A simplified mathematical model of directional DNA site-specific recombination by serine integrases

Alexandra Pokhilko1, Jia Zhao1, W. Marshall Stark1, Sean D. Colloms1 and Oliver Ebenhoh2

1Institute of Molecular, Cell and Systems Biology, University of Glasgow, Glasgow G12 8QQ, UK
2Cluster of Excellence on Plant Sciences (CEPLAS), Heinrich-Heine-University, Universitaetsstrasse 1, 40225 Dusseldorf, Germany

Serine integrases catalyse site-specific recombination to integrate and excise bacteriophage genomes into and out of their host’s genome. These enzymes exhibit remarkable directionality; in the presence of the integrase alone, recombination between attP and attB DNA sites is efficient and irreversible, giving attL and attR products which do not recombine further. However, in the presence of the bacteriophage-encoded recombination directionality factor (RDF), integrase efficiently promotes recombination between attL and attR to re-form attP and attB. The DNA substrates and products of both reactions are approximately isoenergetic, and no cofactors (such as adenosine triphosphate) are required for recombination. The thermodynamic driving force for directionality of these reactions is thus enigmatic.

Here, we present a minimal mathematical model which can explain the directionality and regulation of both ‘forward’ and ‘reverse’ reactions. In this model, the substrates of the ‘forbidden’ reactions (between attL and attR in the absence of RDF, attP and attB in the presence of RDF) are trapped as inactive protein–DNA complexes, ensuring that these ‘forbidden’ reactions are extremely slow. The model is in good agreement with the observed in vitro kinetics of recombination by dC31 integrase, and defines core features of the system necessary and sufficient for directionality.

1. Introduction

Serine integrases catalyse integration of a circular bacteriophage genomic DNA molecule into the bacterial host chromosomal DNA, by recombination between an attP site in the phage DNA and an attB site in the host DNA. In the resulting ‘lysogenic’ state, the phage genome is integrated in the host genome, and is flanked by recombinant attL and attR sites, each consisting of an attP half and an attB half (figure 1a). To resume its replicative life cycle, the phage DNA must be excised from its bacterial host genome. To accomplish this, a phage-encoded recombination directionality factor (the RDF protein) is expressed together with the integrase protein. RDF interacts with integrase and alters its properties so that it recombines the attL and attR sites to release the circular phage genomic DNA with an attP site, and leave an attB site in the host genome (figure 1a). (Hereafter, we refer to recombination between attP and attB as P × B recombination, and recombination between attL and attR as L × R recombination.)

Serine integrase-mediated recombination can be reconstituted in vitro, using purified integrase, RDF and DNA substrates [2–5]. The in vitro reactions strikingly reproduce the directionality observed in vivo. P × B recombination is efficient (i.e. most of the substrate molecules are recombined) when only integrase protein is present, and the reaction is unidirectional (i.e. no L × R recombination is observed.

© 2017 The Authors. Published by the Royal Society under the terms of the Creative Commons Attribution License http://creativecommons.org/licenses/by/4.0/, which permits unrestricted use, provided the original author and source are credited.
Figure 1. Schematic representations of serine integrase-mediated recombination reactions. (a) Overview, showing integrase (I) reacting to convert a PB (attP + attB) substrate to LR (attL + attR) product in the absence of RDF (red), whereas integrase plus RDF (I + R) converts LR (attL + attR) to PB (blue). (b) Scheme of reaction steps in Model M. Blue and red solid arrows show integrase-catalysed steps with and without RDF respectively. PB, PBIR1, PBIR2, LR1, LR2, LRIR are complexes containing four molecules of the integrase protein with PB or LR DNA substrates (with or without four molecules of RDF). The \( P \times B(-R) \) reaction starts with the binding of four molecules of integrase (I) to PB substrate, followed by a recombination step and formation of the final product LR1. The \( L \times R(+R) \) reaction starts with the binding of four molecules of an integrase–RDF complex (IR) to LR substrates, followed by a recombination step and formation of the final product PBIR1. The ‘forbidden’ \( P \times B(+R) \) and \( L \times R(-R) \) reactions form ‘blocked’ LR1 and PBIR1 complexes, which delay recombination due to their very slow conformational change to the productive LR1 and PBIR1 complexes (grey dotted arrows). The favourable directions of reaction steps are shown by big arrowheads. Steps names are shown near arrows. The cartoons show hypothetical structures of intermediates; note that alternative structures might be involved. See [1] for further details; coiled-coil domains of integrase shown by yellow sticks (or by blue sticks in presence of RDF). (c) Core structure of reactions after reduction of the fast variables. Substrates and products are shown by cartoons. Very slow steps are indicated by grey lines.

