

Development and Experimental Validation of a Mechanistic Model of *in vitro* DNA Recombination*

Jack Bowyer¹, Jia Zhao², Susan Rosser³, Sean Colloms² and Declan Bates¹

Abstract—Engineering cellular memory is a key area of research in which Synthetic Biology has already begun to make significant impacts. Recent work elucidating transcriptional memory devices has paved the way for the creation of bi-stable genetic switches based on DNA recombination. Attempts to experimentally design and build synthetic systems using recombinases have thus far been hindered by a lack of validated computational models that capture the mechanistic basis of DNA recombination. The predictive capabilities of such models could be exploited by Synthetic Biologists to reduce the number of iterative cycles required to align experimental results with design performance requirements. Here, we develop and validate the first detailed mechanistic model of DNA recombination, with a focus on how efficiently recombination can occur, and the model features required to replicate and predict experimental data.

I. INTRODUCTION

Synthetic Biology is a highly interdisciplinary field with the aim of establishing engineering protocols for the construction of synthetic biological circuits. Natural systems of genetic parts such as promoters and repressors can be rearranged in order to produce desired transcriptional responses to transient stimuli. One of the first synthetic biological devices to be built in this way was the genetic toggle switch in *E. coli* [1]. The toggle switch exhibits bistability, outputting two distinct transcriptional states, ‘on’ and ‘off’, presenting immense potential in biocomputing [3]. Recent experimental developments have seen a movement towards the use of site-specific recombinases (SSRs) to effectively manipulate genetic material. SSRs mediate distinct recombination events that give rise to two stable DNA states. Hence, DNA recombination not only has huge potential in biological data storage [2], but also in medicine since the efficacy of many diseases is dependent upon the inheritance of cellular states. Engineering mammalian cellular memory in ways that permit a programmed response to harmful stimuli has numerous possible medical applications, such as T cell therapy for cancer patients [4]. The standardisation of both bacterial DNA recombination and the conditions required to induce switching events that are both highly efficient and stable

over significant time frames will be the first steps towards realising the human benefit of this research.

SSRs can enable precise DNA manipulation both *in vitro* and *in vivo* [8]. These enzymes, known as integrase and excisionase, catalyse two recombination events termed integration and excision. Integrase alone is sufficient to mediate the integration reaction between an attachment site encoded within the host chromosome, attB, and an attachment site on the phage chromosome, attP. The phage genome is inserted into the host chromosome and is flanked by the newly formed attachment sites attL and attR. The excision reaction is mediated in the presence of integrase and a recombination directionality factor (RDF) also known as excisionase, restoring the independent substrates as well as the original attB and attP sites (Fig. 1). A pre-integration state consisting of attB and attP is referred to as the BP state, whereas a post-integration/pre-excision state containing attL and attR is known as the LR state. DNA recombination efficiency in switching between these two states is dependent on the concentrations of integrase and RDF in the system.

II. A MECHANISTIC MODEL OF DNA RECOMBINATION

An extensive review of the experimental literature was carried out to synthesise current knowledge of the mechanistic basis of recombination, see [5], [6], [7] and references

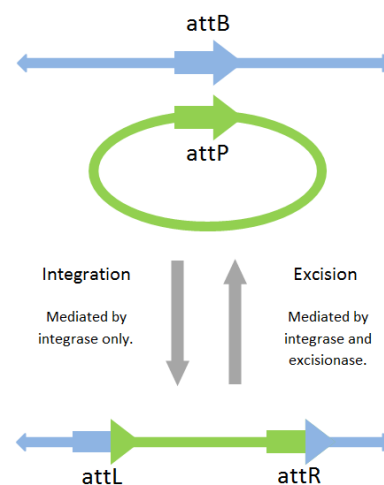


Fig. 1. Schematic diagram of phage integration and excision. The phage genome attachment site, attP, is integrated into the host chromosome attachment site, attB. Integration gives rise to attL and attR, each formed of half of attB and attP. Excision restores attB and attP, removing the integrated phage genome from the host chromosome.

