations of all species that are amplified during PCR. The ability to answer the following questions will allow us both to optimize the PCR process and interpret the PCR results: What products are amplified when sequence mixtures are present, containing multiple, closely related targets and multiplexed primers, which may hybridize with sequence mismatches? What are the effects of time, temperature, and DNA concentrations on the concentrations of products? A better understanding of these issues will improve the design and interpretation of PCR reactions.

The status of the VPCR project after 1.5 years of funding is consistent with the goals of the overall project which was scoped for 3 years of funding. At half way through the projected timeline of the project we have an early beta version of the VPCR code. We have begun investigating means to improve the robustness of the code, performed preliminary experiments to test the code and begun drafting manuscripts for publication. Although an experimental protocol for testing the code was developed, the preliminary experiments were tainted by contaminated products received from the manufacturer. Much knowledge has been gained in the development of the code thus far, but without final debugging, increasing its robustness and verifying it against experimental results, the papers which we have drafted to share our findings still require the final data necessary for publication. The following sections summarize our final progress on VPCR as it stands after 1.5 years of effort on an ambitious project scoped for a 3 year period. We have additional details of the methods than are provided here, but would like to have legal protection in place before releasing them.

Project Goals

The result of this project, a suite of programs that predict PCR products as a function of reaction conditions and sequences, will be used to address outstanding questions in pathogen detection and forensics at LLNL. VPCR should enable scientists to optimize PCR protocols in terms of time, temperature, ion concentration, and primer sequences and concentrations, and to estimate products and error rates in advance of performing experiments. Our proposed capabilities are well ahead of all currently available technologies, which do not model non-equilibrium kinetics, polymerase extension, or predict multiple or undesired PCR products. We are currently seeking DHS funding to complete the project, at which time licensing opportunities will be explored,

an updated patent application will be prepared, and a publication will be submitted. A provisional and a full patent application have already been filed (1).

Mission Relevance

VPCR supports LLNL missions in homeland security, Genomes to Life (GtL), and human health. Any field that uses PCR, including bioforensics, biodetection, basic research in GtL, and disease research (e.g. cancer) will benefit. The challenges that we plan to address using VPCR include 1) computational optimization of signatures for pathogen detection, particularly of multiplexed signatures, for a significant savings in cost and time over purely empirical assay optimization, and 2) assessment of signatures for forensic discrimination of closely related sequences.

Approach

Modeling Hybridization and Polymerization

The VPCR tools that we were developing will offer, for the first time, a complete model of the PCR process (Figure 1). Specifically, the kinetics and thermodynamics of DNA denaturization and renaturization, as well as the kinetics of polymerase association and extension, are being modeled for *all* significant reaction pathways that arise during PCR. Our new approach will enable us to predict the spectrum of PCR products given a set of reaction conditions and multiple primer pairs and genomes.

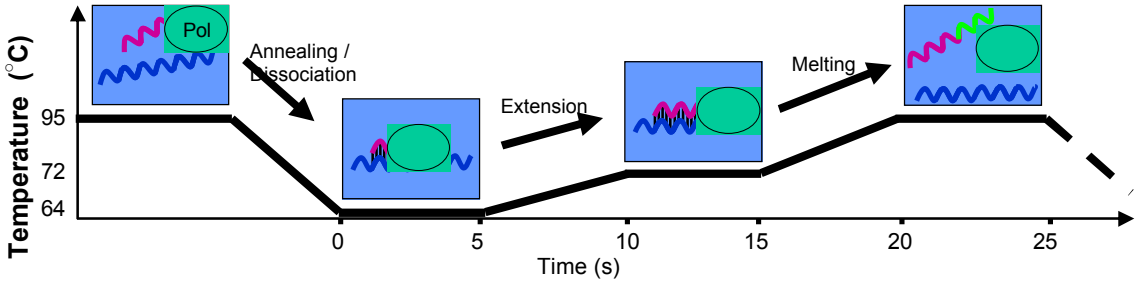

Figure 1: The temperature vs time profile of a single thermocycle. We model 4 types of reactions: annealing, dissociation, polymerase extension, and melting. Although each of these predominantly occurs in a specific temperature regime, they are in constant competition throughout the thermocycle. A typical PCR reaction contains 30-50 thermocycles. The actual temperatures and durations of each soak and ramp may differ from the example protocol shown here, and optimization of each protocol for a given DNA target sequence and set of primers requires trial and error at the bench.

