Biophys J BioFAST, published on July 25, 2008 as doi:10.1529/biophysj.108.135061

Monte Carlo simulations of protein assembly, disassembly, and linear motion on DNA

Thijn van der Heijden[1]
Kavli Institute of Nanoscience,
Delft University of Technology, Delft, The Netherlands

Cees Dekker[2]
Kavli Institute of Nanoscience,
Delft University of Technology, Delft, The Netherlands

July 19, 2008

[1]Present address: Leiden Institute of Physics, 2333 CA Leiden, The Netherlands

[2]Corresponding author. Address: Kavli Institute of Nanoscience, Delft University of Technology, Lorentzweg 1, Delft, The Netherlands, Tel.: +31 (0)15 278-6094, Fax: +31 (0)15 278-1202

Abstract

We use Monte Carlo simulations to analyze the simultaneous interactions of multiple proteins to a long DNA molecule. We study the time dependence of protein organization on DNA for different regimes that comprise (non)cooperative sequence-independent protein assembly, dissociation, and linear motion. A range of different behaviors is observed for the dynamics, final coverage, and cluster size distributions. We observe that the DNA substrate is almost never completely covered by protein when taking into account only (non-)cooperative binding, because gaps remain on the substrate that are smaller than the binding site size of the protein. Due to these gaps, the apparent binding size of a protein during non-cooperative binding can be overestimated by up to 30%. During dissociation of cooperatively bound proteins the dissociation curve can be exponentially shaped even when allowing only end-dependent dissociation. We discuss the potential of our method for the analysis of a number of single-molecule experiments, for example, the binding of the DNA-repair proteins RecA and Rad51 to DNA.

Key words: RSA model; RecA; Rad51; binding; protein-DNA interaction

Introduction

In the last decade, new experimental techniques have opened the way to study protein-DNA or protein-protein interactions at the level of single molecules. In contrast to bulk experiments, single-molecule experiments do not suffer from averaging multiple events, thereby allowing much more detailed characterization. The interaction between proteins and DNA involves a variety of relevant processes, e.g. binding, dissociation, translocation, shape deformation, etc. To describe the dynamic interactions between a protein and DNA, one can study systems where the protein-DNA interaction is restricted to a specific site, e.g. a recognition sequence for a restriction enzyme (1). For this type of experiments with a single protein interacting with DNA, models were developed to extract the relevant kinetic interaction rates.

For proteins that bind nonspecifically to DNA, however, the situation may be much more complex. Often, many proteins interact with DNA and other proteins simultaneously. One approach to study these systems in detail has been to avoid multiple events by severely reducing the amount of target area. For example, the length of available DNA substrate can be limited to only tens of bases, or the concentration of protein present in the reaction can be substantially lowered with respect to the target area. Another approach to study the dynamics of multiple non-specific interactions of proteins with DNA or other proteins is to develop models that go beyond the description of single-entity binding.

In the classic protein-binding model of McGhee and von Hippel (MVH) (2), two cases of non-specific protein-DNA binding are addressed: non-cooperative and cooperative binding. In the former, proteins bind randomly to the lattice without any preference to bind adjacent to an already bound protein. In cooperative binding, however, a nucleation event is followed by an extension phase where proteins preferably bind next to an occupied lattice position (Fig. 1a). In their analytical approach, MVH assumed an infinite lattice to which proteins can bind (non-)cooperatively, without taking into account disassembly. Depending on the protein concentration and the strength of the cooperativity the fractional final coverage was deduced at equilibrium, yielding a value for the binding constant of the protein to the lattice.

Although this model is valuable and widely used to determine the magnitude of cooperativity for a certain protein-DNA system (3–5), it has certain limitations. In a single-molecule experiment one can measure the fractional

coverage as it develops in time. The approach of MVH does not allow to describe this dynamics or to extract kinetic parameters from the experimental single-molecule data, because it restricts the description to the final equilibrated system. Furthermore, the model proposed by MVH assumes that for B proteins that bind non-cooperatively to the lattice, there are $B+1$ 'gaps' of bare DNA in between the bound proteins. This is a priori not true, because even in the non-cooperative case, proteins can bind next to an already bound protein. Finally, the obtained coverage for cooperative binding in the MVH model is always complete. This outcome is incorrect because gaps smaller than the binding size of the protein remain on the lattice due to the random nucleation of proteins along the lattice.

Recently, an analytical tool based upon hidden Markov modeling was developed and applied to extract kinetic rates from single-molecule fluorescence data (6, 7) and ion-channel data (8, 9). A regular Markov model consists of a series of states where at each time the system may change from the state it was in the moment before, or may stay in the same state. These states are directly visible to the observer. In a hidden Markov model, however, the model contains an underlying stochastic process that is not observable (it is hidden) (10). A correct interpretation of single-molecule data using hidden Markov modeling, depends on the number of states in the model and the corresponding probabilities involved (11). Furthermore, the states within the model should be independent of each other. For example, hidden Markov modeling does not work well for RNA secondary-structure analysis (10). Studying protein binding to a lattice using hidden Markov modeling causes a similar problem because an already bound protein can influence a different protein cluster.

Here, we develop a new analysis method based upon Monte Carlo simulations that allows a description of both the dynamics and the final states of systems with multiple protein-DNA or protein-protein interactions via non-specific target areas. In these Monte Carlo simulations, a Markov chain of different states is calculated. The simulations allow following the interaction of multiple proteins with a single DNA or protein substrate modeled as a one-dimensional lattice in time. We separately show the results for proteins that bind non-cooperatively or cooperatively, dissociate, or reorganize along the DNA substrate. Furthermore, combinations of these three different interaction modes are implemented in the Monte Carlo simulations. The usefulness of our method was recently illustrated with a comparison to the interaction between the recombinase Rad51 with single- and double-stranded

DNA, where a fit of the model to the experimental data allows to extract a variety of protein-DNA interaction parameters which could not be obtained otherwise (12). Finally, we suggest a number of different systems to which this method can be applied.

Description of the model

We model protein-DNA interaction using Monte Carlo simulations (13–16). In our Monte Carlo simulations, we model the interaction between protein and DNA with a Markov chain where the next state of the protein-DNA complex depends on the current state. The transition probabilities between different states are given, and a certain stochastic pathway results. The Monte Carlo approach allows to study both the dynamics as well as equilibrium states.

We first describe the concept for simulating the simplest two-state process that can be written as

$$A \xrightarrow{k_1} B. \tag{1}$$

The reaction rate k_1 is coupled to a transition probability p_1 in a Markov Chain as

$$k_1 = \frac{p_1}{\Delta t}, \tag{2}$$

where Δt is the duration of a single simulation step in the Monte Carlo simulations. The duration of a simulation step is taken such that (i) the transition probability within a single simulation step is always much smaller than unity, and (ii) the chance of having two local transitions within a single simulation step is negligible. In the Monte Carlo simulations, a transition to the new state occurs when the transition probability is larger than a random value extracted from a uniform distribution between 0 and 1.

The interaction between proteins and DNA is implemented as follows in the Monte Carlo simulations (for details see Methods). The DNA substrate is represented as a long one-dimensional array with the number of elements equivalent to the number of nucleotides or base pairs available. Upon binding, a protein occupies a certain number of elements corresponding to its binding site size corresponding to the most simple model of irreversible adsorption known as random sequential adsorption (RSA). The RSA model has proven to be quite successful in describing a number of systems (17), lattice and continuum limits have been studied (18, 19). Subsequently, the protein

can dissociate or move along the substrate, respectively freeing or occupying other elements of the array. The protocol for binding, dissociating, and moving is repeated for each protein interacting with the DNA. Monitoring the transitions in time can be done by evaluating observables such as the lattice occupancy, contour length, or stepping of a single labeled protein.

Methods

Binding of protein onto a DNA substrate was modeled using Monte Carlo simulations implemented in Interactive Data Language (RSI, Boulder CO). A one-dimensional array was used to represent the DNA substrate containing a number of elements equivalent to the number of nucleotides or base pairs of the DNA molecule of interest. Simulations were done with various binding sizes for the protein. Cooperative binding was described by nucleation followed by growth that extended the nucleation point, whereas non-cooperative binding involved nucleation only.