under these conditions). However, when RDF is also present, \( L \times R \) recombination is efficient and most of the sites are converted to attP and attB products. In addition to stimulating the \( L \times R \) reaction, the presence of RDF inhibits the \( P \times B \) reaction. No high-energy cofactors (such as adenosine triphosphate) are needed for recombination, and the unbound DNA substrate and product molecules are expected to be approximately isoenergetic. The molecular basis of the thermodynamic ‘driving force’ that favours \( P \times B \) recombination in the absence of RDF, but \( L \times R \) recombination in the presence of RDF, is unknown.

Serine integrases have recently attracted much attention as potential tools for experimental and applied genetic manipulations, because of their recombination efficiency, their short DNA recombination sites (att sites) (typically 40–50 bp) and absence of host factor requirements [5,6]. Mathematical and biochemical analysis of recombination directionality in these systems is therefore timely. We recently presented a detailed mathematical model of recombination by \( \Phi C31 \) integrase (the first serine integrase to be identified, and the best-characterized to date) [5,7,8], which aimed to account as far as possible for the available biochemical, molecular and structural data [1] (here called ‘Model A’). This model comprises 35 ordinary differential equations (ODEs). Although it provides a good match to the \textit{in vitro} kinetics data, its complexity makes it difficult to identify the key steps that determine directionality. Additionally, the large number of parameters in our previous model complicates analysis of their specific effects on reaction kinetics. We were therefore motivated to create a highly simplified data-driven mathematical model of serine integrase-mediated recombination. Such a minimal model with a simple structure and minimal number of parameters should be useful in analysis of the key principles of unidirectional reversible genetic transformations, which might be applicable to other biological systems.

Here, we present our simplified model, which consists of only three ODEs and assumes that all other steps of the reaction are in rapid quasi-equilibrium. This simplified model clearly illustrates the key theoretical assumptions required for the directionality and regulation of recombination by serine integrases. Moreover, owing to its simplicity, this model is more generally applicable and is easily adaptable to other integrase-mediated recombination systems.

2. Model description

The minimal model (referred to hereafter as ‘Model M’) was fitted to \textit{in vitro} experimental data on recombination by \( \Phi C31 \) integrase (I) with its RDF gp3 (R) [1] (figure 2). In the experiments used to produce these data, the extent of recombination at different concentrations of integrase and RDF was determined after 3 h reactions. Recombination was intramolecular, between attP and attB sites (PB) or attL and attR sites (LR) in inverted repeat orientation on supercoiled plasmid substrates. Recombination inverts the orientation of the DNA sequence flanked by the att sites, but the size of the plasmid is unchanged. Model M therefore assumes intramolecular recombination of a PB or LR plasmid substrate, but it would be easily adaptable for other types of substrates. The reaction scheme is a simplification of that used in the previously reported full model, Model A [1] (figure 1b). Model M consists of a set of reversible reaction steps, each described by first or second order kinetics.

The \( P \times B(-R) \) reaction starts by binding of four molecules of I to the PB substrate (two molecules to each att site) and formation of the PB synapse (PBI), in which the attP and attB sites in the PB plasmid are held together by an integrase tetramer. Model M describes the formation of PBI synapse from free PB substrate as a single binding step (step ‘b1’), although the process requires multiple molecular events, which were represented as two steps (b1 and s1) in our earlier Model A. This simplification was based on our earlier assumption that integrase binding is faster than synopsis and therefore is not a rate-limiting step [1]. The next step (recombination ‘r1’)

\[ L \times R(-R) \rightarrow IR \]

represents the formation of the integrase–RDF complex, which can then associate with the PB substrate (step ‘b2’). This reaction is rate-limiting in the presence of RDF, and the reaction is very slow in its absence. The ‘forbidden’ reaction \( P \times B(+R) \) occurs with a very low rate, allowing the reaction to proceed in the desired direction. The reaction \( L \times R(-R) \) is also rate-limiting, and the reaction proceeds in the desired direction. The reaction \( L \times R(+R) \) occurs with a very low rate, allowing the reaction to proceed in the desired direction. The reaction \( L \times R(-R) \) is also rate-limiting, and the reaction proceeds in the desired direction.
transforms PB1 to LR1, comprising recombinant attL and attR sites bound together by an integrase tetramer which synapses the two sites (figure 1b). Here again, this single step describes a series of molecular events (r1 and mod in Model A). Even though the rate of step mod in Model A is unknown, it was reasonable to assume that it is faster than the recombination step r1, which includes multiple complex changes in DNA state, such as cleavage, rotation and strand exchange [1,5,6]. Model M proposes that LR1 is the predominant endpoint of the P × B (−R) reaction at short reaction times (of the order of 3 h), because the next step on the pathway is very slow (see below).