While many programs are available for PCR primer design, they are all limited to selecting a "best" primer or, most recently, examining how reaction conditions affect the equilibrium reactions based only on the first thermocycle (out of 30-50 thermocycles). Numerous bioinfomatics programs such as Primer3, PRIMO, Primer Design, and Oligonucleotide Analyzer return a list of best primers for a specific target based on sequence similarities, primer length, and melting temperatures, using simple GC content equations or the Nearest Neighbor parameters (2). The recently released software Visual OMP goes one step further to predict the concentration of hybridized species for the first thermocycle for a given set of reaction conditions. This software is limited in that it does

not include extension or multiple thermocycles, and it assumes that all reactions proceed to equilibrium at a (single) annealing temperature. To model the dynamic, non-equilibrium conditions that actually comprise a PCR reaction, we developed a novel kinetic approach. As stated, the current state of the art is to estimate PCR products based solely on thermodynamics; however, to gain insight into the dynamics of the process and to optimize cycle processing conditions, we need to employ reaction kinetics.

Traditionally, researchers (3) have described the reaction kinetics using deterministic methods. While these approaches are valuable, it is well known that biochemical processes are stochastic in nature (4). This was further recognized by (5) who applied probabilistic methods to modify rates of reaction to describe the DNA shuffling process. We, however, for the first time, are going a step further to numerically solve the Master Stochastic equation (6) and derive probabilities of competitive events to describe the rates of stochastic events. This approach clearly distinguishes our effort and provides a generalized framework that can be applied to describe the time-evolution of a host of biochemical processes (4). Not only is there a lack of computational tools to model the complete PCR process, but there is also a need for these tools.

The literature provides many of the parameters necessary to develop the proposed comprehensive kinetic and thermodynamic framework, namely, empirically-based free energy parameters for base pair annealing, temperature-dependent rate constants of polymerase extension, and rates of polymerase decay (2,7-8). As these parameters are functions of DNA sequence identities and concentrations, salt concentrations, and temperature, the model will allow us to study the effects that these parameters have on PCR products with detail that is experimentally unfeasible. We are building a new capability that enables product yield information over multiple thermocycles and time-dependent yield within a single thermocycle.

Simulations track all thermodynamically favorable annealing events (i.e. perfect matches and those with mismatches) and extensions to determine the equilibrium distribution of PCR products after each thermocycle. All products present in the solution can cross-react, thus amplifying undesired segments of DNA. The process of iteratively simulating DNA annealing and extension will allow us to track both the sequences and concentrations of all reaction products over multiple thermocycles as a function of reaction conditions. Because we are following all of the possible reaction pathways as they evolve during thermocycles, it is necessary to have a robust method of solving large numbers of highly coupled, non-linear equations. To accomplish this, we have designed algorithms that link to KINSOL, a parallel equation solver developed by CASC researchers at LLNL.

The purpose of the kinetics module is to determine, to single-molecule resolution, the numbers of molecules of each species in the PCR reaction vessel, and to track how these numbers change with time and reaction conditions, e.g., temperature and solution properties. Given a set of possible reactions (as determined by the thermodynamics module), along with the state of the reaction vessel, we first determine the probability of each reaction using our model. Once reaction probabilities are determined, we calculate the actual reactions completed using an algorithm largely derived from Gillespie's tau-leap stochastic method (6,9), incorporating new improvements to the algorithm to reduce simulation times (10). The tau-leaping method has been shown to reduce computational times in the kinetic simulations while giving accurate results.

A significant part of this work involves determining the model(s) to assign probabilities of each event occurring in the PCR reaction process. A complete probabilistic description of primer-to-oligo annealing and denaturation events involves the rate of reaction if the sequences were perfectly matched, attenuated by mismatches. We calculate this reaction probability as the product of two factors: 1) the rate of annealing with perfectly matched base-pairs, 2) the probability of reaction based on the relative free energy of the annealing pathway given the mismatches that may be present (3). The kinetics takes into account changes in processing temperature and fluid conditions within a thermocycle. We are also modeling polymerase activity, based on published experimental data for Taq DNA polymerase (7-8). The simulations randomly assign polymerase molecules to extendable DNA dimers, and polymerization rate depends on temperature. In addition, we are modeling the temperature- and time-dependent degradation of the polymerase enzyme as the PCR reaction proceeds.

Bioinformatics analyses are used to determine the reactants to use as input to the simulations that are relevant to the applications described in the next section, and to examine the products at the completion of all thermocycles to predict whether one would make a false positive or negative call if the computed results were observed in the lab.

Accomplishments

Prototype code for much of the VPCR software has been written (Figure 2), and we have drafted several sections of a manuscript. The code incorporates simulation during the temperature ramps as well as the constant-temperature soaks, competition between multiple (dozens to hundreds) of potential reaction pathways including both hybridization and denaturation (Figure 3), and several alternative formulations of stochastic kinetic simulation algorithms that we have compared and optimized to balance speed and accuracy to handle the complexity of our library of test cases. In addition, we model temperature-dependent extension of DNA and decay of the Taq polymerase enzyme, and track all reaction particles and their concentrations through time. This code is not yet fully debugged and tested against experimental results, so we are not yet ready to write up the results section of the paper in preparation.

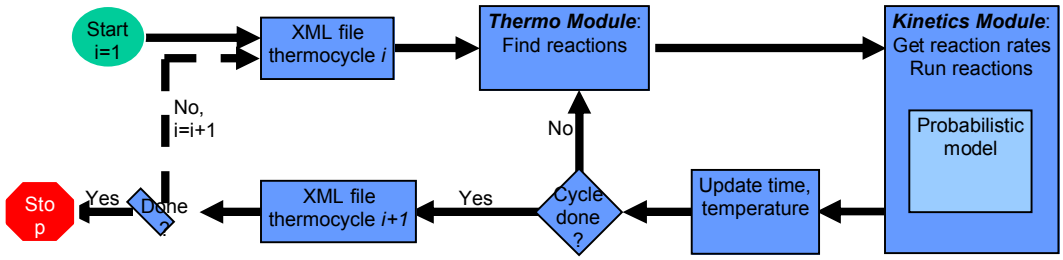

Figure 2: Simplified code overview.. First, based on thermodynamic principles, all possible reactions and their associated rate constants are computed, Then Gillespie's tau-leap trapezoidal stochastic simulation algorithm is used to model the kinetics and track the evolution of the reaction particles. KINSOL is used to solve the resulting nonlinear system of equations at each time step.

A + B ↔ AB (forward primer to genome)
C + D ↔ CD (reverse primer to genome)
B + D ↔ BD (genome to genome)
E + B ↔ EB (probe to genome)
A + A ↔ 2A (primer dimer)
A + C ↔ AC (forward primer to reverse primer)
A ↔ A' (primer hairpin)
AB + E ↔ ABE (probe to primer and genome)

X0 = 8.00 - 9.01e-14 X0 X1 - 9.43e-15 X0 X3 + 1.62e-03 X4
X1 = 2.00 - 9.01e-14 X0 X1 - 1.11e-14 X1 X5 + 1.07e-02 X6
X2 = 9.01e-14 X0 X1
X3 = 2.01e+12 - 9.43e-15 X0 X3 + 1.62e-03 X4 + 5.02e+03 X3 - 4.25e+03 X7
X4 = 3.00 + 9.43e-15 X0 X3 - 1.62e-03 * X4
X5 = ...

Figure 3: Example of the competing hybridization and denaturation reactions, and the resulting nonlinear system of equations.

In 2005, we began to include hairpin kinetics describing DNA secondary structure in the kinetic simulations, but this is not completed. Experiments were performed this summer by a summer student to examine a number of primer sequences at different temperatures and template concentrations, but it appears that the reactions were most likely contaminated, possibly by unwanted DNA from the polymerase enzyme that was delivered to us from the manufacturer (11-12), as even the negative controls yielded signals. We had planned to have the unexpected products sequenced and to redo the experiments with clean DNA, but due to the termination of the project we were not able to wrap up this work.

We do not currently have the funding to debug and verify all of the modules of the code required to generate reliable predictions, and we are seeking ROI and patent protection for additional details regarding the methods. Figure 4 shows some predictions from an early version of the software, illustrating the utility of the code for optimizing the annealing temperature of the PCR protocol to balance the competing tradeoff of higher temperature for higher specificity and lower temperature for higher yield.

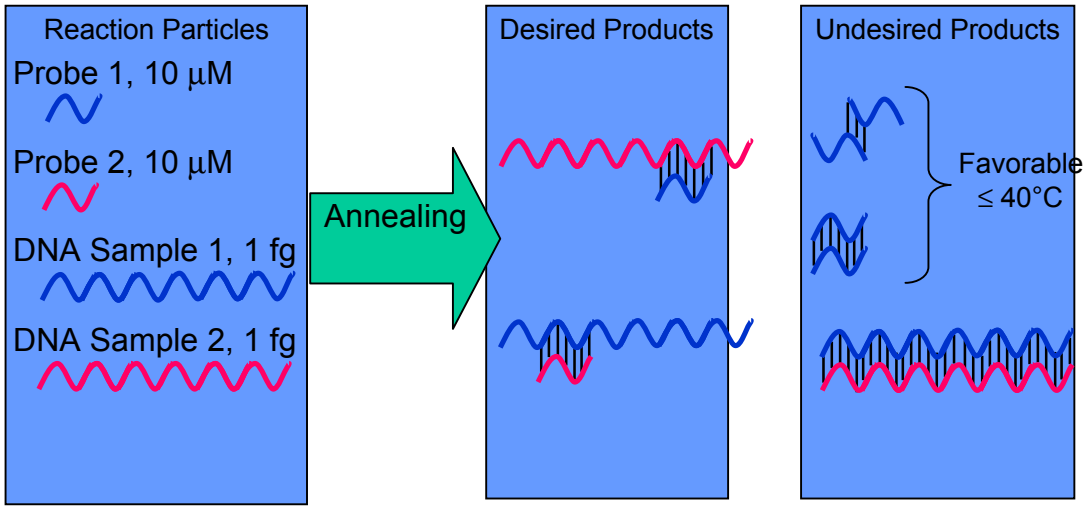

Figure 4A: Example of competing annealing reactions. At lower soak temperatures, annealing reactions are generally more favorable, opening pathways to undesired products.

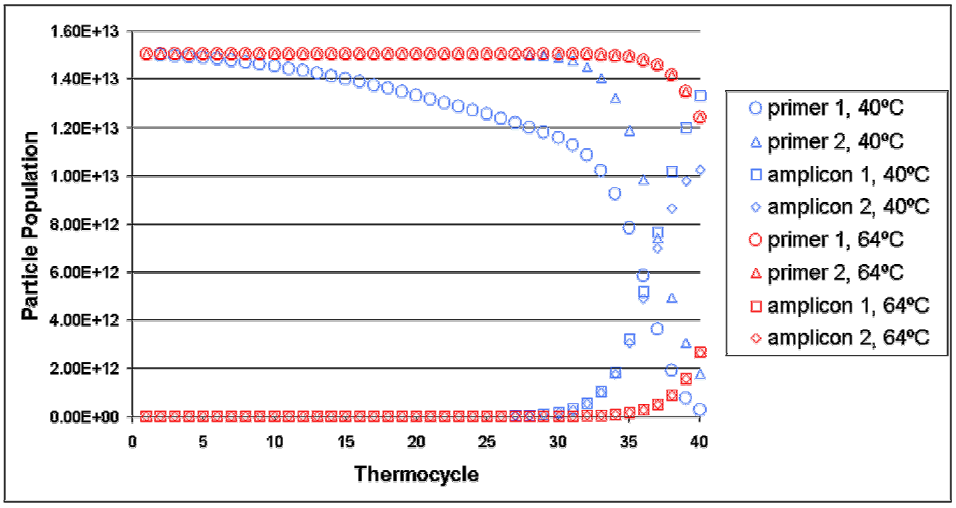

Figure 4B: The system diagrammed in 4A is simulated through 40 thermocycles, at soak temperatures of 40°C and 64 °C. Note that more desired amplicons are output at 40 °C, where pathways to undesired products exist. In this case, the annealing temperature tradeoff favors higher yield at the lower annealing temperature.

Conclusion

While ending prematurely, the VPCR project produced software that is near a patentable and licensable state. Additionally, the accompanying publication is near completion and only requires final revisions to the software and validation with controlled PCR experiments. This novel capability, once completed, will be extremely valuable to the biology community. The team is actively seeking other sources of funding to complete this work.

Acknowledgements

The team would like to acknowledge the LSTO office and the contributing directorates (NAI, Biology and Engineering), for supporting this work.

References

1) Vandersall, JA, Gardner, SN, and Clague, DS. (May 2004) “Computational method and system for modeling, analyzing, and optimizing DNA amplification and synthesis”. Patent application IL-11192. Provisional patent application was filed a year earlier, in May 2003.
2) Allawi, HT and Santalucia, J, Jr. (1997) “Thermodynamics and NMR of internal G.T mismatches in DNA”. Biochem 36, 10581-10594.
3) Wetmur JG. (1991). "DNA probes: applications of the principles of nucleic acid hybridization". Crit Rev Biochem Mol Biol 26, 227-59.
4) McAdams, HH and Arckin, A. (1997) “Stochastic mechanisms in gene expression”. Proc Natl Acad Sci 94, 814-819.
5) Maheshri, N and Schaffer, DV. (2003). “Computational and Experimental Analysis of DNA Shuffling”. Proc Natl Acad Sci USA 100, 3071-3076.
6) Gillespie, D. (1976). "A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions". J Comp Physics 22, 403-434.
7) Innis, MA, Myambo, KB, Gelfand, DH, and Brow, MAD. (1988) DNA sequencing with Thermus aquaticus DNA polymerase and direct sequencing of polymerase chain reaction-amplified DNA. Proc Natl Acad Sci USA 85: 9436-9440.
8) http://www.invitrogen.com/content.cfm?pageid=453
9) Gillespie, D and Petzold LR. (2003). "Improved Leap-Size Selection for Accelerated Stochastic Simulation". J Chem Phys 119, 8229-8234.
10) Cao, Y, Gillespie, DT, and Petzold, LR. (in prep). The adaptive explicit-implicit tau-leaping method with automatic tau selection.
11) Corless, CE, Guiver, M, borrow, R, Edwards-Jones, V., Kaczmarski, EB and Fox, AJ. (2000). Contamination and sensitivity issues with a real-time universal 16S rRNA PCR. J Clin Microbiol 38: 1747-1752.
12) Newsome, T., Li, BJ, Zou, N and Lo, SC. (2004). Presence of bacterial phage-like DNA sequences in commercial Taq DNA polymerase reagents. J Clin. Microbiol 42: 2264-2267.