Nucleation was allowed to occur at any point along the entire molecule. In the Monte Carlo simulations, the nucleation step was simulated as follows: a value was randomly extracted from a uniform distribution yielding a value between 0 and 1. If this value was smaller than a given threshold corresponding to the set nucleation rate for the entire molecule, a protein was bound. The binding location was deduced from a second random number between 0 and 1, that was extracted from a uniform distribution that was multiplied by the number of elements in the one-dimensional array. Binding occurred only when this site plus the following $n-1$ sites were not covered by another protein, to account for the fact that each protein covers n nucleotides or base pairs.

For cooperative binding, we evaluated all nucleation sites where protein patch extension could occur. For each site, a value was extracted from an uniform distribution and compared to a given threshold corresponding to the set rate of extension for a single protein patch. If this value was smaller than the threshold, the protein patch was extended if the next n nucleotides or base pairs were not already covered by protein. Extension was only permitted into the direction of higher numbers in the one-dimensional array.

The probabilities for nucleation and growth per time step were taken so small, that the chance of two binding events within a single Monte Carlo step was negligible. For comparison to experiments, the threshold values, which

are rates expressed in units (Monte Carlo step)$^{-1}$, can convert into kinetic rates expressed in s^{-1} by adjusting the time axis of the Monte Carlo growth curve to the experimental growth data. Whereas our simple modeling involved protein patch extension and disassembly in an unidirectional fashion, the model can be extended using protein patch extension and/or disassembly in both directions. Essentially the same results are found if extension and disassembly occur in both directions, albeit with two slightly different values for the rates that change by a factor up to 2.

In those cases where disassembly was considered, we additionally allowed dissociation to occur after the protein patch extension step. At each end of a protein patch (i.e. a protein cluster consisting of m protein monomers, with $m \geq 1$) opposite to the protein-patch-extension end (i.e. towards lower numbers in the array), a value was extracted from a uniform distribution and if this value was smaller than the threshold set by the dissociation rate, the protein dissociated and a vacancy was created. In the case of diffusion of these end-bound monomers, the protein remained bound to the lattice. Alternatively, a second route was considered where dissociation was allowed at all monomer sites i.e., also in the middle of protein patches. Here, the above procedure was extended to all bound proteins.

Reorganization of individual proteins or protein patches along the DNA substrate was incorporated as follows: a value was randomly extracted from a uniform distribution yielding a value between 0 and 1. If this value was smaller than a given threshold corresponding to the reorganization rate, a step of the protein patch was made of one nucleotide or base pair. For unidirectional translocation, the direction was chosen uniformly, towards lower numbers in the array. For diffusive motion, the stepping direction was randomly towards higher/lower numbers in the array, when an extracted value from a uniform distribution was larger/smaller than 0.5. Diffusive motion of end-bound monomers after detachment was done similarly. Upon collision with individual proteins or protein patches, the diffusive motion was stopped.

To ensure the robustness of the code, all simulations were run a number of times (with different seeds) to validate that the outcome was similar for different runs. Typical data of the different scenarios is shown in the corresponding figures.

Results

We modeled the interaction between proteins and DNA for a variety of processes, i.e. binding, dissociation, reorganization, and combinations of these. Protein-DNA binding can be divided into two different schemes, non-cooperative and cooperative binding (see Fig. 1a) (20–23). We first present the results for non-cooperative binding of proteins to DNA.

Non-cooperative binding

Non-cooperative binding of proteins to a DNA molecule is modeled in the Monte Carlo simulations as random binding to a one-dimensional lattice. Upon binding, the protein covers a binding site of multiple nucleotides or base pairs. First, only binding is considered; i.e., once bound, a protein does not disassemble or rearrange. The occupancy of the lattice is followed in time (see Fig. 2a). The resulting protein coverage displays an exponential growth profile, independent of the binding site size of the protein (see Fig. 2b). The final occupancy, however, varies with respect to the chosen size of the binding site of the protein (see Fig. 2d). For a binding site size of 1 nt, full occupancy is obtained, as expected. However, the fractional occupancy decreases for an increasing binding site size, reaching a plateau of approximately 0.76 (Fig. 2d). Due to the finite size of the binding site, gaps of unoccupied lattice elements with a size smaller than the binding site remain throughout the lattice (see bottom panel in Fig. 2a for an example). The actual number of bound proteins to the lattice is therefore smaller than when all proteins would mutually align such that no gaps would remain on the lattice. Division of the length of the lattice by the number of bound proteins yields the apparent binding size for the protein, which is larger than its intrinsic binding size due to the existence of gaps. This leads to an increase in the apparent binding size of $29.5 \pm 0.2\%$ compared to the actual binding size (see Fig. 2e).

The kinetics of simple non-cooperative binding can analytically be described as follows (21). The binding process is limited by the amount of free base pairs available on the DNA molecule (see Fig. 2a). During growth the amount of free base pairs N_{free} decreases according to $\frac{dN_{\text{free}}}{dt} = -aN_{\text{free}}$, where a is the binding rate of the protein to the lattice, which together with the boundary condition of $N_{\text{free}}(0) = N$ yields $N_{\text{free}} = N\text{e}^{-at}$. The time-

dependent occupancy θ becomes

$$\theta(t) \equiv \frac{N_{\text{bound}}(t)}{N} = 1 - \mathrm{e}^{-at}, \tag{3}$$

showing an exponential binding profile in excellent agreement with the profiles obtained in the Monte Carlo simulations (see black lines in Fig. 2b).

The final occupancy depends on the binding size n of the protein. A lattice that consists of N possible binding sites allows binding up to $\frac{N}{n}$ proteins. During a non-cooperative binding process, 'gaps' of size i ($1 \leq i \leq n-1$) are created throughout the entire lattice reducing the final amount of proteins on the DNA. For the final state, an effective binding size $n^* = n + s_{\text{gap}}$ can be defined, where s_{gap} is the average gap size between proteins. This average gap size between proteins is not equivalent to $\frac{1}{2}(n-1)$ but instead $s_{\text{gap}} = \sum_{i=1}^{n-1} \frac{i}{n+i}$. When the binding size increases, the possible gap sizes increase accordingly. Therefore, one needs to take into account the actual number of proteins with an adjacent gap size i that is bound to the lattice decreasing as $(n+i)^{-1}$ and not as n^{-1}. Together, this yields for the fractional occupancy

$$\theta = \frac{N/n^*}{N/n} = \frac{n}{n + \sum_{i=0}^{n-1} \frac{i}{n+i}}. \tag{4}$$

which can be simplified into

$$\theta = \frac{1}{2 - \Psi(2n) + \Psi(n)}, \tag{5}$$

where $\Psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$ and $\Gamma(x)$ is the Gamma function. This relation between fractional coverage and binding site indeed describes the observed behavior from our Monte Carlo simulations well, cf. the solid red line in Fig. 2d. The fractional coverage for a protein covering two sites ($n = 2$) is 0.857, close to the result of 0.865 derived by (24) using combinatorial techniques. For large binding sites, non-cooperative binding is similar to the 'car parking problem', where one-dimensional cars are parked randomly in a linear array

(25, 26). Eq. 5 yields a fractional coverage of 0.765 for $n \to \infty$ in fairly good agreement with the result of 0.748 obtained for the car parking problem (25).

Cooperative binding

Cooperative binding of proteins to a DNA molecule is modeled in the Monte Carlo simulations in two steps: nucleation followed by extension. 'Nucleation' denotes protein binding at an unoccupied DNA position not adjacent to already bound proteins, whereas we define 'extension' as binding to a site directly adjacent to one that is already occupied. We can follow the binding process to the lattice in the Monte Carlo simulations in time by visualizing the binding of every individual protein (or protein cluster if binding occurs via multimers), see Fig. 3a and b. Using these simulations, we obtained lattice occupancy profiles at different ratios between extension and nucleation (Fig. 3c and d). Different qualitative behavior is observed, depending on the ratio between rates for extension and nucleation, henceforth called the cooperativity number ω. For a high cooperativity number (i.e. when nucleation is rare, see black line in Fig. 3a and b), nothing happens until a first nucleation event occurs, after which the coverage of the DNA molecule increases linearly due to the extension of the protein patch. By contrast, at low cooperativity numbers, many nucleation loci are created followed by extension into multiple protein patches. The process ends when the molecule has no more free binding sites that are large enough to accommodate binding of another protein or protein cluster. With increasing cooperativity numbers, the obtained time-dependent binding profiles change from an exponential (for a ratio of zero, equivalent to non-cooperative binding) to a linear relationship (for ratios larger than 10^6) (see Fig. 3c and d). The final fractional coverage of the lattice increases for increasing cooperativity values, because the final amount of gaps is reduced (Fig. 3g) (20, 27). Although extension can be orders of magnitude larger than nucleation, full coverage is hardly ever obtained for $n \geq 2$.