In Model M, the ‘forbidden’ attL × attR reaction in the absence of RDF (L × R (−R)) starts with the binding of four I molecules to LR (step b2 in figure 1b), forming the LR1 complex. Crucially, the LR1 complex is conformationally distinct from LR1, the immediate product of P × B recombination. Conversion of LR1 into LR1 (step ‘syn’) and its reverse are assumed to be very slow reactions, of the order of days (as in Model A where it is referred to as step s2). A hypothetical structure-based interpretation of these two complexes has been presented [1] (figure 1c). The very slow interconversion of LR1 and LR2 complexes is the key feature that explains why the L × R (−R) reaction does not yield detectable levels of PB recombination product in short reactions (of the order of hours).

In Model M, the reaction of the LR substrate in the presence of RDF (L × R (+R)) starts by binding of integrase–RDF complexes to the attL and attR sites. We assume (based on published data) that each integrase monomer interacts in solution with one RDF monomer to form a 1:1 complex (IR) [1,3,9]. Therefore, four IR complexes bind to LR, forming a synaptic complex LRIR (figure 1b, step ‘b3’ (steps b3 and s3 in Model A)). Synapsis is followed by strand exchange step ‘r2’ (steps r2 and modr in Model A), forming a PB product synapse comprising IR-bound recombinant attP and attB sites, PBIR1 (figure 1b). Analogously to our hypothesis for P × B (−R) recombination (see above), we propose that PBIR1 is the typical endpoint of the L × R (+R) reaction. In the ‘forbidden’ P × B (+R) reaction, we assume that a different complex PBIR2 is formed by binding of four IR complexes to PB, and that conversion of PBIR2 to PBIR1 (step ‘synr’) and its reverse (equivalent to s4 in Model A) are very slow.

Model M also includes the possible formation of unproductive complexes which contain a synaptic integrase tetramer, but fewer than four RDF molecules (PBIR, LRIR; not shown on figure 1b, for clarity). For simplicity, the model only includes one representative version of each species, one for LR and one for PB, each containing two RDF and four integrase monomers. These two complexes correspond to multiple unproductive complexes in the full model (Model A) and are assumed to be completely unproductive for recombination. The inclusion of complexes PBIR, and LRIR, in Model M was sufficient to describe the sharp response of the reactions to small changes in the ratio of RDF:integrase protein concentrations when this ratio is close to 1 (figure 2).

We assume that all reaction steps except recombination (r1, r2) and the slow synaptic conformational change steps (syn,
the independent slow dynamic variables, with total complexes as functions of free \( I, R, \) and \( LR \)-integrase–RDF complex (required to be lower than three unproductive complexes, with one, two and three mol–RDF per four molecules of integrase (in Model A there were \( K_{LRi} \) total concentrations of integrase (\( I_{\text{tot}} \)), RDF (\( R_{\text{tot}} \)) and DNA (\( D_{\text{tot}} \)). Applying the rapid equilibrium approximation yields expressions for the concentrations of the equilibrating complexes as functions of free \( I, R, LR \) and \( PB \) as follows:

\[
\begin{align*}
[I] &= \frac{[I] \cdot [R]}{K_{IR}}, \\
[PB] &= \frac{[I]^4 \cdot [PB]}{K_{PB}}, \\
[LR_1] &= \frac{[I]^4 \cdot [LR]}{K_{LR}}, \\
[LRIR] &= \frac{[IR]^4 \cdot [LR]}{K_{IR}}, \\
[PBIR_2] &= \frac{[IR]^4 \cdot [PB]}{K_{PB}}, \\
[PBIR_3] &= \frac{[I]^2 \cdot [IR]^2 \cdot [PB]}{K_{PBIR_3}}, \\
[LRIR_2] &= \frac{[I]^2 \cdot [IR]^2 \cdot [LR]}{K_{LRIR_2}}, \\
[LRIR_3] &= \frac{[I]^2 \cdot [IR]^2 \cdot [LR]}{K_{LRIR_3}},
\end{align*}
\]

where \( K_{IR}, K_{PB} \) and \( K_{LR} \) are the dissociation constants for the respective complexes.