*Research supported by EPSRC via a DTA studentship to J. Bowyer

¹J. Bowyer and D. Bates are with the Warwick Integrative Synthetic Biology Centre and School of Engineering, University of Warwick J.E.Bowyer@warwick.ac.uk, D.Bates@warwick.ac.uk

²J. Zhao and S Colloms are with the Institute of Molecular, Cell and Systems Biology, University of Glasgow j.zhao.1@research.gla.ac.uk, Sean.Colloms@glasgow.ac.uk

³S. Rosser is with Synthetic and Systems Biology, School of Biological Sciences, University of Edinburgh Susan.Rosser@ed.ac.uk

therein. Many structural features of the DNA recombination reaction network have widespread experimental support in the literature. Integrase forms dimers in solution [9] with one dimer bound to attB and attP necessary to mediate the integration reaction [10], which is unidirectional [11]. RDF does not bind directly to DNA attachment sites [12]. One integrase dimer and an additional RDF monomer bound to both attL and attR is necessary to mediate excision, restoring attB and attP [13]. However, we were unable to find consensus regarding three significant biological details:

- The directionality of the excision reaction.
- RDF dimerisation and subsequent tetramerisation in solution.
- Integrase monomer binding to DNA substrates.

By constructing a variety of models based on the well established properties of the reaction network that differed only according to the inclusion/exclusion of these three features, we were able to examine which model best matched our experimental data and hence investigate the most plausible representation of the true biological structure (Fig. 2). We found that including unidirectional excision [14], monomeric RDF that neither dimerises nor tetramerises in solution [15], and integrase monomer binding [16], resulted in optimal data fitting and prediction capabilities in our model. These results were determined via global optimisation of the parameters for each candidate model against the data (see Results section). Early results revealed that our models consistently produced greater recombination efficiency than that observed in our data. Hence we included a biologically plausible mechanism for the sequestration of integrase: the formation of a dysfunctional integrase dimer at a rate k_{intX} (Fig. 2, top). Integrase is necessary to mediate both integration and excision, therefore fewer functional integrase dimers present within the system decreases the efficiency of both reactions. As a result, overall recombination efficiencies are reduced and the model is able to replicate the data with greater accuracy.

Our modelling investigation focuses on DNA recombination mediated by ϕ C31 serine integrase and its RDF, gp3, *in vitro*. The network consists of twenty-two distinct biological entities and twenty-eight reaction constants. The rate of change in concentration of each entity is governed by a corresponding ordinary differential equation (ODE) in the model. We apply mass action kinetics to the biological interactions in order to derive the ODEs that comprise the model. Analysing the efficiency with which the system switches from BP state to LR state involves determining the total register of the system in either state. This equates to summing each of the ODEs that govern the network entities in the DNA state of interest (see Fig. 5 for the full list of ODEs). Model simulations are therefore the numerical solutions to the following equations:

$$\frac{d[D_{LRtot}]}{dt} = k_R[D_{BP}I_4] - k_R[D_{LR}I_4R_2]. \quad (1)$$

$$\frac{d[D_{BPtot}]}{dt} = -k_R[D_{BP}I_4] + k_R[D_{LR}I_4R_2], \quad (2)$$

where D_{LRtot} , D_{BPtot} represent the total concentration of DNA in LR state and BP state respectively, $D_{BP}I_4$ represents the concentration of the protein:DNA complex consisting of four integrase monomers bound to DNA in the BP state, $D_{LR}I_4R_2$ represents the concentration of the protein:DNA complex consisting of four integrase monomers and two excisionase monomers bound to DNA in the LR state and k_R represents the rate of recombination.

Experiments performed *in vitro* allow the initial concentrations of the recombinases and substrates to be quantified exactly and therefore provide the initial conditions with which to solve the system of ODEs. Hence, we do not require knowledge of integrase and excisionase expression levels, as was necessary for the simple model of *in vivo* recombination dynamics developed in [2].