The final distribution of protein clusters along the lattice can be quantified. Different protein-patch length distributions can be obtained depending on the size of the protein cluster that binds during nucleation and extension and the ratio between nucleation and extension rate (see Fig. 3e and f). For the non-cooperative case, the distribution of protein patch sizes peaks around the size of the binding unit with a long tail towards longer protein patches (see Fig. 2c). An analytical expression for the protein-patch-length

distribution F_c has been proposed by (27)

$$F_c = \left[1 - \frac{1-(n-2\omega+1)\frac{\theta}{n}-R}{\frac{2\theta}{n}(\omega-1)}\right] \times \left[\frac{1-(n-2\omega+1)\frac{\theta}{n}-R}{\frac{2\theta}{n}(\omega-1)}\right]^{c-1}, \quad (6)$$

where

$$R = \sqrt{\left[1-(n+1)\frac{\theta}{n}\right]^2 + \frac{4\omega\theta}{n}(1-\theta)}, \quad (7)$$

and c the length of the protein patch. In the saturated state ($\theta = 1$), the cluster size approaches infinity, independent of the binding site of the protein (27). This prediction for the non-cooperative case of $\omega = 1$ is not in agreement with the observed behavior where the distribution shows a Poissonian profile (see Fig. 2c) due to the existence of gaps. We can fit the distributions for $n \geq 2$ with equation 6, where one can reduce the number of free parameters to one, i.e. the cooperativity number ω, because the relative coverage θ is given by $\frac{1}{N}\sum_{c=1}^{N} b_c c$ with b_c the number of appearances of a protein patch with length c. Protein-patch length distributions for $n = 3$ were fit for varying cooperativity numbers (ω_{in}) with Eq. 6 to obtain an apparent value for the cooperativity number (ω_{out}). Interestingly, the fits yield significant differences – even by orders of magnitude – between the values entered (ω_{in}) and obtained after fitting (ω_{out}) for the cooperativity number, see Fig. 3h. Due to the finite lattice length and incomplete coverage, the fit severely underestimated the cooperativity number in all cases.

Multimeric binding and Hill coefficient

In the MVH model, it is assumed that the binding unit of the protein during nucleation and extension is the same. Both processes, however, can in principle involve different protein multimers. The binding unit can be determined from concentration-dependent binding reactions where the binding

rate in either nucleation or extension, is determined with respect to the protein concentration (see Fig. 4a). This behavior can be described by the Hill equation

$$k_i = \frac{k_{i,\max}[A]^{n_H}}{S_{0.5}^{n_H} + [A]^{n_H}} \tag{8}$$

where n_H is the Hill coefficient and $S_{0.5}$ the concentration where half-maximum activity occurs. The Hill coefficient can be interpreted as the minimal size of the binding unit, i.e. for $n_H = 1$ the protein binds as a monomer to the lattice, whereas for larger values of n_H, the protein binds as a n_H-mer (28). This coordination between proteins, for example by binding of preformed multimers, is sometimes called cooperative binding, but this is entirely unrelated to the cooperative binding defined above (the ratio between extension and nucleation in protein patch formation). Within the Monte Carlo simulations, the binding unit in nucleation and extension can be varied independently. In the case where the binding units for nucleation and extension are equivalent, i.e., when the Hill coefficients are identical, the lattice occupancy profiles remain the same independent of protein concentration. On the other hand, if the binding units are not equivalent, the growth profiles and final occupancy change depending on the protein concentration (see Fig. 4c).

Dissociation

In the above binding schemes, the binding was taken to be irreversible. However, proteins bound to a lattice can have a probability to detach from the lattice, i.e., they dissociate (22). Two different scenarios can be envisioned (see Fig. 1b). As longer protein patches are formed on the lattice, (i) only proteins located at an end of a patch are allowed to dissociate (29, 30), or (ii) all proteins are allowed to dissociate regardless of their position within the patch (31). For the latter, the Monte Carlo simulations show an exponentially shaped disassembly curve (see red line in Fig. 5), as expected. This dissociation behavior is independent of the assembly history. Dissociation is however linear when dissociation occurs only at the end of a single protein patch (see black line in Fig. 5). If growth has resulted in a multitude of small patches, end dissociation, however, also leads to an exponentially shaped profile (see green line in Fig. 5) (30).

These observations can be understood straightforwardly. For non-coo-

perative binding, bound proteins do not gain from protein-protein interactions and equivalently, once bound, every protein has an equal probability to dissociate. Following a similar reasoning as above for binding (Eq. 3), this yields an exponential dissociation profile, in agreement with the Monte Carlo simulations. In the case of cooperatively bound proteins, dissociation results in a linear profile only if a single protein patch exists on the lattice, because the proteins can only dissociate from one end, as indeed observed in the Monte Carlo simulations. Multiple patches result in multiple end-dissociation points, and the broad size distribution of the patches then leads to an approximately exponential dissociation curve (29, 30).

Rearrangements

We also consider in our simulations the spatial rearrangement of proteins on the DNA, where bound proteins can move linearly along the DNA. This is modeled by three different pathways (see Fig. 1c): (i) the protein cluster moves diffusively or (ii) translocates unidirectionally along the DNA molecule, or (iii) end-bound monomers detach and move diffusively towards the neighboring protein cluster. Diffusive movement of the protein leads to a random walk in the Monte Carlo simulations (Fig. 6b). Unidirectional motion leads to an approximately linear relation between traveled distance and time, as expected (see Fig. 6a). Unidirectional motion is of course only possible at the expense of an available energy source, e.g. ATP hydrolysis.

One-dimensional diffusive motion of a protein cluster along the lattice can be written as $\langle x^2 \rangle = 2Dt$ where $\langle x^2 \rangle$ is the average mean-square displacement, D the diffusion coefficient, and t the time that the protein is moving along the lattice. As shown in Fig. 6c, the average mean-square displacement of a single protein over a given time window indeed follows this relation.

So far, we have considered that, upon dissociation, an end-bound monomer detaches from the protein cluster on the DNA and vanishes to bulk solution. Instead of dissociation into the bulk solution, the detached monomer can also remain bound to the DNA molecule as sketched in scenario (iii), (lower mode in Fig. 1c). After end detachment, the monomer diffuses freely between two protein clusters (32). When the monomer reaches either protein cluster, it will bind.

Combination of processes

Above we have shown different interaction modes for a protein with the DNA substrate, i.e. (non-)cooperative binding, dissociation, and reorganization. In the simulations, these modes can be combined in various ways to sort out the different processes that contribute to the occupancy of the lattice. Combination of these pathways can yield very different results. Here, we visualize these using kymographs, graphs that represent the one-dimensional lattice occupancy on one axis and time on the other (33).

Fig. 7a shows a kymograph for cooperative binding, where the final lattice is not completely covered because gaps between protein clusters remain on the lattice. The grayscale in the kymographs indicate protein-bound (white) and protein-free (black) DNA substrate. The permanent gaps in Fig. 7a thus are seen as the black horizontal lines that persist over time. Cooperative binding in the presence of end-dependent disassembly yields a different behavior (Fig. 7b). As can be seen in Fig. 7b, protein patches appear and disappear in time at different positions on the lattice. The dissociation rate is chosen such that the fractional coverage on the lattice remains approximately constant in time.

Upon allowing reorganization of detached end-bound monomers or bound protein patches either by diffusive or unidirectional motion, a completely covered lattice is obtained (see Fig. 7c, e, and g, respectively). Unidirectional movement of the protein patches is observed in the kymographs by linear stripes in the downward direction (Fig. 7g) which represent protein patches that shift and fuse with other patches at the bottom. Diffusive motion of monomers or protein patches leads to a more strongly fluctuating behavior (Fig. 7c and e). Also, upon combining all three modes, cooperative growth, rearrangement and end-dependent disassembly, the final lattice is completely covered (Fig. 7d, f, and h). Note that the time scale now has significantly increased due to the presence of protein dissociation before saturation is obtained. The downward motion observed for 1D-diffusion is caused by protein monomers that erode from patches and after a diffusive walk end up at the next patch. A similar behavior is observed in Fig. 7b.

