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 Ebenhöh2

1Institute of Molecular, Cell and Systems Biology, University of Glasgow, Glasgow G12 8QO, UK
2Cluster of Excellence on Plant Sciences (CEPLAS), Heinrich-Heine-University, Universitätsstraße 1, 40225 Düsseldorf, 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 φC31 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...
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 φC31 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 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 in vitro experimental data on recombination by φC31 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) 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 $P \times B(\neg R)$ 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 synapsis and therefore is not a rate-limiting step [1]. The next step (recombination ‘r1’)
from LR

Model M proposes that state, such as cleavage, rotation and strand exchange [1,5,6].

step reasonable to assume that it is faster than the recombination of RDF (see below).

LR Conversion of attL RDF complexes to the productive for recombination. The inclusion of complexes attL and attR 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 PBIR1 and LIR1, 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 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 PBIR1 and LIR1, 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,
synr) are fast. This assumption allows us to reduce the dimensionality of the model by applying rapid equilibrium approximations to the other steps, including formation of integrase–RDF complex IR in solution and binding of integrase (with or without RDF) to DNA. After these simplifications, four slowly changing variables (the concentrations of LRI, PBIRs, total PB and total LR) remain, of which only three are independent, because of conservation of the total DNA pool. We chose the three quantities LRI, PBIR and total PB as the independent slow dynamic variables, with total LR being determined by the difference between total DNA and total PB.

The concentrations of rapidly equilibrating species (PB, PBI, PBIR, LR, LRI, LRIR, and PBIR) can be analytically determined from the slowly changing variables along with the conserved total concentrations of integrase (Itot), RDF (Rtot) and DNA (Dtot). 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:

\[
[I] = \frac{[I] \cdot [R]}{K_{ir}},
\]

\[
[PB1] = \frac{[I]^2 \cdot [PB]}{K_{ir}},
\]

\[
[LRI] = \frac{[I]^2 \cdot [LR]}{K_{lr}},
\]

\[
[LRIR] = \frac{[I]^4 \cdot [LR]}{K_{lr}^2},
\]

\[
[PBIR_2] = \frac{[I]^4 \cdot [PB]}{K_{ir}^2},
\]

\[
[PBIR] = \frac{[I]^2 \cdot [IR^2 \cdot [PB]}{K_{lr}},
\]

and

\[
[LRIR] = \frac{[I]^2 \cdot [IR^2 \cdot [LR]}{K_{lr}^2},
\]

where \(K_{ir}, K_{lr} \) 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 b1, b2, b3, b4, and also unproductive complex PBIR) have the same dissociation constant \(K_{lr} \). The dissociation constant \(K_{lr} \) for the unproductive DNA complex LRIR was required to be lower than \(K_{lr} \). 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

\[
[I]_{tot} = [I] + [IR] + 4 \cdot ([PB1] + [PBIR_1] + [PBIR_2]) + 2 \cdot ([LRIR] + [PBIR]),
\]

and

\[
[R]_{tot} = [R] + 4 \cdot ([LRIR] + [PBIR_1] + [PBIR_2]) + 2 \cdot ([LRIR] + [PBIR]).
\]

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

\[
[I]_{tot} \approx [I] + [IR] = [I] + \frac{[I] \cdot [R]}{K_{ir}},
\]

and

\[
[R]_{tot} \approx [R] + [IR] = [R] + \frac{[I] \cdot [R]}{K_{ir}}.
\]

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 \(I_{tot}\) and \(R_{tot}\)

\[
[R] = [R]_{tot} - [IR] = [R]_{tot} - [I]_{tot} + [I].
\]

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

\[
[I]^2 + [I] \cdot ([I]_{tot} - [I]_{tot} + K_{ir}) - [I]_{tot} \cdot K_{ir} = 0,
\]

resulting in

\[
[I] = 0.5 \cdot \left( \sqrt{b^2 + 4 \cdot [I]_{tot} \cdot K_{ir} - b} \right);
\]

\[
b = [R]_{tot} - [I]_{tot} + K_{ir}.
\]

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

\[
[PB_{tot}] = [PB] + [PB1] + [PBIR_1] + [PBIR_2] + [PBIR].
\]

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 \([PB]\):

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

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

\[
[D_{tot}] = [LR] + [LRIR] + [LRIR_1] + [LRIR_2] + [LRIR],
\]

leading to the equation

\[
[LR] = \frac{[D_{tot}] - [PB_{tot}] - [PBIR]_1 - [PBIR_2]}{1 + ([I]^4/K_{ir}) + ([IR]^4/K_{ir}) + ([IR]^2 \cdot [IR]/K_{lr})}.
\]

The kinetics of the slowly changing variables (the concentrations of LRIR, PBIR, PBIR) are governed by the following
three ODEs:
\[
\frac{d[LRI_1]}{dt} = k_{r1} \cdot [PBIR] - k_{s1} \cdot [LRI_1] + k_{-s} \cdot [LRI_2]
\]
\[
\frac{d[PBI]}{dt} = k_{s1} \cdot [LRI] - k_{r1} \cdot [PBIR] + k_{-s} \cdot [LRI_2]
\]
\[
\frac{d[PBIR]}{dt} = \frac{k_{r1} \cdot [LRI] - k_{r2} \cdot [PBIR] + k_{-s} \cdot [LRI_2]}{-k_{s2} \cdot [PBIR]}
\]
(2.14)
(2.15)

and
\[
\frac{d[PBIR]}{dt} = k_{s1} \cdot [LRI_1] - k_{s2} \cdot [PBIR] + k_{-s} \cdot [LRI]
\]
(2.16)

The parameters \( k_{s1}, k_{s2}, k_{-s}, k_{-r1}, k_{-r2}, k_{-s}, k_{-s2} \) 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 → LR for (−R) and LR → PB for (+R) reactions. The forward rate constants of steps r1 and r2 are assumed to be equal (\( k_{s1} = k_{s2} \)), for simplicity. This assumption resulted in a good fit to our data (figure 2). Equilibrium constants \( K_{syn} \) and \( K_{synr} \) were determined from the observed extent of recombination after 3 h of reactions (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 LRI is formed after strand exchange during the PB × B(−R) reaction, whereas the complex LRI is formed when integrase is added to the ‘naked’ LR plasmid (L × R(−R) reaction; figures 1b and 3a,c). Similarly, the complex PBIR is formed after strand exchange during the LR × R(+R) reaction, whereas the complex PBIR is formed upon binding of integrase–RDF to the PB plasmid (P × B(+R) reaction; figures 1b and 3b,d). The interconversion of LRI and LRI is assumed to be very slow, as is interconversion of PBIR and PBIR (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 × B(−R) and L × R(+R) quickly approach quasi-equilibrium, with the accumulation of LRI and PBIR complexes respectively, which can equilibrate only very slowly with the LRI and PBIR complexes (figure 3a,b). The ‘forbidden’ reactions L × R(−R) and P × B(+R) are predicted to be far from equilibrium for long times, due to initial accumulation of non-recombinant, catalytically inactive LRI and PBIR complexes, respectively, which convert very slowly to the catalytically active LRI and PBIR complexes (figure 3c,d; note the logarithmic time scale). This trapping of the substrates of ‘forbidden’ reactions in the blocked LRI and PBIR complexes makes the reactions practically irreversible in vivo, because such complexes will usually be destroyed by cellular

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 synapsis 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 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 independently slowly changing variables, describing concentrations of LRI, 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).
visually for the kinetics of recombinase by PB (RDF-containing) complex (figures 1 and 2 for panels (c,d) 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 LR1 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 2). Similarly, PB substrate forms the productive PBIR complex in the absence of RDF, but the blocked (RDF-containing) PBIR complex becomes predominant as RDF concentration is increased (figure 2).

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 (LR1 and LR2) are formed depending on whether integrase is added to PB or LR DNA. However, these different products have not yet been detected in the experimental data, such as the observed sharp stimulation of the LRIR reaction (figure 2) and inhibition of the PBIR reaction (figure 2).}

**Figure 3.** Simulated time courses of the amounts of abundant DNA-containing products. Panels (ab) show the kinetics of the ‘allowed’ reactions (P × B (–R) and L × R (+R)), whereas panels (cd) 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)).

**Figure 4.** Simulated effect of addition of RDF (R) to the P × B (–R) reaction after 1 h (solid line). The kinetics of the L × 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 (L × R(+) reaction, (d)) or K_r2 (L × R(+) reaction, (b)). (c–f) Dependence of 'half-time' (T_1/2) (the time required to reach 50% of the maximum attainable product level) on the rate constant of the 'recombination' steps (k_rr) for permitted reactions (P × B(-R), (c) and L × R(+)R, (d)) or on the rate constants of the conformational changes (k_syn for permitted reactions (L × R(+)R, (e) and P × B(+)R, (f)). Changes in K_r1 or K_r2 were accompanied by compensating changes in the equilibrium constants of slow steps (K_syn or K_synr respectively), to maintain agreement with the energy conservation equations. Changes in the rate constants (k_syn and k_synr) were accompanied by equal changes in the reverse rate constants (k_−syn and k_−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 LR1 product. The addition of RDF at this point leads to reversal of the recombination step and rapid equilibration with the energetically more favourable PBIR2 complex. This leads to accumulation of PB DNA to levels that transiently approach 100% (figure 4; solid line), by the reverse of the P × B(−R) pathway. By contrast, if integrase is added to LR to form LR2, addition of RDF after 1 h leads to formation of PBIR1 via the L × R(+)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_r1 and K_r2), which combine various steps including DNA strand exchange and the subsequent modifications. Variation of K_r1 or K_r2 results in large changes in recombination efficiencies (figure 5c,d), 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 × B(−R) and L × R(+)R reactions should be critically dependent on the rates of the recombination steps (rate constant k_rr; figure 5c,d). The rates of 'forbidden' reactions L × R(−R) and P × B(+)R are strongly dependent on the rate constants of the conformational change k_−syn and k_−synr (figure 5f).

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 (LR2 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