III. RESULTS - EXPERIMENTAL VALIDATION

We compared the ability of our model to replicate and predict *in vitro* experimental data with a simple mathematical model of DNA recombination previously developed in Bonnet *et al.* [2]. Using a subset of our steady state data for a variety of different initial concentrations of integrase and gp3, the parameters of both models were optimised using a Genetic Algorithm (GA) to minimise the difference between simulations and data values. The GA converges to the global minimum within the allocated parameter space by evolving an initial population of randomly generated solutions over a large number of generations. The predictive power of both models was then assessed by evaluating their ability to match the remaining steady state data. The subset of data used to optimise the models accounts for one third of the full data set, allowing the remaining two thirds to provide data for validation. The data employed as the optimisation subset were selected to capture the overall dynamic behaviour of the system effectively but were sufficiently sparse to allow us to check the predictive capability of the models. Further optimisation of the model parameters was performed against a distinct set of time course data. Recombination efficiency was recorded at ten time points over a three-hour period for integration and excision reactions of varying initial recombinase concentrations. Since the model of [2] was originally developed for *in vivo* molecular interactions, we made appropriate adaptations for simulating our *in vitro* system dynamics. That is, we removed the parameters governing recombinase expression levels, as the quantities of recombinases in the system is predetermined *in vitro* i.e. they are not expressed within the system itself. Also, the sequestration mechanism, shown to improve performance of our models, was added in order to allow direct comparisons between the models. As shown in Fig. 3A, the simple model is unable to provide a reasonable replication of the experimental steady state data subset and has poor predictive capability. The more detailed model, however, exhibits strong predictive power, accurately reproducing the experimental steady state recombination efficiencies across all initial concentrations of integrase and gp3 (Fig. 3B). Similar results were seen for the time course data (Fig. 4). The parameter space within which

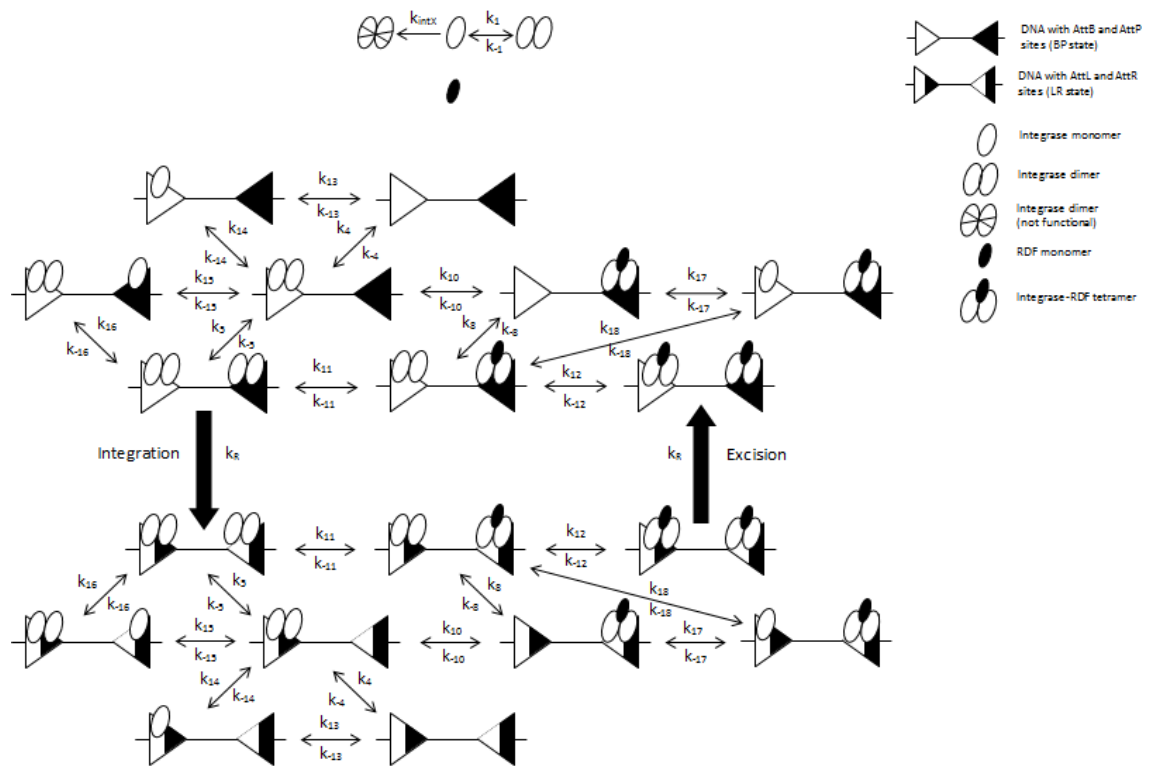


Fig. 2. The DNA recombination reaction network used to construct the model. The network is based on the mechanisms underlying DNA recombination that have been verified in the current experimental literature. We model the dynamics of ϕ C31 integrase and its RDF, gp3. Reactions and their rate constants are depicted by arrows and their corresponding numbered k . The rate of recombination is k_R .

both models are optimised is large, since evidence regarding the numerical values of many reaction rate constants is currently lacking in the literature. Future work will focus on further investigation of model parameters to ensure their biological plausibility and interpret their implications.