Discussion

Using Monte Carlo simulations we have modeled different interactions between protein and DNA, i.e. (non-)cooperative binding, dissociation, and reorganization. The flexibility of the Monte Carlo simulations allows using different binding site sizes of the protein during nucleation, filament extension, or dissociation. The Monte Carlo simulations of the different interactions yielded interesting results. First, the DNA substrate is almost never completely covered by protein when taking into account only binding. Gaps remain on the substrate smaller than the binding site size of the protein. Second, the apparent binding size of a protein during non-cooperative binding can be overestimated by up to 30% due to the existence of gaps. Furthermore, the fractional coverage increases for higher numbers of cooperativity. Finally, the dissociation behavior of cooperatively bound protein can lead to an exponentially shaped dissociation curve even when allowing only end-dependent dissociation (30, 31).

We can compare the benefits of our Monte Carlo simulations to the MVH model and hidden Markov modeling. MVH derived equations to describe the binding of a protein to a lattice while taking into account cooperativity. This model has been applied numerous times in equilibrium studies to extract binding constants and the cooperativity number. The model fails however to address the kinetics of individual proteins when dissociation and translocation of bound proteins are relevant. Furthermore, the final occupancy of the lattice is (incorrectly) always complete despite the finite binding site size of the protein involved. Indeed, our Monte Carlo results show that the MVH cooperativity number extracted from filament length distributions did not correspond to the input value. Finally, the binding site sizes of the protein during nucleation and protein patch extension in the original MVH model are identical, whereas in an experiment they can be different.

In contrast to the MVH model, hidden Markov modeling allows addressing the reaction kinetics. Due to the modular setup of a Markov chain, different interaction pathways can be modeled. However, hidden Markov modeling cannot cope with systems where different pathways influence the outcome of each other.

With Monte Carlo simulations, protein-DNA interactions can be followed in time for each protein involved. A disadvantage is that it takes a fair amount of computational power to simulate the kinetics of complex pathways in Monte Carlo simulations, but not unreasonably so. (The current results

were obtained on a pc with an integrated computation time of 3 weeks).

Application of Monte Carlo modeling

The Monte Carlo simulations described here can be applied to a variety of experimental systems. We briefly mention a few examples. A system to which this method was applied, was the interaction between the RecA-like recombinase RAD51 and DNA (12) (see Fig. 8a). Using magnetic tweezers, the end-to-end distance of a tethered DNA molecule was followed in time while RAD51 was allowed to bind forming a nucleoprotein filament. Upon binding to the DNA substrate, RAD51 induces a change in end-to-end distance yielding a measure for the lattice occupancy. The induced changes in end-to-end distance were fit with Monte Carlo simulated binding profiles yielding all relevant single RAD51 (dis)assembly rates.

Another useful application would be RNA-dependent RNA polymerases. These polymerases can either generate template-long duplexes by synthesizing full-length RNA chains in one run, or generate many short duplexes by synthesizing short complementary RNA oligonucleotides scattered along the RNA template, known as abortive initiation (see Fig. 8b) (34). The former is a highly cooperative binding mode, whereas abortive initiation corresponds to a low-cooperative binding mode. This can be experimentally measured because the creation of duplex RNA from a single strand template increases the stiffness of the RNA molecule yielding a change in end-to-end distance of a tethered molecule in e.g. a tweezers setup. These changes in end-to-end distance can be analyzed with the Monte Carlo simulations yielding values for the rates of initiation and duplex extension.

Other protein-DNA binding reactions can be analyzed as well. The case of single-stranded binding proteins like SSB and RPA, or nucleosome binding to DNA is conceptually very similar to the RAD51 binding that we have already described. Another example is structural maintenance of chromosome (SMC) proteins which are the central components of several multiprotein complexes that help to organize chromosomes throughout the cell cycle (35) (Fig. 8c). The analysis presented here provides a basis for quantification of, for example, the binding size, the presence of any cooperativity involved, or possible reorganization during binding of SMC proteins with DNA.

Conclusion

To overcome certain limitations of the classic MVH model and the hidden Markov model, we have used Monte Carlo simulations to model ligand-lattice interaction. These Monte Carlo simulations allow determination of protein-related binding rates even when multiple proteins interact simultaneously with the lattice. This tool was applied to understand RAD51-DNA interaction. Application of this analytical tool can be extended to other systems where cooperativity plays a crucial role, like single-stranded binding proteins, polymerases, and SMC proteins.

Acknowledgments

We have benefited greatly from many discussions with Claire Wyman and Roland Kanaar over the years. We thank Marijn van Loenhout for useful discussions and a critical reading of the manuscript. This work was supported by grants from the "Stichting voor Fundamenteel Onderzoek der Materie (FOM)", which is financially supported by the "Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO)".

References

1. Seidel, R., J. van Noort, C. van der Scheer, J. G. P. Bloom, N. H. Dekker, C. F. Dutta, A. Blundell, T. Robinson, K. Firman, and C. Dekker, 2004. Real-time observation of DNA translocation by the type I restriction modification enzyme EcoR124I. *Nature Structural & Molecular Biology* 11:838–843.

2. McGhee, J. D., and P. H. von Hippel, 1974. Theoretical aspects of DNA-protein interactions - Cooperative and non-cooperative binding of large ligands to a one-dimensional homogeneous lattice. *Journal of Molecular Biology* 86:469–489.

3. Lonberg, N., S. C. Kowalczykowski, L. S. Paul, and P. H. von Hippel, 1981. Interactions of bacteriophage T4-coded gene 32 protein with nucleic-acids: III. Binding-properties of 2 specific proteolytic digestion products of the protein (G32P*I and G32P*III). *Journal of Molecular Biology* 145:123–138.

4. Ando, R. A., and S. W. Morrical, 1998. Single-stranded DNA binding properties of the UvsX recombinase of bacteriophage T4: Binding parameters and effects of nucleotides. *Journal of Molecular Biology* 283:785–796.

5. Moreno-Herrero, F., L. Holtzer, D. A. Koster, S. Shuman, C. Dekker, and N. H. Dekker, 2005. Atomic force microscopy shows that vaccinia topoisomerase IB generates filaments on DNA in a cooperative fashion. *Nucleic Acids Research* 33:5945–5953.

6. Yang, H., G. B. Luo, P. Karnchanaphanurach, T. M. Louie, I. Rech, S. Cova, L. Y. Xun, and X. S. Xie, 2003. Protein conformational dynamics probed by single-molecule electron transfer. *Science* 302:262–266.

7. Joo, C., S. A. McKinney, M. Nakamura, I. Rasnik, S. Myong, and T. Ha, 2006. Real-time observation of RecA filament dynamics with single monomer resolution. *Cell* 126:515–527.

8. Qin, F., A. Auerbach, and F. Sachs, 2000. A direct optimization approach to hidden Markov modeling for single channel kinetics. *Biophysical Journal* 79:1915–1927.

9. Qin, F., A. Auerbach, and F. Sachs, 2000. Hidden Markov modeling for single channel kinetics with filtering and correlated noise. *Biophysical Journal* 79:1928–1944.

10. Eddy, S. R., 2004. What is a hidden Markov model? *Nature Biotechnology* 22:1315–1316.

11. Rabiner, L. R., 1989. A tutorial on hidden Markov-models and selected applications in speech recognition. *Proceedings of the IEEE* 77:257–286.

12. van der Heijden, T., R. Seidel, M. Modesti, R. Kanaar, C. Wyman, and C. Dekker, 2007. Real-time assembly and disassembly of human RAD51 filaments on individual DNA molecules. *Nucleic Acids Research* 35:5646–5657.

13. Metropolis, N., and S. Ulam, 1949. The Monte Carlo method. *Journal of the American Statistical Association* 44:335–341.

14. Halton, J. H., 1970. A retrospective and prospective survey of Monte Carlo method. *Siam Review* 12:1–63.

15. Binder, K., 1986. Monte Carlo methods in statistical physics. Topics in current physics. Springer-Verlag, Berlin ; New York, 2nd edition.

