**Kinetic Modeling and Parameter Estimation of DNA Polymerases**

Discrete time formulation of polymerase processivity

Single hit conditions: enzyme concentrations are sufficiently low that the probability of reassociation \approx 0. Hence once a template dissociates, it never reassociates.

Let i index the sequence positions on the template.

Let p denote the conditional probability of the polymerase –not- dissociating at position/time I. The probability of dissociation at position/time i is



Plot p\_off(i) vs i; log p is slope. p is called the microscopic processivity parameter.

Expected position of dissociation of the polymerase (expected stopping time; sometimes called the processivity).



* Above assumes infinite length template. Long templates used in processivity experiments to estimate p.
*  is sometimes reported as the processivity instead of the microscopic processivity parameter
* For finite length, 
* The (time-independent) theory of processivity gives no absolute indication as to how processivity affects reaction time for extension under multihit conditions
* Dissociation is said to be (much) more likely during the translocation phase of extension (see below), due to only weak electrostatic interactions maintaining polymerase-template binding. Reference: On the Processivity of Polymerases.

PETER H. VON HIPPEL, FREDERIC R. FAIRFIELD, AND MARY KAY DOLEJSI.

* For heterogeneous templates, p will vary with position. From processivity experiments, one can obtain the p at each position since we will have  for all i. The equations can be used to solve uniquely for each pi. However, it is impractical (for the same reason identified in annealing kinetics) to do processivity experiments for each new template. Hence, one can do processivity experiments on templates with different types of nearest neighbor motifs (including hairpins) for a given polymerase, and then nearest neighbor processivity parameters can be used in modeling of an arbitrary sequence.
* If one prefers to use a single p to simplify modeling, p can be obtained from the first moment of the sequence length distribution for the template, according to. A single p model is also preferred if one is estimating p from time series (nonequilibrium) data; see below. Single p’s are often reported for polymerase extension on poly-A sequences.
* Equilibrium processivity theory plays a role not unlike that of equilibrium thermodynamic modeling of DNA hybridization, in that is provides a system of equations that can be used along with additional constraints from initial or transient rate data to estimate kinetic parameters.
* *Literature data on Taq processivity at 72 C: E[ioff]=22* (reference: Wang et al., A novel strategy to engineering DNA polymerases for enhanced processivity and improved performance. Nucl. Acids Res. 32: 1197-1207).

**Continuous time formulations of polymerase processivity**

The goal of this work is to obtain a predictive model for the transient dynamics of polymerase-mediated extension. Previous reports on thermostable polymerases have not presented a methodology for determining the average extension time or “minimal extension time”, meaning the time at which a particular probability of conversion to full-length DNA is achieved. Prior data has indicated only the rate of change of the total number of nucleotides incorporated. Data reported has indicated rate of change of total number of nucleotides. We are interested in



where P(n|0,t) denotes the probability of conversion to full length DNA in time t, conditional on the initial state being primer-template hybrid, and     indicates that the unconditional probability of observing full-length DNA at time t is equal to a specified (objective) value of the probability.

We present an experimental methodology as well as theoretical/computational methods for solving the problem. These methods can ultimately be applied to any type of sequencing problem, including but not limited to PCR.

We consider two CT systems that can produce the observed distribution of partially extended DNAs observed in processivity experiments: a model without a polymerase translocation step after base addition (i.e., polymerase moves forward at the same time it adds a base), and a model with a polymerase translocation step after base addition. The former provides an estimate of k-1, whereas the latter does not; but in the latter, the value of k-1 does not significantly affect the transient dynamics of extension. Each should be assessed in terms of its consistency with processivity experiments as well as time series extension experiments under standard (PCR) conditions.

Single hit conditions, no translocation

-Single hit conditions are used to estimate polymerase processivity parameters (with an appropriate kinetic model, associated rate constants or equations relating the rate constants)

- Assumptions: a) E + Di -> E.Di is omitted from the model. Justification: single hit conditions do not allow enzyme reassociation. Hence enzyme-template association occurs only during initial equilibration of enzyme with SP.

- Here, dissociation is assumed to occur with equal conditional probability (equal rate) at all times according to the reaction scheme:

We denote the enzyme dissociation rate constant by k-1 because it may not be the same dissociation constant relevant to processivity experiments (see below). Let k = , and apply the steady state approximation for intermediates (validity considered elsewhere). Omit  from state space model for simplicity; does not affect equilibrium and we are not estimating it.







 : conditional probability of transition from E.Di -> E.Di+1 assuming system did not remain in state E.Di at time t+dt (i.e., the polymerase advanced by one base during time dt)

Conditional probability of multiple transitions in time dt = 0.

Sequence heterogeneity can be accommodated by expressing each k-1 above in terms of a measured position-dependent pi.



* Possible issues with formulation: *literature is unclear on whether distribution of partially extended sequences produced during extension under single hit conditions depends on salt concentration as well as nucleotide concentration.* Formulation above requires processivity to change with [N]. Some literature indicates that processivity strongly depends on salt concentration, due to polymerase binding during translocation being mediated primarily by electrostatic interactions.
* Advantage: provides an absolute k-1(T).
* Solve for k-1 from processivity data for Taq at 72C, calculate k1 using Keq(72C) and simulate under MM conditions. Check accuracy of predictions against time series data at 72C.
* Conditions for processivity experiments are being investigated by CJ. Some literature indicates dissociation occurs during translocation, otherwise do not allow for dissociation during translocation
* Note the equations can be generalized accommodate a heterogeneous sequence with different k-1’s at every position, estimated via processivity data.

Transition probability matrix for predicting concentrations of partially completed sequences at any time:



Single hit conditions, without translocation, without steady state assumption

The simplified state space model above, which uses the steady state assumption for nucleotide addition, is degenerate and can be solved analytically with concomitant advantages. It would be justified under certain conditions as follows:

If kcat >> k-2, kcat/Kn [N] \approx k2[N] and if kcat >> k2[N], kcat/Kn [N] model is accurate, since intermediate forms slowly and is rapidly converted. We have only Kn, kcat and [N] from MM experiments. Assuming kcat >> k-2, Kn/[N] >> 1 would improve steady state model accuracy. However, we don't have data on k-2 for thermostable polymerases of interest. This approximation (which is useful for parameter estimation but not necessarily accurate in later stages of reaction where transient model is important) is another potential source of errors in the time series prediction.

*Hence we should test the predictive accuracy of the simplified steady state model, starting at low [N].* If inaccurate, we can estimate k2 using time series data and available processivity data at 72C (see below). (CJ is looking up literature values for some of these rate constants for other polymerases to help determine accuracy of approximation.)



This provides an expression relating k-1, k2,k-2, kcat/Kn to processivity parameter.

In addition to the experimentally known processivity parameter, we have Ke(T) = k1/k-1, and Kn(T) = (k-2 + kcat) / k2. There is one additional unknown parameter compared to the model with the steady state assumption. See below for more details on parameter estimation.

[To be completed after testing steady state models]

Single hit conditions, with translocation

- Assumptions: a) E + Di <-> E.Di is omitted from the model. Justification: single hit conditions do not allow enzyme reassociation. Hence enzyme-template association occurs only during initial equilibration of enzyme with SP. b) (E.Di)’ dissociation to E+Di, but not E.Di dissociation to E+Di, is included. Justification: under single hit conditions, [N] apparently does not affect equilibrium distribution of products. Since E.Di dissociation would depend on kcat/Kn[N] as shown above, it cannot contribute appreciably compared to the dissociation of (E.Di)’. Hence it is not included below to simplify the model, but it can be included with a minor modification.

- Due to the above change in the dissociation model, we use k-e to denote the dissociation rate constant during translocation.

****



Omit  from state space model for simplicity; given that MM experiments do not provide an estimate for this parameter

With appropriate choice of k-e, kt (see below), E.Di’ disappears rapidly, validating assumption . Then, omit the associated state vector components from the state space model (not possible in a control setting), since those components are not of interest.

Dissociated DNA appears rapidly in the Di channels

Conditional probability of E.Di -> E.Di+1 transition in time dt: 

Conditional probability of E.Di -> E + Di+1 transition in time dt: 

 : conditional probability of transition from E.Di -> E.Di+1 assuming system did not remain in state E.Di at time t+dt (i.e., the polymerase advanced by one base during time dt). Does not depend on [N].

* Choose any  >>  (for the highest [N] that arises in PCR) to justify assumption .



Multihit conditions, with translocation. PCR stage 1 (significant enzyme excess)

* Use for prediction of time series data from CJ’s experiments
* Assumptions: E + Di associates to form E.Di, but only (E.Di)’ dissociates to form E.Di. Justification: under single hit conditions, [N] apparently does not affect equilibrium distribution of products. Since E.Di dissociation would depend on kcat/Kn[N] as shown above, it cannot contribute appreciably compared to the dissociation of (E.Di)’. Hence it is not included below to simplify the model, but it can be included with a minor modification.



Omit  from transient model for simplicity, given that MM experiments do not provide estimate for this parameter

No information about parameters can be obtained from equilibrium data under multihit conditions. However, time series data can be used to help estimate parameters, as discussed below.

This system with degenerate eigenvalues can be solved analytically; we can solve for minimal extension time without numerical decomposition or simulation. Multihit systems without translocation have nondegenerate eigenvalues. The two systems display important differences in their dynamics. These differences should be detectable in time series data.

MM equations for parameter estimation of models with/without translocation

(MM equations only apply under special multihit conditions.)

*Assumptions of extension kinetics MM model*

Equation (10) below describes the steady state approximation, which is used in the MM derivation of (5):



This assumption could be validated through solution of the model with trial values of the on/off rate constants k2 and k-2, that it is a common result in MM kinetics assuming that substrate (nucleotide) concentration is sufficiently high (as it typically is in,e.g., PCR reactions). See below for further details.

Also,



The time over which (11) holds (during which the initial rate measurements must be made) can be investigated through the use of a full “state space” model of the extension reaction network.

**Steady state assumptions should be validated by simulation.** Such tests are important to this paper due to the (long) structure of the reaction network, which is unconventional in MM. See below.

*No translocation:*

****



(The above can also be shown via conservation of mass and the state space matrix for the full extension network, along with the steady state assumptions regarding intermediates)

Simpler derivation, but only valid under assumption :

Let



Since the concentrations of the two substrates, Di and N, are high as typically assumed in MM kinetics formulations, a steady state approximation for the intermediates can be applied (see Supporting Information) and hence,



Equation (3) is valid under the approximation. (By contrast, in a standard ordered bireactants model ([17](#_ENREF_17)), the product is assumed to form during the initial rate measurement. The differences between the polymerase extension MM formulation and the standard ordered bireactants model is considered further in the Supporting Information.)

By analogy to standard MM kinetics, the initial rate is given by



where KN denotes the Michaelis constant for the nucleotide addition reaction and Keq,1 denotes the equilibrium constant of the enzyme binding reaction.

Substituting the definitions of [S1], [S2], and the associated enzyme complexes into the expressions for v and [E] above, the sum of the rates of conversion of each E.Di, into E.Di+1, i=1,…,n, which is equal to the rate of nucleotide substrate depletion, can then be written



In the initial stage of reaction:

 (6)

In equation (6) [N]0 is the total concentration of dNTP; [SP]0 is the total concentration of substrate/primer complex. Non-linear regression of v0 versus [N]0 would result in kcat and apparent Michaelis constant which equals to .

Additionally, kcat and Keq,1 can be determined given an estimate of KN by plotting 1/v against 1/[S1]= at constant [N]. We find that according to the relationship

,

the slope of the double reciprocal plot is  and the intercept is .

*Steady-state analysis with translocation*

In the multihit translocation model, the following revised steady state calculation must be used for MM kinetics. (Note that this steady state analysis includes both dissociation during translocation and dissociation after translocation.)



Thus  should replace Keq,1 in the bireactants MM model in this case. k-1 may be omitted for the reasons cited above, or it may be written in terms of Keq,1(T) and k1.  can be obtained from processivity data. Thus by varying [N]\_0 at constant [SP]\_0 and varying [SP]\_0 and constant [N]\_0, in two sets of initial rate experiments, we can solve the resulting equations simultaneously for  and .

If processivity data is not available (e.g. at a given temperature for Taq), one can use time series data along with MM initial rate data to solve for k1 and  (see below).

The steady state assumptions for each model can be checked numerically through simulation.

Expected time required for addition of n nucleobases (full extension)

In stage 1, the expected time required for full length extension of a given template starting from SP hybrid can be obtained from E[t]=\int\_0^\infty t \* P\_n0(t) in the multihit model.

Otherwise, for any specified initial probability vector p(0), one can solve for t for a desired probability of full-length extension (fraction of templates fully extended).

**Models and parameters:**

* *Single hit conditions, no translocation*

Parameters: k-1, kcat/Kn

Solve for k-1 in terms of kcat/Kn using experimental processivity data and relation between k-1, kcat/Kn, and p above.

Obtain kcat,Kn from MM experiments (using Keq(T) from literature).

From k-1 and Keq(T) for enzyme binding, obtain k1 for use in multi hit simulation.

Hence if processivity data is available, all parameters can be identified; otherwise, time series estimation is required.

* *Single hit conditions, no translocation, no steady state assumption*

Parameters: k-1, k2, k-2,k2, kcat/Kn

Solve for k-1 in terms of kcat/Kn, k2, and k-2 using experimental processivity data and relation between these parameters and p above. Apply Kn = (k-2+ kcat)/k2 to eliminate one unknown.

Obtain kcat,Kn from MM experiments (using Keq(T) from literature in MM equations).

From k-1 and Keq(T), obtain k1 in terms of k2. Use the resulting expressions for parameters in terms of k2 in multi hit simulation/estimation below. Time series estimation is required to identify all parameters.

* *Single hit conditions, with translocation*

Experimental processivity data does not provide unique value of k-e, but allows

us to choose relative values for k-e and kt that will determine the transient dynamics

k-1 can be omitted from model. Cannot use Keq(T) to relate k1, k-e.

From bireactant MM experiments, obtain two equations in the unknowns k1, kcat/Kn, and k-e.

Hence if processivity data is available (hence k-e was selected), all parameters can be identified; otherwise, time series estimation is required.

* *Multi hit conditions, no translocation*

In absence of processivity data, we must estimate one parameter via time series estimation.

* *Multi hit conditions, no translocation, no steady state assumption*

If use of kcat/Kn is not a good approximation, must estimate one additional parameter for transient analysis of nucleotide addition step.

With processivity data, we must estimate one parameter via time series estimation. For example, we can estimate k2 using time series data and available processivity data.

Without processivity data, we must estimate two parameters via time series estimation.

* *Multi hit conditions, with translocation*

In absence of processivity data, we must estimate one parameter via time series estimation.

Proposed estimation procedure:

The Km1,app, Kn, and kcat are estimated using MM data and equations above. (As shown above, Km1,app is approximately equal to Keq for enzyme binding in models without translocation, but depends on the processivity parameter in models with translocation.) Then use time series estimation to estimate remaining parameters (or vice versa, whichever is easier for uncertainty analysis).

*Single parameter time series estimation:*

Consider as an example the multi hit model with translocation. kcat/Kn, k1, p appear in state space model. Using MM estimation, all rate constants in the state space model would be expressed in terms of one (eg kcat/Kn) that is then estimated using time series data. Eg solve for k1,p in terms of kcat/Kn using the MM eqns then solve for kcat/Kn (which will appear in some form in each function of time in state space soln) using time series data.

It may be shown that for a single unknown parameter in the multihit model, and a specified value of the total DNA concentration at a specified time, a unique solution exists for the parameter estimate. This will be demonstrated later. The parameter estimates could be compared to those from processivity at a temperature (eg 72 C) where processivity data is available. These approaches to the use of time series data for dynamic estimation can also lead to more rapid characterization of polymerases, which is especially important for thermostable polymerases.

Start by getting the estimates for Km1,app, Kn, kcat from MM and then use the resulting estimates for all parameters in terms of one unknown parameter that appears in state space model (e.g., kcat/Kn) without their uncertainties in the time series moment condition. Solve for kcat/Kn using the time series condition and assign its uncertainty using nonlinear least squares. Then solve for k1,p and kcat,Kn estimates and uncertainties algebraically using the MM equations.

*Next steps*

Note that the MM estimation method depends on validity of both steady state approximations in initial stage of reaction. Simulation may be used to test 2nd steady state assumption now, but only if we choose trial values of the rate constants k2,k-2 (alternatively, we could do the estimation of k2 above first, then test assumption by simulation, but this could waste time). Such tests are important to this paper due to the (long) structure of the reaction network, which is unconventional in MM.

KM may test the various MM assumptions by simulation and carry out time series estimation. (CJ is currently doing MM estimation – a) in procedure above). First, use processivity data at 72 C with the models that do not require estimation of additional parameters to test the predictions of these models and compare to experimental data. Consider parameter estimation for model without steady state assumption for nucleotide addition if the other models cannot accurately predict experimental time series data.