For simplicity and because it was sufficient to describe the data (figure 2), we assume that all of the complexes of integrase–RDF (or integrase and RDF) with DNA (complexes formed in steps \( b_1, b_2, b_3, b_4 \)) and also unproductive complex \( PBIR \) have the same dissociation constant \( K_{PB} \). The dissociation constant \( K_{LR} \) for the unproductive DNA complex \( LRIR \) was required to be lower than \( K_{PB} \). This is a consequence of the simplified assumption that there is only one type of unproductive LR-integrase–RDF complex (LRIR), with two molecules of RDF per four molecules of integrase (in Model A there were three unproductive complexes, with one, two and three molecules of RDF). A low value of \( K_{LR} \) prevents formation of \( PB \) product in the \( L \times R (+) \) reaction when RDF concentrations are lower than integrase, in agreement with the data [1], which show sharply reduced \( L \times R (+) \) recombination when RDF is lower than integrase (figure 2).

The total concentrations of integrase and RDF can be expressed as the sum of the concentrations of all complexes containing these species

\[
[I_{\text{tot}}] = [I] + [IR] + 4 \cdot ([PB] + [LRIR] + [LR_1] + [LR_2] + [PBIR_3] + [PBIR_2] + 2 \cdot ([LRIR] + [PBIR]). \quad (2.8')
\]

and

\[
[R_{\text{tot}}] = [R] + [IR] + 4 \cdot ([LRIR] + [PBIR_3] + [PBIR_2] + 2 \cdot ([LRIR] + [PBIR]). \quad (2.9')
\]

While correct expressions for \([I] \) and \([R] \) can in principle be derived, their analytic forms are highly complex. We therefore make the approximation that the concentrations of DNA-bound integrase and RDF species are negligible. Thus considering only unbound integrase and RDF

\[
[I_{\text{tot}}] \approx [I] + [IR] = [I] + \left[\frac{[R]}{K_{IR}}\right]. \quad (2.8)
\]

and

\[
[R_{\text{tot}}] \approx [R] + [IR] = [R] + \left[\frac{[I]}{K_{IR}}\right]. \quad (2.9)
\]

This approximation is justified by the experimental observations that the concentrations of integrase and RDF required for efficient recombination are typically much higher (more than 10-fold) than the concentration of the DNA substrate. For example, in our experimental conditions with DNA concentrations of 10 mM, integrase concentrations above 200 nM were required for efficient recombination [1]. Thus, even if every plasmid binds four integrase molecules (two molecules to each \( att \) site), the free integrase pool is reduced by only 20%.

From the above equations, \( R \) is easily expressed via \( I_{\text{tot}} \) and \( R_{\text{tot}} \)

\[
[R] = [R_{\text{tot}}] - [IR] = [R_{\text{tot}}] - [I_{\text{tot}}] + [I]. \quad (2.10)
\]

Inserting equation (2.10) into equation (2.8) leads to the quadratic equation

\[
[I]^2 + [I] \cdot ([I_{\text{tot}}] - [I_{\text{tot}}] + K_{IR} - [I_{\text{tot}}] \cdot K_{IR} = 0,
\]

resulting in

\[
[I] = 0.5 \cdot \left( \sqrt{b^2 + 4 \cdot [I_{\text{tot}}] \cdot K_{IR} - b} \right); \quad (2.11)
\]

\[
b = [R_{\text{tot}}] - [I_{\text{tot}}] + K_{IR}.
\]

The total concentration of \( PB \) can be expressed as a sum of the concentrations of all species containing \( PB \)

\[
[PB_{\text{tot}}] = [PB] + [PB] + [PBIR_3] + [PBIR_2] + [PBIR].
\]

After expressing the concentrations of all quickly changing variables via \([I], [R], [PB] \) (equations (2.1)–(2.7)), we can get the following expression for \([PB] \):

\[
[PB] = \frac{[PB_{\text{tot}}] - [PBIR]}{1 + ([I]^4/K_{IR}) + ([IR]^4/K_{IR}) + ([IR]^2 \cdot [IR]^2/K_{PBIR})}. \quad (2.12)
\]

Similarly, the total concentration of DNA (\( D_{\text{tot}} \)) can be expressed as the sum of concentrations of all PB and all LR containing species

\[
[D_{\text{tot}}] = [LR] + [LRIR] + [LR_1] + [LR_2] + [LRIR_2] + [LRIR_3] + [PBIR_{\text{tot}}],
\]

leading to the equation

\[
[LR] = \frac{[D_{\text{tot}}] - [PBIR_{\text{tot}}] - [LR_1]}{1 + ([I]^4/K_{IR}) + ([IR]^4/K_{IR}) + ([IR]^2 \cdot [IR]^2/K_{LRIR})}. \quad (2.13)
\]

The kinetics of the slowly changing variables (the concentrations of \( LR_1, PBIR_3, PBIR_{\text{tot}} \)) are governed by the following
three ODEs:

\[
\frac{d[LR_1]}{dt} = k_{a1} \cdot [PBI] - k_{b1} \cdot [LR_1] + k_{-syn} \cdot [LR_2] - k_{syn} \cdot [LR_1],
\]

(2.14)

\[
\frac{d[PBIR]}{dt} = k_{a1} \cdot [LRIR] - k_{b1} \cdot [PBIR] + k_{-syn} \cdot [PBIR_2] - k_{syn} \cdot [PBIR_1]
\]

(2.15)

\[
\frac{d[PB_{tot}]}{dt} = k_{a1} \cdot [LR_1] - k_{b1} \cdot [PBI] + k_{a1} \cdot [LRIR] - k_{b1} \cdot [PBIR].
\]

(2.16)

The parameters \( k_{a1} \), \( k_{b1} \) and \( k_{syn} \) are respectively the forward and reverse rate constants of the strand exchange steps \((r1, r2)\) and modification steps \((syn, synr)\), with the forward direction defined as \( PB \to LR \) for \((-R)\) and \( LR \to PB \) for \((+R)\) reactions. The forward rate constants of steps \( r1 \) and \( r2 \) are assumed to be equal \((k_{a1}, k_{b1})\), for simplicity. This assumption resulted in a good fit to our data (figure 2). Equilibrium constants \( K_{a1}, K_{b1}, K_{syn} \) are calculated as the ratios of forward and reverse rate constants:

\[
K_{a1} = \frac{k_{a1}}{k_{b1}}; \quad K_{b1} = \frac{k_{b1}}{k_{a1}}; \quad K_{syn} = \frac{k_{syn}}{k_{syn}}.
\]

All concentrations are expressed in micromolars; the time units are hours.

Equations (2.1)–(2.16) comprise the complete set of equations describing the dynamics of the system. Concentrations were set to reflect the conditions used experimentally (figure 2). Thus, the total DNA concentration DNA_{tot} was 10 nM, while integrase and RDF concentrations were varied.

Model M includes 10 parameters (the seven forward and reverse rate constants from differential equations (2.14)–(2.16), the DNA binding dissociation constants \( K_{d1}, K_{d2}, K_{d3} \), and the dissociation constant of IR, \( K_d \) presented in electronic supplementary material, table S1. These parameters were chosen to fit the experimental data [1]. In particular, the equilibrium constants of the strand exchange steps \( K_{a1}, K_{b1} \) were fitted to the observed extent of recombination after 3 h of the \( P \times B(-R) \) and \( L \times R(+R) \) reactions under saturated integrase (and RDF for \( L \times R(-R) \) reaction) concentrations. The rate constant of the slow step ‘syn’ \((k_{syn})\) was fitted to the observed extent of recombination after 3 h of the ‘forbidden’ \( P \times B(+R) \) reaction (figure 2). For the rate constant of the ‘syn’ step \((k_{syn})\), only the upper bound was estimated based on the observed absence of recombination after 3 h of the ‘forbidden’ \( L \times R(-R) \) reaction (figure 2). The rate constants of strand exchange steps were fitted to the observed rates of recombination (see the Results section). Finally, the equilibrium constants \( K_{syn} \) and \( K_{syn} \) were determined from the conservation of energy during the \( P \times B \) and \( L \times R \) reactions, which requires that the product of all equilibrium constants of the reactions equals 1 according to Wegscheider’s condition [10]:

\[
\frac{K_{a1} \cdot K_{syn} \cdot K_{d1}}{K_{d1}} = 1; \quad \frac{K_{a2} \cdot K_{syn} \cdot K_{d2}}{K_{d2}} = 1, \text{ leading to }
\]

\[
K_{syn} = \frac{1}{K_{a1}}; \quad \text{and } K_{syn} = \frac{1}{K_{a2}}.
\]

The presence of these constraints results in a balance between energetically favourable and non-favourable steps. For example, the favourable binding of integrase to PB substrate and formation of the \( LR_1 \) product of the \( P \times B(-R) \) reaction is balanced by the unfavourable transition from \( LR_1 \) to \( LR_2 \) and dissociation of integrase from \( LR_1 \) (figure 1b).

The system of ODEs was solved using Matlab, integrated with the stiff solver ode15s (The MathWorks UK, Cambridge). Matlab code of the main model is provided in electronic supplementary material, text S1 and is freely available at https://github.com/ (the full URL is https://github.com/QTB-HHU/integrateModel/tree/master/minimalModel).

3. Results and discussion

Our Model M is based on the simplified reaction scheme of figure 1b. Several steps in this scheme are formed by combining a number of steps in Model A [1]. For example, the formation of the substrate synapse \( PBI \) from free \( PB \) and four integrase molecules is considered as a single step in Model M, combining intermediate steps (binding of integrase monomers and/or dimers to individual att sites, and synapsis). In addition, the presence of the relatively slow recombination and desynapsis steps \((r1, r2, syn \text{ and } synr)\) allows us to apply rapid equilibrium approximations to the other reactions, further reducing the dimensionality of the model (figure 1c). As a result Model M has only three ODEs for the independent slowly changing variables, describing concentrations of \( LR_1, PBIR \) and conversion of \( PB \) to \( LR \) by the recombination steps (equations (2.14)–(2.16)), with other faster changing variables being expressed as functions of these slow changing ones. The initial rates of the ‘allowed’ recombination reactions are determined by the rates of the recombination steps (figure 3a,b), while the final approach to equilibrium is determined by very slow changes in the conformations of the product complexes, which might be related to slow desynapsis as proposed in [1] (figure 1b).

Despite the much reduced complexity of Model M compared to Model A, the key feature of different, slowly interconverting, tetrameric integrase complexes formed from either \( PB \) or \( LR \) substrates is preserved. The complex \( LR_1 \) is formed after strand exchange during the \( P \times B(-R) \) reaction, whereas the complex \( LR_2 \) is formed when integrase is added to the ‘naked’ \( LR \) plasmid \((L \times R(-R) \text{ reaction; figures } 1b \text{ and } 3b,c))\). Similarly, the complex \( PBIR \) is formed after strand exchange during the \( L \times R(+R) \) reaction, whereas the complex \( PBIR_2 \) is formed upon binding of integrase–RDF to the \( PB \) plasmid \((P \times B(+R) \text{ reaction; figures } 1b \text{ and } 3b,d))\). The interconversion of \( LR_1 \) and \( LR_2 \) is assumed to be very slow, as is interconversion of \( PBIR_1 \) and \( PBIR_2 \) (figure 3). These slow interconversions, which may be interpreted as indicating a high activation energy barrier, allowed us to describe directionality (as in Model A). Thus, the ‘allowed’ reactions \( P \times B(-R) \) and \( L \times R(+R) \) quickly approach quasi-equilibrium, with the accumulation of \( LR_1 \) and \( PBIR_1 \) complexes respectively, which can equilibrate only very slowly with the \( LR_2 \) and \( PBIR_2 \) complexes (figure 3a,b). The ‘forbidden’ reactions \( L \times R(-R) \) and \( P \times B(+R) \) are predicted to be far from equilibrium for long times, due to initial accumulation of non-recombinant, catalytically inactive \( LR_2 \) and \( PBIR_2 \) complexes respectively, which convert very slowly to the catalytically active \( LR_1 \) and \( PBIR_1 \) complexes (figure 3c,d; note the logarithmic time scale). This trapping of the substrates of ‘forbidden’ reactions in the blocked \( LR_2 \) and \( PBIR_2 \) complexes makes the reactions practically irreversible in vivo, because such complexes will usually be destroyed by cellular
Figure 3. Simulated time courses of the amounts of abundant DNA-containing products. Panels (a,b) show the kinetics of the ‘allowed’ reactions \((P \times B(−R))\) and \((L \times R(+R))\), whereas panels (c,d) display the kinetics of the ‘forbidden’ reactions. Solid red and blue lines show the total amount of LR and PB plasmid molecules, respectively. The dotted and dashed lines show the amounts of integrase–DNA complexes as indicated at the left in each panel. The amounts of all DNA-containing species are expressed relative to the total DNA content. The calculations were performed for 10 nM of plasmid substrate, 400 nM of integrase and 800 nM of RDF (for panels (b) and (d)).

processes such as DNA replication before they can convert to \(LR_1\) or \(PBIR_1\) and undergo recombination.

Model M provides a good match to the in vitro experimental data on the kinetics of recombination by \(\phi C31\) integrase and its RDF gp3 [1] (figure 2). In these experiments, the levels of recombinant products were measured after 3 h of the \(P \times B\) and \(L \times R\) reactions under different concentrations of integrase and RDF (figure 2a,c). The simulations with Model M quantitatively describe the key features of the data, such as the observed sharp stimulation of the \(L \times R(+)\) reaction (figure 2a,b) and inhibition of the \(P \times B(−)\) (figure 2a,b) reaction when the concentration of RDF reaches that of integrase. These effects are predicted to result from competition between RDF-containing and RDF-free complexes of integrase for binding to DNA substrates. Thus, LR substrate forms the blocked \(LR_1\) complex when integrase is present in the absence of RDF, but increasing RDF concentration shifts the balance towards the productive \(LRIR\) complex (figures 1 and 2c,b). Similarly, PB substrate forms the productive \(PBIR\) complex in the absence of RDF, but the blocked (RDF-containing) \(PBIR_1\) complex becomes predominant as RDF concentration is increased (figure 2a,b).

Certain minor features of our experimental data, which were accounted for in the full model (Model A), are not described by Model M, due to its simplicity. These features include the observed slight decrease of the maximal amount of recombination products when integrase concentrations are raised to higher than 200 nM (figure 2). This was described in Model A by the inclusion of unproductive complexes of integrase tetramers bound at single recombination sites. Also, Model M does not describe the observed small deviation of the initial kinetics of the reactions from simple exponential kinetics (electronic supplementary material, figure S1a). The observed two-exponential kinetics is better described by Model A (electronic supplementary material, figure S1b) because two steps (synapsis and recombination) limit the faster and slower exponentials respectively.

As described above, our model predicts that two distinct LR products \((LR_1\) and \(LR_2)\) are formed depending on whether integrase is added to \(PB\) or \(LR\) DNA. However, these different products have not yet been detected.

Figure 4. Simulated effect of addition of RDF (R) to the \(P \times B(−)\) reaction after 1 h (solid line). The kinetics of the \(L \times R(+R)\) reaction is shown for comparison by a dotted line. The computations were performed with 10 nM plasmid DNA, 400 nM integrase and 800 nM RDF. (Online version in colour.)
The computations were performed for 10 nM substrate, 400 nM integrase and 800 nM RDF (for the 'recombination' steps (\(k_{\text{f}}\)) or slow steps (\(k_{\text{b}}\), \(L \times R\) reaction, \(\beta\)). \((c-f)\) Dependence of 'half-time' (\(T_{1/2}\)) on the rate constant of the 'recombination' steps (\(k_{\text{f}}\) or \(k_{\text{b}}\)) for permitted reactions (\(P \times B\), (c) and \(L \times R\), (d)) or on the rate constants of the conformational changes (\(k_{\text{syn}}\) or \(k_{\text{synr}}\)) for non-permitted reactions (\(L \times R\), (e) and \(P \times B\), (f)). Changes in \(k_{\text{f}}\) or \(k_{\text{b}}\) were accompanied by compensating changes in the equilibrium constants of slow steps (\(K_{\text{syn}}\) or \(K_{\text{synr}}\) respectively), to maintain agreement with the energy conservation equations. Changes in the rate constants (\(k_{\text{f}}\), \(k_{\text{b}}\), \(k_{\text{syn}}\) and \(k_{\text{synr}}\)) were accompanied by equal changes in the reverse rate constants (\(k_{\text{f}}\), \(k_{\text{b}}\), \(k_{\text{syn}}\) and \(k_{\text{synr}}\)) to keep the equilibrium constants unchanged for these steps. The computations were performed for 10 nM substrate, 400 nM integrase and 800 nM RDF (for \(b,d,f\)).

Figure 5. Effects of changes in Model M parameters on the key characteristics of integrase reactions. Simulations with unchanged parameters are indicated by dotted lines. \((a,b)\) Dependence of extent of recombination at equilibrium on the equilibrium constants of 'recombination' steps, \(K_{\text{f}}\) (\(P \times B\), (–R) reaction, \(a\)) or \(K_{\text{b}}\) (\(L \times R\), (–R) reaction, \(b\)). \((c-f)\) Dependence of 'half-time' (\(T_{1/2}\)) on the rate constant of the 'recombination' steps (\(k_{\text{f}}\)) for permitted reactions (\(P \times B\), (c) and \(L \times R\), (d)) or on the rate constants of the conformational changes (\(k_{\text{syn}}\) or \(k_{\text{synr}}\)) for non-permitted reactions (\(L \times R\), (e) and \(P \times B\), (f)). Changes in \(k_{\text{f}}\) or \(k_{\text{b}}\) were accompanied by compensating changes in the equilibrium constants of slow steps (\(K_{\text{syn}}\) or \(K_{\text{synr}}\) respectively), to maintain agreement with the energy conservation equations. Changes in the rate constants (\(k_{\text{f}}\), \(k_{\text{b}}\), \(k_{\text{syn}}\) and \(k_{\text{synr}}\)) were accompanied by equal changes in the reverse rate constants (\(k_{\text{f}}\), \(k_{\text{b}}\), \(k_{\text{syn}}\) and \(k_{\text{synr}}\)) to keep the equilibrium constants unchanged for these steps. The computations were performed for 10 nM substrate, 400 nM integrase and 800 nM RDF (for \(b,d,f\)).

experimentally. Model M makes a prediction that could allow us to test experimentally our hypothesis for the directionality of recombination. According to the model, when integrase is added to \(PB\) for 1 h, most of the DNA will form the \(LR\) product. The addition of RDF at this point leads to reversal of the recombination step and rapid equilibration with the energetically more favourable \(PBIR\) complex. This leads to accumulation of \(PB\) DNA to levels that transiently approach 100% (figure 4; solid line), by the reverse of the \(P \times B\) (–R) pathway. By contrast, if integrase is added to \(LR\) to form \(LRIR\), addition of RDF after 1 h leads to formation of \(PBIR\) via the \(L \times R\) pathway. This reaction never produces more than 74% \(PB\) product (figure 4; dotted line). Such high levels of \(PB\) product in a RDF-mediated reaction (near 100%) have not been observed to date in reactions catalysed by dC31 integrase, and would provide new evidence for our model for directionality.

We next used our model to explore the effects of different reaction steps on the overall kinetics. For practical applications, the recombination efficiency (that is, the maximum extent of conversion of substrate to product) is especially important; this can vary dramatically between different integrases [11]. Our analysis demonstrates that the reactions are most critically affected by the equilibrium constants of the \(r1\) and \(r2\) steps (\(K_{\text{f}}\) and \(K_{\text{b}}\)), which combine various steps including DNA strand exchange and the subsequent modifications. Variation of \(K_{\text{f}}\) or \(K_{\text{b}}\) results in large changes in recombination efficiencies (figure 5a,b), suggesting that variations in the efficiencies of different integrases might be primarily due to differences in these steps. Another important characteristic of an integrase system is its reaction rate. Our model predicts that the rates of the ‘allowed’ \(P \times B\) (–R) and \(L \times R\) (–R) reactions should be critically dependent on the rates of the recombination steps (rate constant \(k_{\text{f}}\); figure 5c,d). The rates of ‘forbidden’ reactions \(L \times R\) (–R) and \(P \times B\) (–R) are strongly dependent on the rate constants of the conformational change \(k_{\text{syn}}\) and \(k_{\text{synr}}\) (figure 5e,f).

4. Conclusion

The minimal model (Model M) for integrase-mediated DNA recombination that we have presented above is able to account for the puzzling directionality of these systems. Model M is much simpler than our previously reported model (Model A), with only three independent variables governed by ODEs,
but it is sufficient to quantitatively describe the in vitro experimental data on the kinetics of recombination by dC31 integrase and its RDF gp3. The model explains the observed directionality by the formation of stable synaptic complexes as end products of the ‘allowed’ recombination reactions (with or without RDF; LR1 and PBIR1; figure 1), and kinetically stable inactive complexes with the substrates of ‘forbidden’ reactions (LR12 and PBIR2; figure 1). This might represent a mechanism by which phages avoid spontaneous ‘reversal’ of integration and excision reactions.

Our minimal model emphasizes the key features of integrase-mediated reactions that bring about directionality, and might serve as a paradigm for those studying other biological systems with directional properties.

Authors’ contributions. A.P., W.M.S., S.D.C. and O.E. designed computational analysis; A.P. performed computational analysis; W.M.S. and S.D.C. designed experiments; J.Z. performed experiments; A.P., W.M.S., S.D.C. and O.E. wrote the paper.

Competing interests. We declare we have no competing interests.

Funding. This research was funded by Biotechnology and Biosciences Research Council BB/K003356/1. Deutsche Forschungsgemeinschaft, Cluster of Excellence on Plant Sciences CEPLAS (EXC 1028) to O.E.

References