16. Evans, J. W., 1993. Random and cooperative sequential adsorption. *Reviews of Modern Physics* 65:1281–1329.

17. Onoda, G. Y., and E. G. Liniger, 1986. Experimental determination of the random-parking limit in two dimensions. *Physical Review A* 33:715–716.

18. Widom, B., 1966. Random sequential addition of hard spheres to a volume. *Journal of Chemical Physics* 44:3888–3894.

19. Gonzales, J. J., P. C. Hemmer, and J. S. Hoye, 1974. Cooperative effects in random sequential polymer reactions. *Chemical Physics* 3:228–238.

20. Epstein, I. R., 1978. Cooperative and non-cooperative binding of large ligands to a finite one-dimensional lattice - Model for ligand-oligonucleotide interactions. *Biophysical Chemistry* 8:327–339.

21. Epstein, I. R., 1979. Kinetics of large-ligand binding to one-dimensional lattices - Theory of irreversible binding. *Biopolymers* 18:765–788.

22. Epstein, I. R., 1979. Kinetics of nucleic acid large ligand interactions - Exact Monte-Carlo treatment and limiting cases of reversible binding. *Biopolymers* 18:2037–2050.

23. Dateo, C., and I. R. Epstein, 1981. Kinetics of nucleic-acid large ligand interactions - Multiplet-closure approximations and matrix-iteration techniques. *Biopolymers* 20:1651–1669.

24. Flory, P. J., 1939. Intramolecular reaction between neighboring substituents of vinyl polymers. *Journal of the American Chemical Society* 61:1518.

25. Renyi, A., 1958. On a one-dimensional problem concerning random space-filling. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* 3:109–127.

26. Solomon, H., and H. Weiner, 1986. A Review of the Packing Problem. *Communications in statistics - Theory and methods* 15:2571–2607.

27. Kowalczykowski, S. C., L. S. Paul, N. Lonberg, J. W. Newport, J. A. McSwiggen, and P. H. von Hippel, 1986. Cooperative and noncooperative binding of protein ligands to nucleic-acid lattices - Experimental approaches to the determination of thermodynamic parameters. *Biochemistry* 25:1226–1240.

28. Weiss, J. N., 1997. The Hill equation revisited: uses and misuses. *Faseb Journal* 11:835–841.

29. Lohman, T. M., 1984. Kinetics and mechanism of dissociation of cooperatively bound T4-gene-32-protein single-stranded nucleic-acid complexes: I. Irreversible dissociation induced by sodium-chloride concentration jumps. *Biochemistry* 23:4656–4665.

30. Lohman, T. M., 1983. Model for the irreversible dissociation kinetics of cooperatively bound protein nucleic-acid complexes. *Biopolymers* 22:1697–1713.

31. Balazs, A. C., and I. R. Epstein, 1984. Kinetics of irreversible dissociation for proteins bound cooperatively to DNA. *Biopolymers* 23:1249–1259.

32. Lohman, T. M., 1984. Kinetics and mechanism of dissociation of cooperatively bound T4-gene-32-protein single-stranded nucleic-acid complexes: II. Changes in mechanism as a function of sodium-chloride concentration and other solution variables. *Biochemistry* 23:4665–4675.

33. Waterman-Storer, C. M., A. Desai, J. C. Bulinski, and E. D. Salmon, 1998. Fluorescent speckle microscopy, a method to visualize the dynamics of protein assemblies in living cells. *Current Biology* 8:1227–1230.

34. Makeyev, E. V., and D. H. Bamford, 2002. Cellular RNA-dependent RNA polymerase involved posttranscriptional gene silencing has two distinct activity modes. *Molecular Cell* 10:1417–1427.

35. Hirano, T., 2006. At the heart of the chromosome: SMC proteins in action. *Nature Reviews Molecular Cell Biology* 7:311–322.

Figure Legends

Figure 1.

Schematic drawings of different pathways for protein-DNA kinetics. (a) Assembly of a non-specifically binding protein on its DNA substrate can be divided into two modes, non-cooperative and cooperative. In the former (left panel), the protein binds randomly, whereas in the latter (right panel) a preference exist to bind next to an already bound protein. (b) Disassembly of bound proteins can also be divided into two different modes, end-dependent and position-independent dissociation. In the first case (left), only proteins located at the end of a protein complex can dissociate, whereas in the second case (right) all bound proteins, regardless of their position within the protein complex, have the same probability to dissociate. (c) Linear motion of a protein patch can be described by either a diffusive (left) or a unidirectional (right) mode. In the mode depicted at the bottom of the panel, end-bound monomers are allowed to detach from a protein patch and move diffusively towards a neighboring protein patch.

Figure 2.

Non-cooperative binding (a) Snapshots of the DNA occupation by proteins at different times during a Monte Carlo simulation for non-cooperative protein binding. This simulation is carried out for $k = 5{\cdot}10^{-6}$ site^{-1} (MC step)$^{-1}$. As a protein covers 3 nt or 3 bp upon binding, binding can only occur if sufficient space is available. In the bottom panel, the simulation has reached its final state since no further proteins can bind. Gaps of 1 or 2 nt/bp are clear. (b) Time-dependent binding profiles are simulated for different binding sizes, i.e. $n = 1$, 3, and 10 (respectively red, green, and blue lines), showing an exponentially shaped growth curve (red line). Only for $n = 1$, full coverage is obtained. (c) The protein-patch length distribution of bound proteins in the saturated state has a maximum around a dimeric protein-patch length. The solid black line denotes the fit of Eq. 6 yielding a cooperativity number of 1.0 ± 0.3. (d) After protein coverage has saturated, the final occupancy of the substrate was determined. With increasing binding site size of the protein, the final occupancy decreased and finally reaches a plateau of approximately 76 %. The dependence is quite well described by Eq. 5 (red line). The dashed red line indicates the dependence when the gap size corresponds to

$\frac{1}{2}(n-1)$, which clearly fails to describe the data. (e) The apparent binding site size of the protein can deviate from the actual binding site size due to the existence of gaps between bound proteins. In the Monte Carlo simulations the apparent binding site size is equivalent to the real size (red line). The obtained value in the MVH model overestimate the actual value by about 30% (black line).

Figure 3.

Cooperative binding (a) Snapshots of the DNA occupation by protein at different times during a Monte Carlo simulation for cooperative protein binding. This simulation is carried out for $k_{\mathrm{nucl}} = 1 \cdot 10^{-6}$ site^{-1} (MC step)$^{-1}$, $k_{\mathrm{ext}} = 5 \cdot 10^{-5}$ (MC step)$^{-1}$ and a binding size $n = 1$. Due to the fast nucleation rate, multiple protein patches are formed along the DNA substrate. Because the protein covers only a single nucleotide or base pair, the final state (bottom panel) is a fully saturated lattice. (b) For a binding size larger than one, here $n = 5$, a similar intermediate state is observed for equivalent binding rates, but the final state contains gaps since no further proteins can bind. (c) Time-dependent lattice occupancy profiles are obtained from the simulations for different levels of cooperativity. If only random binding (nucleation) occurs along the contour length of the DNA molecule (non-cooperative binding, see top left), an exponential lattice occupancy profile is obtained. However, if protein-patch extension is fast compared to nucleation, e.g., for a ratio larger than 10^6 (strong cooperative binding, see bottom right), the lattice occupancy profile becomes linear and the molecule can be fully covered by the protein. For intermediate ratios between protein-patch extension and nucleation, sigmoidally shaped lattice occupancy profiles are observed. All lattice occupancy profiles reach complete saturation because the binding site size of the protein is one nucleotide or base pair in this case. (d) For $n = 5$ a similar change in binding profiles is observed, but complete saturation is not obtained. (e) The protein-patch-length distribution for a protein with a binding site size of 3 nucleotides or base pairs at a cooperativity number of $\omega_{\mathrm{in}} = 100$. The solid line denotes the best fit obtained with Eq. 6 yielding a cooperativity ω_{out} of 2.3 ± 0.3. (f) Similar scheme, but for a protein with a binding site size of 15 nucleotides or base pairs, yielded a cooperativity number of 10.8 ± 1.0 using Eq. 6. (g) Final occupancy of the substrate for varying numbers of cooperativity. If the binding

site size is 1 nucleotide or base pair full coverage is always obtained. For larger binding site sizes, the final occupancy increases with the applied cooperativity number approaching the full 100 % at very high $k_{\mathrm{ext}}/k_{\mathrm{nucl}}$. (h) Apparent cooperativity number ω_{out} versus actual cooperativity number ω_{in}. For varying cooperativity numbers (ω_{in}), the protein-patch-length distribution is determined for $n = 3$. Subsequently, the simulated distributions are fit with Eq. 6 to obtain a measure for the apparent cooperativity number (ω_{out}). For nucleation-driven reactions ($\omega_{\mathrm{in}} = 1$) the fit yields a value close to one. For extension-driven reactions where the cooperativity number is larger than one, however, the obtained value ω_{out} deviates significantly from the input value ω_{in}.