IV. CONCLUSIONS

We have developed the first detailed mechanistic model of *in vitro* DNA recombination. The predictive power of the model was validated against a large set of experimental data of recombination efficiencies for different initial concentrations of integrase and gp3. The proposed model sheds light on a number of mechanistic features of DNA recombination for which there is currently no consensus in the experimental literature, and will be a valuable design tool for Synthetic Biologists working on the construction of recombinase-based genetic circuitry, potentially producing significant reductions in development times. Future work will extend our modelling investigation to the *in vivo* system in order to examine model performance under cellular biological conditions.

V. MATERIALS AND METHODS

Proteins (ϕ C31 integrase and gp3) were purified as described in [16], [17], [18]. Integrase and gp3 were diluted in integrase dilution buffer [25 mM Tris-HCl (pH 7.5), 1 mM DTT, 1 M NaCl, and 50% (vol/vol) glycerol] and kept at -20°C . Substrate plasmids containing inverted repeat recombination sites (pZJ56off with attB and attP; pZJ56on

with attR and attL) used for *in vitro* assay were prepared from *E. coli* DH5, using a plasmid miniprep kit (Qiagen). DNA concentrations were determined by measuring the absorbance at 260 nm.

In a typical reaction, premixed integrase and gp3 with $10\times$ their final concentrations were added to a solution of substrate plasmid (~ 10 nM) in a reaction buffer [50 mM Tris-HCl (pH 7.5), 0.1mM EDTA, 5mM spermidine, and 0.1 mg/ml BSA]. Reactions were carried out at 30°C , terminated at various time points, by heating the samples to 80°C for 10 min. Samples were digested with restriction enzymes, then, treated with 5 μl of loading buffer [25mM Tris-HCl (pH 8.2), 20% (wt/vol) Ficoll, 0.5% sodium dodecyl sulphate, 5 mg/ml protease K, 0.25 mg/ml bromophenol blue] at 37°C for 30 min prior to loading onto 1.2% (wt/vol) agarose gels. Gels were stained with ethidium bromide, destained in electroporation buffer, and photographed using Bio-Rad UV Transilluminator. Recombinant and non-recombinant DNA bands were quantitated using the volume analysis tool of Quantity One software.

REFERENCES

- [1] T. Gardner, C. Cantor and J. Collins. Nature. 2000 Jan 20;403(6767):339-42.
- [2] J. Bonnet, P. Subsoontorn and D. Endy. Proc Natl Acad Sci USA. 2012 Jun 5;109(23):8884-9. doi: 10.1073/pnas.1202344109.
- [3] A. Fireland, T. Lu, X. Wang, D. Shi, G. Church and J. Collins. Science. 2009 May 29;324(5931):1199-202. doi: 10.1126/science.1172005.

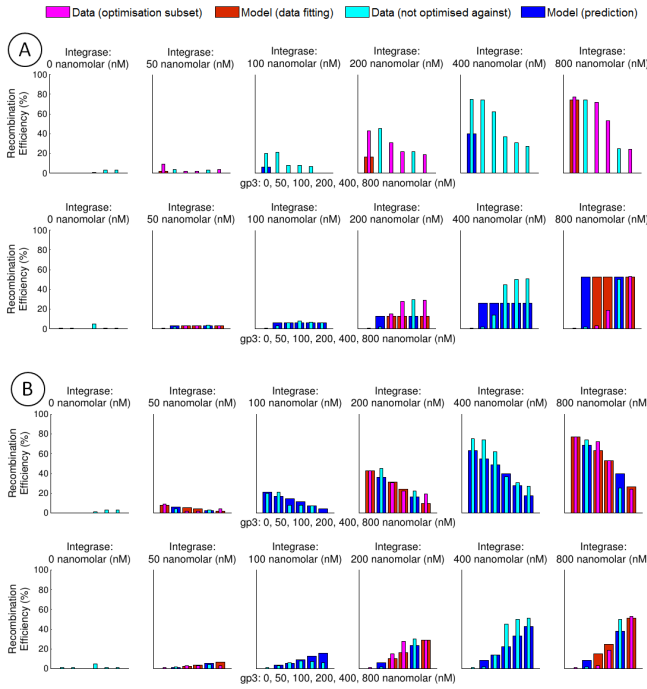


Fig. 3. A: Data fitting / prediction results for the model of [2]. B: Data fitting / prediction for our model. In both A and B the top row of bar graphs represents the integration (BP-LR) reaction and the bottom row of bar graphs represents the excision (LR-BP) reaction. The wider bars represent model simulations and the thinner bars represent data.

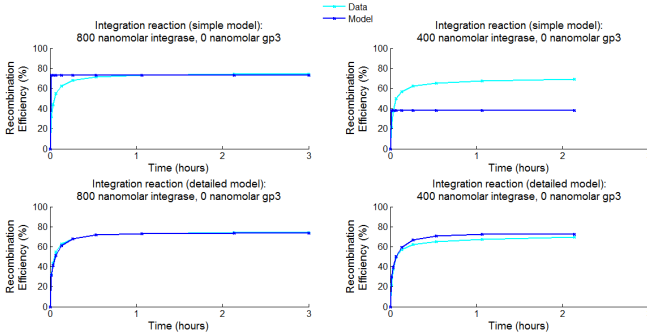


Fig. 4. Typical data fitting results for the model of [2] (top row) and our model (bottom row) against time course data. Model simulations are plotted against two integration reaction time course data sets, one initiated with 800 nanomolar integrase and the other with 400 nanomolar integrase. The cyan line represents data and the blue line represents the model simulations.

[4] D. Burrill and P. Silver. Cell. 2010 Jan 8;140(1):13-8. doi: 10.1016/j.cell.2009.12.034.
 [5] P. Ghosh, L. Wasil and G. Hatfull. PLoS Biol. 2006 Jun;4(6):e186.
 [6] P. Ghosh, N. Pannunzio and G. Hatfull. J Mol Biol. 2005 Jun 3;349(2):331-48. Epub 2005 Apr 7.
 [7] P. Ghosh, L. Bibb and G. Hatfull. Proc Natl Acad Sci U S A. 2008 Mar 4;105(9):3238-43. doi: 10.1073/pnas.0711649105.
 [8] N. Grindley, K. Whiteson and P. Rice. Annu Rev Biochem. 2006;75:567-605.
 [9] P. Ghosh, N. Pannunzio and G. Hatfull. J Mol Biol. 2005 Jun 3;349(2):331-48. Epub 2005 Apr 7.
 [10] T. Miura, Y. Hosaka, Y. Yan-Zhuo, T. Nishizawa, M. Asayama, H. Takahashi and M. Shirai. J Gen Appl Microbiol. 2011;57(1):45-57.
 [11] P. Fogg, S. Colloms, S. Rosser, M. Stark and M. Smith. J Mol Biol. 2014 Jul 29;426(15):2703-16. doi: 10.1016/j.jmb.2014.05.014. Epub 2014 May 22.

$$\begin{aligned}
 \frac{d[I]}{dt} &= 2k_{-1}[I_2] - 2k_1[I][I] - 2k_{\max}[I][I] + \\
 & k_{-13}[D_{BP}I] - k_{13}[D_{BP}I][I] + k_{-13}[D_{LR}I] - k_{13}[D_{LR}I][I] + \\
 & k_{-14}[D_{BP}I_2] - k_{14}[D_{BP}I][I] + k_{-14}[D_{LR}I_2] - k_{14}[D_{LR}I][I] + \\
 & k_{-15}[D_{BP}I_3] - k_{15}[D_{BP}I_2][I] + k_{-15}[D_{LR}I_3] - k_{15}[D_{LR}I_2][I] + \\
 & k_{-16}[D_{BP}I_4] - k_{16}[D_{BP}I_3][I] + k_{-16}[D_{LR}I_4] - k_{16}[D_{LR}I_3][I] + \\
 & k_{-17}[D_{BP}I_5R] - k_{17}[D_{BP}I_4R][I] + k_{-17}[D_{LR}I_5R] - k_{17}[D_{LR}I_4R][I] + \\
 & k_{-18}[D_{BP}I_4R] - k_{18}[D_{BP}I_3R][I] + k_{-18}[D_{LR}I_4R] - k_{18}[D_{LR}I_3R][I], \\
 \frac{d[I_2]}{dt} &= k_1[I][I] - k_{-1}[I_2] + \\
 & k_{-4}[D_{BP}I_2] - k_4[D_{BP}I][I_2] + k_{-5}[D_{BP}I_4] - k_5[D_{BP}I_2][I_2] + \\
 & k_{-8}[D_{BP}I_4R] - k_8[D_{BP}I_2R][I_2] + k_{-4}[D_{LR}I_2] - k_4[D_{LR}I][I_2] + \\
 & k_{-5}[D_{LR}I_4] - k_5[D_{LR}I_2][I_2] - k_8[D_{LR}I_2R][I_2] + k_{-8}[D_{LR}I_4R], \\
 \frac{d[R]}{dt} &= k_{-11}[D_{BP}I_4R] - k_{11}[D_{BP}I_4][R] + k_{-12}[D_{BP}I_4R_2] - k_{12}[D_{BP}I_4R][R] + \\
 & k_{-11}[D_{LR}I_4R] - k_{11}[D_{LR}I_4][R] + k_{-12}[D_{LR}I_4R_2] - k_{12}[D_{LR}I_4R][R] + \\
 & k_{-10}[D_{BP}I_2R] - k_{10}[D_{BP}I_2][R] + k_{-10}[D_{LR}I_2R] - k_{10}[D_{LR}I_2][R], \\
 \frac{d[D_{BP}I]}{dt} &= k_{-4}[D_{BP}I_2] - k_4[D_{BP}I][I_2] + k_{-13}[D_{BP}I] - k_{13}[D_{BP}I][I], \\
 \frac{d[D_{BP}I_2]}{dt} &= k_4[D_{BP}I][I_2] - k_{-4}[D_{BP}I_2] + k_{-5}[D_{BP}I_4] - k_5[D_{BP}I_2][I_2] + \\
 & k_{14}[D_{BP}I][I] - k_{-14}[D_{BP}I_2] + k_{-15}[D_{BP}I_3] - k_{15}[D_{BP}I_2][I] + \\
 & k_{-10}[D_{BP}I_2R] - k_{10}[D_{BP}I_2][R], \\
 \frac{d[D_{BP}I_4]}{dt} &= k_5[D_{BP}I_2][I_2] - k_{-5}[D_{BP}I_4] + k_{-11}[D_{BP}I_4R] - k_{11}[D_{BP}I_4][R] + \\
 & k_{16}[D_{BP}I_3][I] - k_{-16}[D_{BP}I_4] - k_{R}[D_{BP}I_4], \\
 \frac{d[D_{BP}I_2R]}{dt} &= k_{-8}[D_{BP}I_4R] - k_8[D_{BP}I_2R][I_2] + k_{10}[D_{BP}I_2R][I] - k_{-10}[D_{BP}I_2R] + \\
 & k_{-17}[D_{BP}I_3R] - k_{17}[D_{BP}I_2R][I], \\
 \frac{d[D_{BP}I_4R]}{dt} &= k_8[D_{BP}I_2R][I_2] - k_{-8}[D_{BP}I_4R] + k_{18}[D_{BP}I_3R][I] - k_{-18}[D_{BP}I_4R] + \\
 & k_{11}[D_{BP}I_4][R] - k_{-11}[D_{BP}I_4R] + k_{-12}[D_{BP}I_4R_2] - k_{12}[D_{BP}I_4R][R], \\
 \frac{d[D_{BP}I_4R_2]}{dt} &= k_{12}[D_{BP}I_4R][R] - k_{-12}[D_{BP}I_4R_2] + k_{R}[D_{BP}I_4R_2], \\
 \frac{d[D_{LR}I]}{dt} &= k_{-4}[D_{LR}I_2] - k_4[D_{LR}I][I_2] + k_{-13}[D_{LR}I] - k_{13}[D_{LR}I][I], \\
 \frac{d[D_{LR}I_2]}{dt} &= k_4[D_{LR}I][I_2] - k_{-4}[D_{LR}I_2] + k_{-5}[D_{LR}I_4] - k_5[D_{LR}I_2][I_2] + \\
 & k_{14}[D_{LR}I][I] - k_{-14}[D_{LR}I_2] + k_{-15}[D_{LR}I_3] - k_{15}[D_{LR}I_2][I] + \\
 & k_{-10}[D_{LR}I_2R] - k_{10}[D_{LR}I_2][R], \\
 \frac{d[D_{LR}I_4]}{dt} &= k_5[D_{LR}I_2][I_2] - k_{-5}[D_{LR}I_4] + k_{-11}[D_{LR}I_4R] - k_{11}[D_{LR}I_4][R] + \\
 & k_{16}[D_{LR}I_3][I] - k_{-16}[D_{LR}I_4] + k_{R}[D_{LR}I_4], \\
 \frac{d[D_{LR}I_2R]}{dt} &= k_{-8}[D_{LR}I_4R] - k_8[D_{LR}I_2R][I_2] + k_{10}[D_{LR}I_2R][I] - k_{-10}[D_{LR}I_2R] + \\
 & k_{-17}[D_{LR}I_3R] - k_{17}[D_{LR}I_2R][I], \\
 \frac{d[D_{LR}I_4R]}{dt} &= k_8[D_{LR}I_2R][I_2] - k_{-8}[D_{LR}I_4R] + k_{18}[D_{LR}I_3R][I] - k_{-18}[D_{LR}I_4R] + \\
 & k_{11}[D_{LR}I_4][R] - k_{-11}[D_{LR}I_4R] + k_{-12}[D_{LR}I_4R_2] - k_{12}[D_{LR}I_4R][R], \\
 \frac{d[D_{LR}I_4R_2]}{dt} &= k_{12}[D_{LR}I_4R][R] - k_{-12}[D_{LR}I_4R_2] - k_{R}[D_{LR}I_4R_2], \\
 \frac{d[I_{2X}]}{dt} &= k_{\max}[I][I], \\
 \frac{d[D_{BP}I]}{dt} &= k_{13}[D_{BP}I][I] - k_{-13}[D_{BP}I] + k_{-14}[D_{BP}I_2] - k_{14}[D_{BP}I][I], \\
 \frac{d[D_{BP}I_3]}{dt} &= k_{15}[D_{BP}I_2][I] - k_{-15}[D_{BP}I_3] + k_{-16}[D_{BP}I_4] - k_{16}[D_{BP}I_3][I], \\
 \frac{d[D_{BP}I_5R]}{dt} &= k_{17}[D_{BP}I_4R][I] - k_{-17}[D_{BP}I_5R] + k_{-18}[D_{BP}I_4R] - k_{18}[D_{BP}I_5R][I], \\
 \frac{d[D_{LR}I]}{dt} &= k_{13}[D_{LR}I][I] - k_{-13}[D_{LR}I] + k_{-14}[D_{LR}I_2] - k_{14}[D_{LR}I][I], \\
 \frac{d[D_{LR}I_3]}{dt} &= k_{15}[D_{LR}I_2][I] - k_{-15}[D_{LR}I_3] + k_{-16}[D_{LR}I_4] - k_{16}[D_{LR}I_3][I], \\
 \frac{d[D_{LR}I_5R]}{dt} &= k_{17}[D_{LR}I_4R][I] - k_{-17}[D_{LR}I_5R] + k_{-18}[D_{LR}I_4R] - k_{18}[D_{LR}I_5R][I].
 \end{aligned}$$

Fig. 5. Model equations arising from the application of mass action kinetics to the network reactions shown in Fig. 2. The model comprises 22 ODEs. Concentrations are represented by square brackets. DNA is denoted by D with the subscripts BP and LR representing the BP and LR states respectively. Integrase and RDF are denoted by I and R respectively, with numerical subscripts representing the number of monomers present / bound. The dysfunctional integrase dimer is denoted by I_{2X} .

[12] P. Ghosh, L. Bibb and G. Hatfull. Proc Natl Acad Sci U S A. 2008 Mar 4;105(9):3238-43. doi: 10.1073/pnas.0711649105.
 [13] A. Groth and M. Calos. J Mol Biol. 2004 Jan 16;335(3):667-78.
 [14] B. Swalla, E. Cho, R. Gumpert and J. Gardner. Mol Microbiol. 2003 Oct;50(1):89-99.
 [15] R. Keenholz, S. Rowland, M. Boocock, M. Stark and P. Rice. Structure. 2011 Jun 8;19(6):799-809. doi: 10.1016/j.str.2011.03.017.
 [16] M. Smith, R. Till, K. Brady, P. Soutanas, H. Thorpe and M. Smith. Nucleic Acids Res. 2004 May 11;32(8):2607-17. Print 2004.
 [17] F. Olorunniji, D. Buck, S. Colloms, A. McEwan, M. Smith, M. Stark and S. Rosser. Proc. Natl. Acad. Sci. U. S. A. 2012, 109, 1966119666.
 [18] T. Khaleel, E. Younger, A. McEwan, A. Varghese and M. Smith. Mol. Microbiol. 2011, 80, 14501463.