Figure 4.

Influence of the Hill coefficient on the kinetic interaction between protein and DNA. (a) Concentration dependence of the binding rates. If the protein interacts as a monomer with the DNA substrate, the curve follows a Michaelis-Menten dependence (black). However, for larger complexes (Hill coefficient $n_H \geq 2$), the profiles become sigmoidal (red and green for respectively a dimer and pentamer). (b) The ratio between extension and nucleation is concentration dependent when the Hill coefficients differ for extension and nucleation. Blue and magenta denote the ratio between pentameric-monomeric and monomeric-pentameric binding units, respectively. (c) At three different concentrations (in order of increasing concentrations denoted by 1, 2, and 3 in the inset of a), the lattice occupancy profiles for the three independent cases are depicted. The black curves for $n_H = 2$ (middle panel) in both nucleation and extension are the same for various protein concentrations. The magenta and blue curves, for $n_{\mathrm{ext}} : n_{\mathrm{nucl}} = 1 : 5$ and $5 : 1$, respectively, are protein-concentration dependent. It is clear that the lattice occupancy profiles change when the ratio is not constant in the applied concentration regime.

Figure 5.

Protein dissociation After proteins have formed a single continuous filament on the DNA substrate, a linear decrease is observed when the protein disassembles from one end with $k_{\mathrm{dis}} = 0.2$ (MC step)$^{-1}$ (black line). When all bound proteins have the same probability to dissociate irrespective of their

position in the protein complex, an exponentially shaped disassembly curved is obtained with $k_{\text{dis}} = 6.7 \cdot 10^{-4}$ (MC step)$^{-1}$ (red line). If the proteins form multiple short protein patches on the DNA substrate with bare DNA in between, end-dependent disassembly shows again an exponentially shaped disassembly profile with $k_{\text{dis}} = 2.2 \cdot 10^{-3}$ (MC step)$^{-1}$ (green line).

Figure 6.

Linear motion of a protein (a) The position of a protein bound to the DNA substrate is followed while allowing unidirectional motion with $k_{\text{uni}} = 0.01$ (MC step)$^{-1}$. This yields an approximately linear decrease in time. (b) For a diffusive process with $k_{\text{dif}} = 0.01$ (MC step)$^{-1}$, the position of the protein along the DNA substrate displays a random walk. (c) As expected for a diffusive process, the mean-square displacement of a protein increases approximately linearly in time. The obtained diffusion constant is 0.0049 nt^2 (MC step)$^{-1}$ in excellent agreement with the expected rate of diffusion, $D = \frac{1}{2}k_{\text{dif}} = 0.005$ nt^2 (MC step)$^{-1}$.

Figure 7.

Kymographs for various combinations of protein-DNA interactions. (a) Cooperative protein binding is visualized in time, where white corresponds to proteins occupying lattice sites and black denotes unoccupied lattice positions. The simulation is carried out for $n = 3$, $k_{\text{nucl}} = 3 \cdot 10^{-5}$ site^{-1} (MC step)$^{-1}$, and $k_{\text{ext}} = 5 \cdot 10^{-4}$ (MC step)$^{-1}$. In the final saturated state, gaps remain smaller than the binding size of the protein. (b) In the presence of end-dependent disassembly, $k_{\text{dis}} = 7 \cdot 10^{-4}$ (cluster end)$^{-1}$ (MC step)$^{-1}$, protein patches appear and disappear on the lattice. (c) Cooperative binding and diffusive motion of detached end-bound monomers, $k_{\text{det}} = 0.01$ (MC step)$^{-1}$, and $k_{\text{step}} = 0.1$ (MC step)$^{-1}$, yields a completely covered lattice. (d) In the presence of dissociation of detached monomers, $k_{\text{dis}} = 7 \cdot 10^{-4}$ (detached monomer)$^{-1}$ (MC step)$^{-1}$, a combination of cooperative binding, dissociation, and diffusive motion of detached end-bound monomers also yields a completely covered lattice albeit on a longer time scale. (e) Cooperative binding and diffusive motion of protein patches, $k_{\text{dif}} = 0.01$ (MC step)$^{-1}$, yields a single continuous protein complex. (f) A combination of cooperative binding, dissociation, and diffusive motion leads tot the formation of single continuous protein complex albeit on a longer time scale. (g) A similar

saturated end state is observed for cooperative binding and unidirectional motion of protein patches, $k_{\mathrm{uni}} = 0.01$ (MC step)$^{-1}$. (h) Same as (f) but with unidirectional instead of diffusive motion. This also eventually leads to the formation of a single continuous protein complex.

Figure 8.

Different biological systems to which the current Monte Carlo simulations can be applied. (a) The interaction between the RecA-like recombinase RAD51 and DNA was successfully modeled using the analysis described. This showed that RAD51 binds cooperatively to DNA forming short nucleoprotein filaments (12). (b) RNA polymerase transforms a single-stranded template into a double-stranded substrate in the presence of free nucleotides. Two different pathways exist. In the most common pathway, the polymerase creates full-length templates, whereas in the other case, known as abortive initiation, the polymerase forms short oligomers. The pathways are similar to respectively a high- and low-cooperative binding mode. (e) Structural maintenance of chromosomes (SMC) proteins form condensed DNA structures by binding cooperatively to DNA holding two DNA molecules in close proximity.

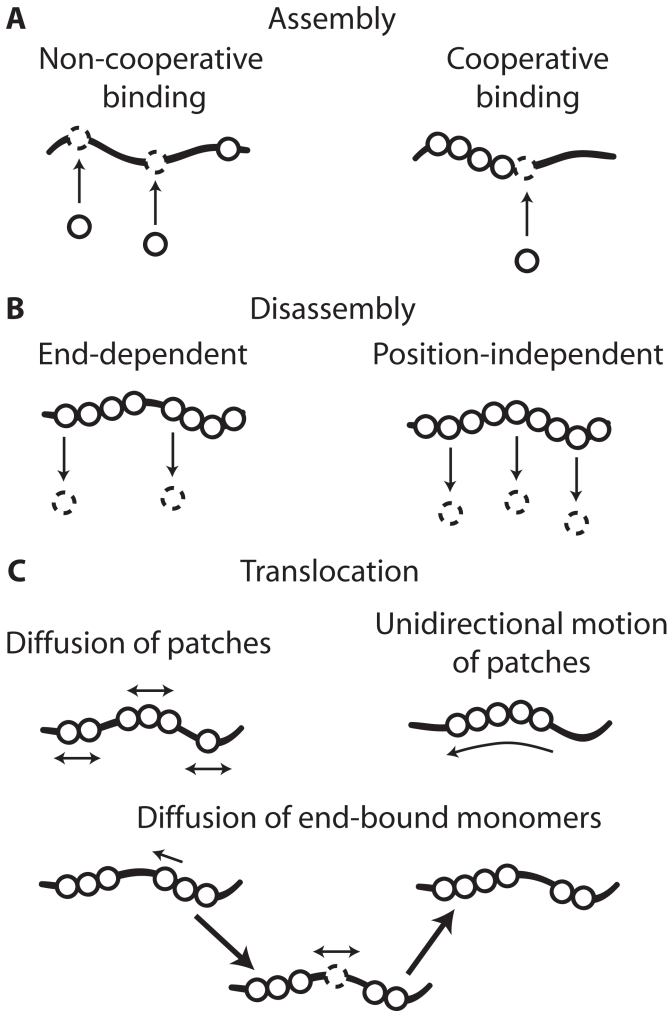
A
Assembly
Non-cooperative binding
Cooperative binding
B
Disassembly
End-dependent
Position-independent
C
Translocation
Diffusion of patches
Unidirectional motion of patches
Diffusion of end-bound monomers

Figure 1:

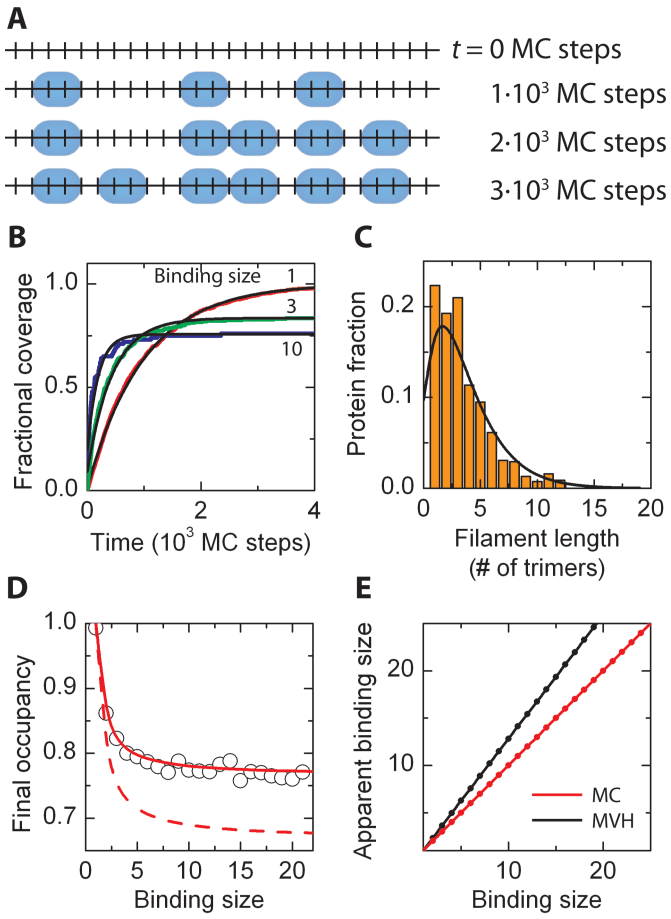
A
t = 0 MC steps
1·10³ MC steps
2·10³ MC steps
3·10³ MC steps
B
Binding size
1
3
10
Fractional coverage
1.0
0.5
0.0
0
2
4
Time (10³ MC steps)
C
Protein fraction
0.2
0.1
0.0
0
5
10
15
20
Filament length
(# of trimers)
D
Final occupancy
1.0
0.9
0.8
0.7
0
5
10
15
20
Binding size
E
Apparent binding size
20
10
MC
MVH
10
20
Binding size

Figure 2:

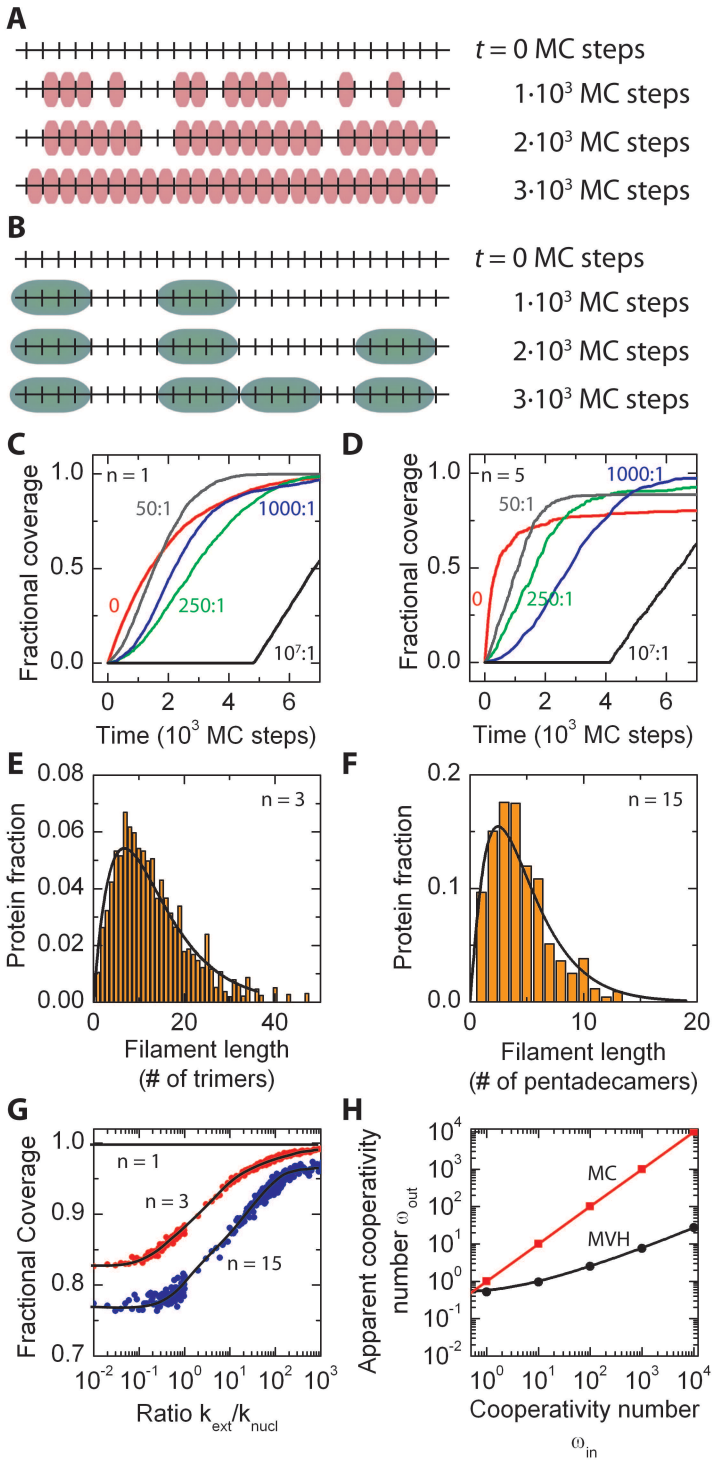
A
t = 0 MC steps
1·10³ MC steps
2·10³ MC steps
3·10³ MC steps
B
t = 0 MC steps
1·10³ MC steps
2·10³ MC steps
3·10³ MC steps
C
Fractional coverage
n = 1
50:1
1000:1
0
250:1
10⁷:1
Time (10³ MC steps)
D
Fractional coverage
n = 5
1000:1
50:1
0
250:1
10⁷:1
Time (10³ MC steps)
E
Protein fraction
n = 3
Filament length
(# of trimers)
F
Protein fraction
n = 15
Filament length
(# of pentadecamers)
G
Fractional Coverage
n = 1
n = 3
n = 15
Ratio k_ext/k_nucl
H
Apparent cooperativity number ω_out
MC
MVH
Cooperativity number ω_in

Figure 3:

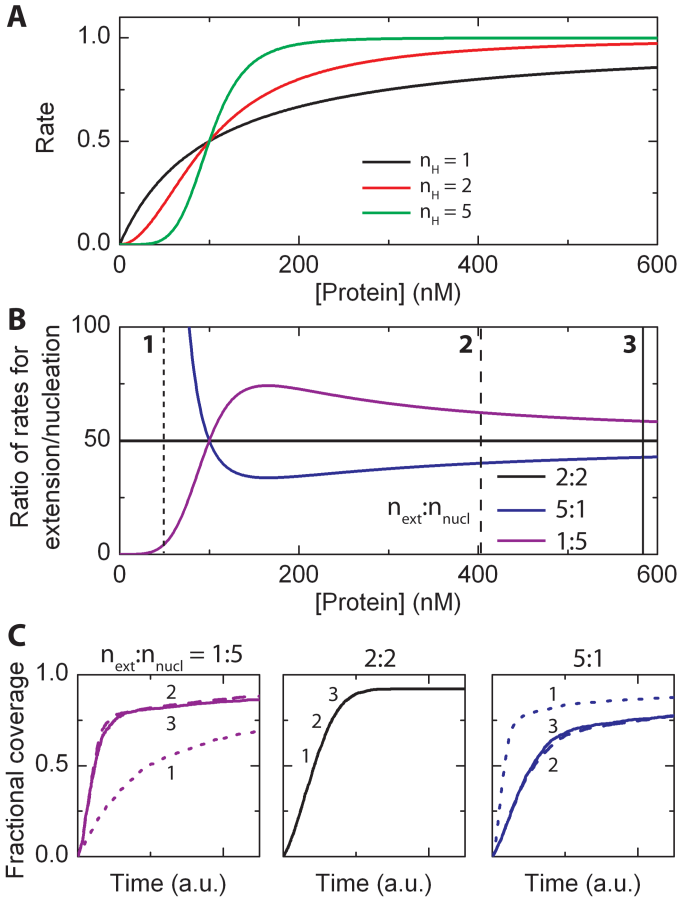
A
Rate
1.0
0.5
0.0
0
200
400
600
[Protein] (nM)
$n_H = 1$
$n_H = 2$
$n_H = 5$
B
Ratio of rates for extension/nucleation
100
50
0
1
2
3
2:2
5:1
1:5
n_{ext}:n_{nucl}
C
Fractional coverage
n_{ext}:n_{nucl} = 1:5
2:2
5:1
Time (a.u.)

Figure 4:

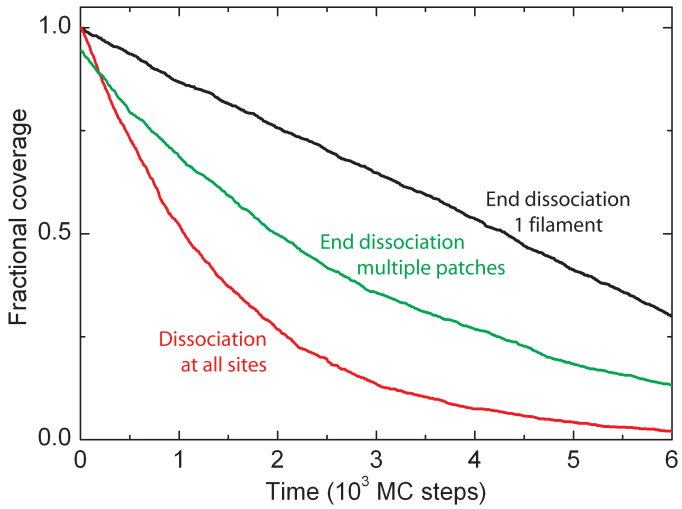
1.0
0.5
0.0
Fractional coverage
0
1
2
3
4
5
6
Time (10^3 MC steps)
End dissociation
1 filament
End dissociation
multiple patches
Dissociation
at all sites

Figure 5:

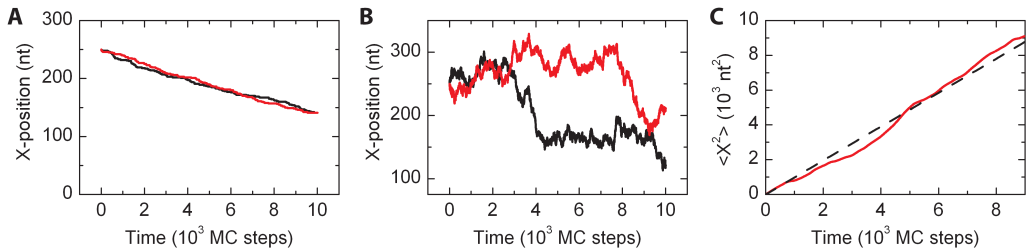

Figure 6:

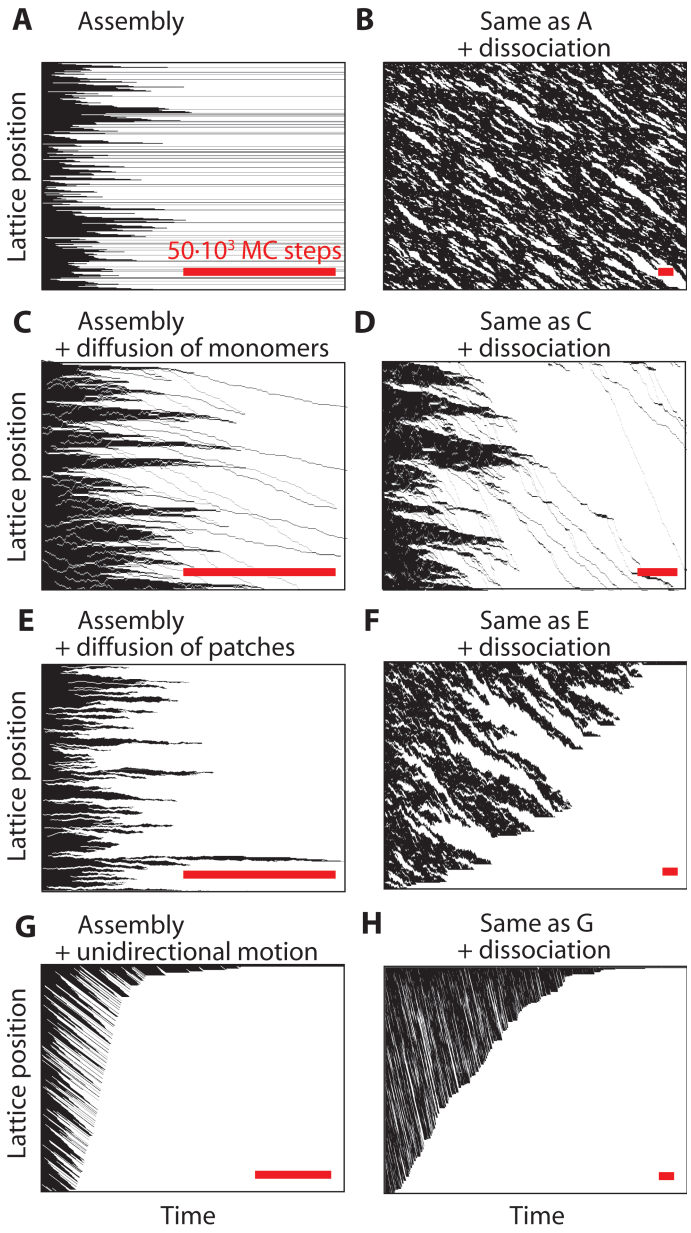
A
Assembly
B
Same as A
+ dissociation
Lattice position
50·10³ MC steps
C
Assembly
+ diffusion of monomers
D
Same as C
+ dissociation
Lattice position
E
Assembly
+ diffusion of patches
F
Same as E
+ dissociation
Lattice position
G
Assembly
+ unidirectional motion
H
Same as G
+ dissociation
Lattice position
Time
Time

Figure 7:

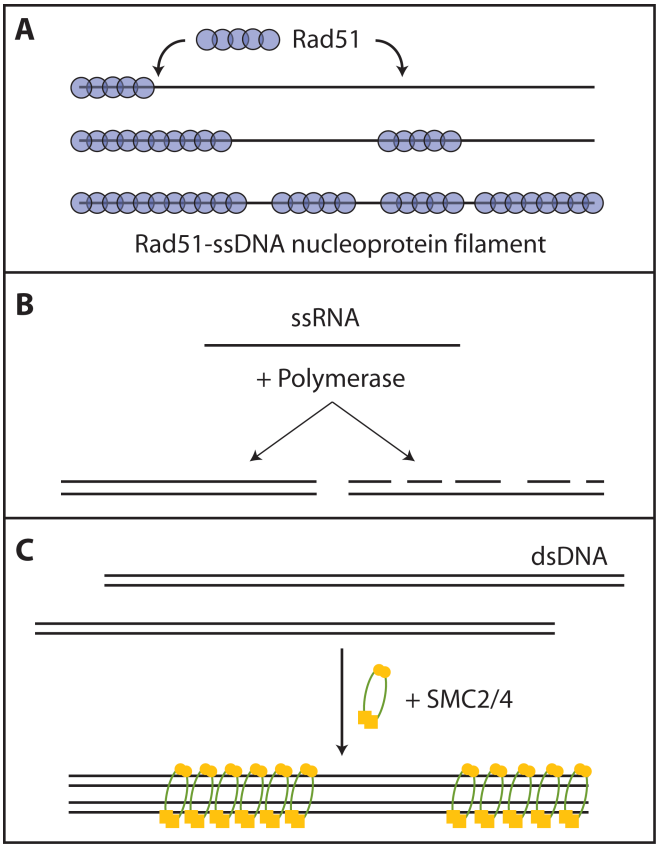
A
Rad51
Rad51-ssDNA nucleoprotein filament
B
ssRNA
+ Polymerase
C
dsDNA
+ SMC2/4

Figure 8: