HHS Public Access

Author manuscript

Phys Rev E Stat Nonlin Soft Matter Phys. Author manuscript; available in PMC 2016 August 03.

Published in final edited form as:

Phys Rev E Stat Nonlin Soft Matter Phys. 2014 August ; 90(2): 022704. doi:10.1103/PhysRevE.90.022704.

RecA-mediated sequence homology recognition as an example of how searching speed in self-assembly systems can be optimized by balancing entropic and enthalpic barriers

Lili Jiang and Mara Prentiss*

Harvard University, Department of Physics, Cambridge, Massachusetts 02138, USA

Abstract

Ideally, self-assembly should rapidly and efficiently produce stable correctly assembled structures. We study the tradeoff between enthalpic and entropic cost in self-assembling systems using RecA-mediated homology search as an example. Earlier work suggested that RecA searches could produce stable final structures with high stringency using a slow testing process that follows an initial rapid search of ~9–15 bases. In this work, we will show that as a result of entropic and enthalpic barriers, simultaneously testing all ~9–15 bases as separate individual units results in a longer overall searching time than testing them in groups and stages.

I. INTRODUCTION

A. Entropy and enthalpy in self-assembling systems

Many systems self-assemble due to decreases in free energy resulting from the correct pairing of corresponding binding sites. It is desirable that the self-assembly accurately produce stable final products. For living systems it is also important that the stable final products form on biologically relevant time scales. For systems with a few binding sites and a sparse target sample, the required speed, stability, and stringency can readily be achieved in a pairing system with only a single bound conformation, where pairing accuracy is tested by allowing all of the binding sites to simultaneously interact as separate distinct units. In contrast, for systems involving only one bound conformation and more than ~3 binding sites, there are conflicting requirements to maximize speed and stability, noted by previous theoretical work as the speed-stability paradox [1–5]. This work aims to consider the speed-stability paradox in the context of entropy and enthalpy cost. In particular, if the sites are considered separately but simultaneously, searching speed is slowed by the significant entropic barrier associated with the large number of possible states that the separate independent sites can assume during a pairing. As we will show, grouping several binding sites into one unit that is tested collectively will reduce the entropic barrier by decreasing the number of possible states; however, grouping the sites increases enthalpic barriers faced during the search since the binding energy for the entire group may be much larger than the individual binding energy for each site in the group. Lowering the energy per binding site

* prentiss@fas.harvard.edu.

PACS number(s): 87.10.Rt, 87.15.ak, 87.16.af, 87.10.Mn

would lower the enthalpic barriers resulting from grouping the sites, but the lower binding energies impair stringency and make the final product less stable.

This work will explore how to optimize the average self-assembly time by balancing the entropy and enthalpy tradeoff through strategies such as grouping or employing sequence dependent barriers that can control transitions between bound states. In particular, we consider a system inspired by RecA-mediated homology recognition as an example. We choose this system because the binding site interaction can be approximated by a one-dimensional model searching the sequence of a bacterial genome which provides a very tractable statistical distribution of energy mismatches that makes optimizing the average searching time fairly simple. The results also provide useful information about a search that is of great biological importance.

Though the physical origins differ, many self-assembling systems employ searches that divide testing into stages and/or group binding sites. These systems have sizes that vary by more than six orders of magnitude: the persistence length constraint in RNA folding that limits initial pairings to the ~ 4 -base initial interaction length [6,7]; the persistence length constraints a thread of mm-sized charged beads separated by uncharged beads [8]; and the electrostatic and hydrophobic force in protein folding [9].

The relationship between speed, stability, and stringency is also a known issue in the experimental pairing of long ssDNA-ssDNA molecules, where kinetic trapping due to pairings containing significant accidental homology makes rapid searching difficult. For example, given an average Watson-Crick pairing energy equal to approximately twice the average thermal energy, the Watson-Crick pairing of two 20 nucleotide sequences containing one mismatch would have a binding energy of almost 40 times the thermal energy; therefore, the time required to unbind this incorrect pairing would be very long. In order to avoid such kinetic trapping, pairing experiments between long ssDNA sequences are frequently done at high temperatures in buffers that reduce the Watson-Crick pairing energy. As a result, the binding energy per correct Watson-Crick pairing is lower than the average thermal energy in such systems, however, factors that reduce kinetic trapping also reduce stringency since the energy penalty per mismatch is lowered by the same factors that reduces the kinetic trapping.

B. The RecA system

RecA and its protein family promote homologous pairing and exchange of DNA strands in prokaryotic and eukaryotic organisms, a process crucial to meiosis and DNA damage repair [10–17]. The RecA protein has two strongly positively charged regions: site I and site II. During the RecA homolog search process, site I is bound to an incoming single-stranded DNA (hereafter referred to as I), which serves as the target sequence for the homology search [18]. Then, a segment of double-stranded DNA (dsDNA) from the same bacterial genome binds to site II of the ssDNA-RecA filament [Fig. 1(b)]. The ssDNA-RecA filament tests whether the dsDNA is sequence matched to the corresponding region in I. When the matched region is found, one strand in the dsDNA exchanges its base pairing from its original partner to a new partner in I [18,19] [Fig. 1(c)]. The strand that exchanges partners is called the complementary strand (C), and its original partner that is left unpaired is called

the outgoing strand (O). RecA homology recognition is believed to involve base flipping, where bases in C flip back and forth to bind with bases on I and O [20–23].

Although seminal work has shown that accurate kinetic proofreading can be achieved if the process is concluded with an irreversible process [24], RecA cannot use such strategies as strand exchange can occur in the absence of adenosine triphosphate (ATP) hydrolysis [25–27]. Hence all of the binding energies must be of the order of the thermal energy $k_B T$.

II. MODEL

A. pre-BRW and BRW stage of the search

Experimental results suggest that the RecA search process has at least two distinctive stages. It starts with a sequence-independent initial stage, where the first few bases C can flip separately to pair with their partners in I [20]. If the first ~9 bases in C have all flipped and paired with I, the strand exchange product becomes metastable [20,28] and homology recognition proceeds much more slowly as an iterative search in units of successive bp triplets [29]. The latter process can be modeled as a biased random walk (BRW) [30]. We will refer to the initial sequence-independent stage as pre-BRW.

This paper will focus primarily the pre-BRW stage. While the detailed modeling of the BRW stage can be referred to Kates-Harbeck's work [30], a summary of the BRW process is the following. Experimentally, fluorescence resonance energy transfer (FRET) has shown that after 12 ± 3 bps are bound, strand exchange proceeds iteratively in units of successive base pair triplets [29]. Thus, we consider the complementary strand as a one-dimensional array. At each step in the BRW, either the triplet at the right hand edge of the strand exchanged dsDNA flips back to the pair with the outgoing strand, or its right hand neighbor flips forward and pairs with the incoming strand. No other triplet can flip. The first process decreases the number of strand-exchanged triplets by 1, which represents a step backward in the random walk. The second process increases the number by 1, which represents a forward step in the random walk. Whether the system steps forward or backward is governed by the thermodynamic equilibrium, depending on whether or not the last flipped base in I is complementary to that in C. Previous work showed that such an iterative search of a bacterial genome could provide stringencies in excess of 10^{-20} [30]. However, the correct match frequently unbinds unless the BRW is preceded by a rapid initial search (pre-BRW) which produces a metastable product. Given a pre-BRW that establishes a metastable strand exchange product consisting of ~9 strand exchange bases, the correct initial pairing progresses to a stable final product with a probability that can easily exceed 90%.

In vivo, the final product is stable because ATP hydrolysis occurs after ~80 bps have strand exchanged [31]; *in vitro* if ATP hydrolysis is strongly suppressed, the strand exchange product can extend for thousands of bases. As a result of the bias in the BRW, once 30 contiguous bps have undergone strand exchange, the processes becomes effectively irreversible [30].

In this model, mismatched sequences that are sent back by the BRW will re-enter the pre-BRW and repeat the process until it is eventually rejected by the pre-BRW and unbinds from

the RecA protein (Fig. 2). This paper will study how the pre-BRW can be optimized to minimize the average total search time over a bacterial genome, as we discuss in the next section.

B. Total time

Given that excellent stringency is provided by the final BRW [30], we will not consider ultimate stringency but focus on the total search time.

To simplify the problem, we first consider a search process with only one decision point that occurs when 9 bps are bound to site II followed by a BRW. The total search time can be divided into three components:

$$\text{total time} = \frac{1}{r_H} (T_1 + T_{\text{pre-BRW}} + T_{\text{BRW}} + T_{B\text{-form}}), \quad (1)$$

where r_H is the probability that the homolog in C remains bound to I in the end of the search. Thus the expected number of times to repeat the search until the ssDNA is stably bound to the homolog is $\frac{1}{r_H}$. T_1 is the total time taken up by all the sequences that enter pre-BRW only once and are rejected before ever entering the BRW phase. $T_{\text{pre-BRW}}$ and T_{BRW} are the total cumulative time spent in pre-BRW and BRW taken up by sequences that have undergone the BRW at least once (Fig. 2). $T_{B\text{-form}}$ is the time during which the dsDNA is in B form and bases in C cannot flip to pair with bases in I. This includes the three-dimensional (3D) and 1D diffusion time, as well as time spent hopping between different bound positions. The division between these different modes of registration shifting is of enormous interest [32]. However, in this work we will only focus on minimizing the first two terms and do not consider $T_{B\text{-form}}$ because we assume the diffusion time is largely independent from the optimization of strand exchange process.

C. Effective absorbing boundaries

In this work, we consider homology searching on a time scale that features two forms of effective irreversibility. (1) The entry of the correct sequence into the BRW is irreversible: previous work showed that once a correctly paired sequence enters the BRW, it has a very high probability of completing strand exchange; in contrast, mismatches entering the BRW will all be reflected back to the pre-BRW after spending some time in the BRW. (2) The unbinding of the dsDNA from the ssDNA-RecA filament is irreversible: due to the large number of possible registrations between the ssDNA-RecA filament and the dsDNA being searched, the probability of returning to the same registration is very small. We also assume that the ssDNA-RecA will bind in a new registration after some time while being governed by diffusion.

D. Pre-BRW: Markov base flipping model with absorbing boundaries

We model the pre-BRW process as a Markov chain with the absorbing and reflecting boundaries described above (Fig. 3). The system randomly selects a base in C to flip to bind with O or I, if the thermal energy fluctuation is enough to break the WC bond and stacking

energy [Fig. 3(c)]. We assume population of the conformations between the absorbing boundaries remain at thermal equilibrium, as the populations slowly disappear into the absorbing boundaries. This is consistent with the experimental result that base flipping is significantly faster than unbinding time and moving forward time [33].

The model can be described by a system of first order differential equations. For a given sequence with j mismatches, consider the growth or decay of the population in unbound, intermediate, and forwarded conformations, $\eta_{\text{unbound},j}$, $\eta_{\text{int},j}$ and $\eta_{\text{pass},j}$:

$$\frac{d\eta_{\text{unbound},j}}{dt} = R_u P_{i,j} \cdot \eta_{\text{int},j}(t), \quad (2)$$

$$\frac{d\eta_{\text{int},j}}{dt} = -(R_u P_{i,j} + R_f P_{t,j}) \cdot \eta_{\text{int},j}(t), \quad (3)$$

$$\frac{d\eta_{\text{pass},j}}{dt} = R_f P_{t,j} \cdot \eta_{\text{int},j}(t), \quad (4)$$

where R_u and R_f are the unbinding and moving forward rate, and $P_{i,j}$ and $P_{t,j}$ are the equilibrium concentration of conformation i and t [Figs. 3(b) and 3(d)] for sequences with j mismatches.

If the initial condition is set to be $\eta_{\text{int},j}(0) = 1$, then the equations are solved to be

$$\eta_{\text{int},j} \rightarrow 0, \quad (5)$$

$$\eta_{\text{unbound},j} \rightarrow \frac{R_u P_{i,j}}{R_u P_{i,j} + R_f P_{t,j}}, \quad (6)$$

$$\eta_{\text{pass},j} \rightarrow \frac{R_f P_{t,j}}{R_u P_{i,j} + R_f P_{t,j}}. \quad (7)$$

For a sequence with j mismatches, let the time for each visit to the pre-BRW stage be $t_{\text{pre-BRW},j}$ which is the time before it is absorbed into either side of the Markov chain. Then $t_{\text{pre-BRW},j}$ is inversely proportional to the decaying rate of $\eta_{\text{int},j}$ namely, $R_u P_{i,j} + R_f P_{t,j}$. Explicitly,

$$t_{\text{pre-BRW},j} = \frac{\tau}{R_u P_{i,k=1,j} + R_f P_{t,k=1,j}}, \quad (8)$$

where τ is a time scaling factor (see Appendix C 1).

If a sequence passes the pre-BRW, it will enter the BRW for time

$$t_{\text{BRW}} = \lambda. \quad (9)$$

λ is the ratio of flipping time per binding site of the deeply bound BRW stage to that of the weakly bound pre-BRW stage, which depends on the relative energy penalty. For simplicity we assume λ is a sequence-independent constant (Appendix B).

Because the second boundary is not entirely absorbing, a mismatched sequence will undergo the pre-BRW and BRW process multiple times until it exits the process through unbinding. T_1 is simply expressed as

$$T_1 = \sum_{j=0}^9 N_j t_{\text{pre-BRW},j}, \quad (10)$$

where N_j represents number of nine-bp-long sequences with j mismatches (see Appendix A 2 for calculation of N_j).

Let r_j be the average number of times sequence j will undergo the BRW, then,

$$T_{\text{pre-BRW}} + T_{\text{BRW}} = \sum_{j=0}^9 N_j r_j (t_{\text{pre-BRW},j} + t_{\text{BRW}}). \quad (11)$$

E. Calculation of parameters

1. r_j —As mentioned previously, entering the BRW is an absorbing boundary for perfectly matched sequences but not for the mismatched ones. Of course, a pairing that is reflected back into the pre-BRW can return again to the BRW before unbinding. Thus, one given mismatched pairing can enter the BRW once, twice, three times, etc.

As each time η_{pass} portion of the sequences (re)enters the BRW, the numbers of sequences entering the BRW for a first, second, third...time form an infinite geometric series: $N_j \eta_{\text{pass}}, N_j \eta_{\text{pass}}^2, N_j \eta_{\text{pass}}^3 \dots$. Thus, summing the series while leaving out N_j (as it is absorbed in $t_{\text{pre-BRW}}$) gives

$$r_j = \frac{\eta_{\text{pass}}}{1 - \eta_{\text{pass}}}. \quad (12)$$

Not surprisingly, r_j is the greatest for near homologs (~ 10 times) and smallest for completely unmatched sequences. Nevertheless, the inclusion of r_j does not have a big impact on the final results.

2. P_i and P_t —For a system in thermal equilibrium P_i the probability of being in conformation i in a multiconformation system is

$$P_i = \frac{\exp(-E_i/k_B T)}{\sum_s \exp(-E_s/k_B T)}, \quad (13)$$

where E is the energy of a conformation, and the denominator is the partition function. P_t can be obtained in a similar way (see Appendix C 2 for an example of calculation).

3. R_u and R_f —If given r_H , the relative values of R_u and R_f the rates for unbinding and moving forward respectively in the Markov chain, can be determined. In particular,

$$r_H = \frac{R_f P_{t,j=0}}{R_u P_{i,j=0} + R_f P_{t,j=0}}, \quad (14)$$

where we assume after passing the pre-BRW, the homolog will 100% remain bound in the BRW. If we set r_H as 0.9, a biologically reasonable value, then R_f : R_u can be solved using the fact that $P_j = P_t$ for the homolog. Therefore Eq. (14) simply becomes

$$r_H = \frac{R_u}{R_u + R_f} = 0.9, \quad (15)$$

Thus we can set $R_u = 1$ and $R_f = 9$. Note that these are in a unitless time scale, not in real time.

4. Binding energy ratio λ —In this model, all results of total time are plotted as a function of λ , which is the ratio of flipping time per binding site in BRW to that in pre-BRW. Experimentally, strand exchange in BRW progresses at a rate of ~ 6 bp/sec [34]. In contrast, given that the entire search process over 10^7 bps is accomplished within 1 h, the average time per registration test must be less than $\frac{3600}{10^7}$ sec, including time spent in free diffusion. We therefore estimate the ratio λ to be at least $\frac{1}{6}$ sec/bp: $\frac{3600}{10^7}$ sec/bp, that is at least of the order of 1000.

III. RESULTS

To highlight a few insights we can gain from the model, we studied the effect of bp pairing energy, stacking energy, flipping group size, and location of the decision point as examples. We also consider the effect of adding two more decision points at 3 and 18 bps.

A. Base pairing energy

Figure 4(a) compares the total time of systems with different bp pairing energy. That is, how much energy is needed to break the bond of a matched base pair. For this section, we set the stacking energy to be $0k_B T$ and assume bases flip individually. The results suggest that over the range of values considered in Fig. 4(a), the optimal pairing energy increases with λ . Of course, the optimal bp pairing energy will change slightly depending on other parameters in the system, such as stacking energy.

High bp pairing energy leads to higher enthalpy, which results in lower P_f and higher P_i for heterologs. This is because a mismatched bp will not regain the bp pairing energy after breaking the bond with O and trying to pair with I. Thus a high bp pairing energy value increases the energy penalty for heterologs and is more likely to send them back. As a result, higher bp pairing energy has better stringency. However, higher bp pairing energy also requires longer equilibrium time. As the enthalpy becomes higher and less likely to be overcome, the system needs more attempts to flip to reach equilibrium.

B. Initial testing length ($N_{i,t}$)

Initial testing length is the length of bps being tested during pre-BRW. We consider whether the total search time would be faster if $N_{i,t}$ were 6, 9, 12, or 15 bps. A longer sequence results in increased entropy while the enthalpy remains the same. This leads to better stringency but worse speed. Figure 4(d) shows that a short initial sequence length is preferred for low λ range but long initial sequence length for a higher range. If $\lambda \approx 10^3$, the optimal decision point should be set at the ninth bp. This is consistent with experimental results [20,28].

C. Flipping group size (N_{flip})

We group sequences in one, two, and four bases for an eight-bp sequence. Bases in a group flip together and their homogeneity is considered simultaneously. The result [Fig. 4(c)] shows that flipping in larger bp groups leads to higher P_f and P_i due to lower entropy, and equilibrium time τ . In low λ range, grouping in 4 is better than grouping in 2, which is better than grouping in 1. In high λ range, they converge to the same total time because different group sizes yield the same stringency due to the same average penalty per mismatch. Similar as before, grouping in 8 would be prohibitively long due to the exceedingly high enthalpy to overcome, thus its result is not shown.

D. Allowing multiple bound conformations

RecA may have more than one stage in pre-BRW, with two additional decision points at 3 and 18 bp based on experimental and statistical evidence (Appendix A 1). Having explored

the properties of a one-stage search in the pre-BRW, we now will consider the case where the search contains multiple stages.

In the one-stage process shown above, all 18 bps are allowed to flip. In a two-stage process, for those sequences past the first decision point at 9, all first 18 bps can make free contact with each other. A three-stage process adds an additional decision point at 3. We simulate the case where triplets flip together, and bp pairing energy is $1k_B T$ per bp (Fig. 5). For simplicity, we will consider only the case where $\lambda = 1000$.

The total time can be generalized from Eq. (1) to be

$$\text{total time} = \frac{1}{r_H^n} \left(\sum_{k=1}^n T_k + (T_{\text{pre-BRW}} + T_{\text{BRW}}) \right), \quad (16)$$

where T_k is the average time spent in testing sequence rejected by recognition stage k , and n is the total number of stages. The reason that r_H is to the n th power is that within each stage, the unbinding and moving forward rate, R_u and R_f , are such that a homolog can pass with r_H probability. When there are n consecutive stages, a homolog will pass with r_H^n probability.

Likewise, the number of sequences passing each stage k can be generalized to be

$$N_{\text{pass},k,j} = \sum_{j'=0}^{l_k-1} N_{\text{pass},k-1,j'} \varepsilon_k(j', j) \eta_{\text{pass},k,j}, \quad (17)$$

where $k \in \{1, 2, 3\}$, l_k is the testing length of stage k measured in number of bps such that $l_1 = 3$, $l_2 = 9$, and $l_3 = 18$. $\eta_{\text{pass},k,j}$ is the probability of sequences with j mismatches in l_k bases passing stage k , given by Eq. (7). $\varepsilon_k(j', j)$ relates j' , the number of mismatches in the first l_{k-1} bases, to j , the number of mismatches in the first l_k bases based on probability distribution (Appendix C 3). For example, if a sequence has three mismatches in the first nine bases, $\varepsilon_3(3, 5)$ gives the probability that the sequence has five mismatches in the first 18 bases (in other words, two additional mismatches in the next nine bases).

Figure 6 shows that dividing the pre-BRW into two stages provides a faster search than a single stage, but that adding a third stage does not improve the searching speed. As the number of stages increases from 1 to 2, both speed and stringency are improved. This is because, in a two-stage process, RecA can combine the strength of having high speed with a shorter initial sequence length at 9 and the strength of having high stringency with a longer initial sequence length at 18. By first filtering at the ninth bp, a limited number of sequences will undergo a free-flipping process of the first 18 bps, the more entropic stage. Although the decision time per sequence is relatively long in the latter stage, overall the time is still negligible comparing to the time spent by the decision point at 9. However, when the number of stages increases from 2 to 3, stringency improves but the search consumes more

time, mainly because of the additional $\frac{1}{r_H}$ factor. Intuitively, since there are more stages, the homolog is more likely to fall out in one of the stages and the whole search needs to start again.

In general, the optimal number of stages depends on the sample being searched and the value of λ .

IV. DISCUSSION

A. Search in reasonably sized groups is faster

Grouping base pairs to flip collectively with reasonable size is shown to reduce the conflict between speed and stability. Large group leads to higher P_b , P_t but also longer equilibrium time τ . In general, the flip time scales as $\exp(N_{\text{flip}})$, where N_{flip} is the size of the flipping group. The former desirable factor outweighs the latter when single bp flipping is compared to triplet flipping. Increasing the flip size from three to all nine bps causes the exponential increase in τ to dominate over the decrease in P_b , P_t .

For systems that do not require bond breaking, the time increase will scale as the number of accidental matches rather than the number in the group, but the search times will still increase exponentially with the size of the tested group. Thus, systems other than RecA will probably optimize at larger binding site groups since they do not start off as paired and thus pay a lower binding penalty. The group size also depends on the math of the mismatches more generally.

However, in reality it is the stacking energy that determines whether the bp flips in triplet or individually. In order for bases to dominantly flip in triplets, the stacking energy must be $\sim 5 k_B T$, much larger than the generally accepted values. Thus, though flipping in triplets is advantageous because it lowers the number of intermediate states, the stacking energy in the RecA system is insufficient to provide the correlation between bases required for them to flip in units of triplets; however, this study illustrates the general advantage that can be obtained by grouping contacts in order to reduce entropic penalties.

B. Decision points may improve search speed and stringency

For the system considered here, a multistage search is the most effective way to break the speed-and-stability paradox for the overall search, although the paradox still applies locally in each stage of the search. This largely benefits from the mismatch distribution in the sample space: the drop of repeated sequences in $l = 3, 9$, and 18 is significant. Of course, up to a certain point, too densely located decision points will not serve the purpose.

In addition, previous work has considered a model system that undergoes the BRW without an initial weakly bound stage [35]. Our simulation shows that such a search cannot discriminate a single mismatch out of the first 18 bps unless the energy per binding site is extremely low; however with such small energy penalty, the homolog is also highly likely to unbind and the search needs to be repeated throughout the genome for ~ 10 times. One fundamental reason is that a BRW lacks absorbing boundaries which can control the forward and unbinding rates such that they can enhance the probability of a homology being bound

without sacrificing speed. Another reason is that unbinding heterologs in BRW takes an enormous amount of time [30].

Such division into a sample-independent weakly bound stage and a second sample-dependent deeply bound stage is not unique to RecA homology recognition. Research suggests that the sex-determining region Y protein [36], RNA polymerase processive transcriptional dynamics [37], and zinc-finger proteins [38] may search for its target DNA in a similar manner, employing a “conformational switch” leading to different stages (or in their language, “search mode” and “recognition mode,” or “sequence-dependent pausing”).

In general, adding another decision point is advantageous if the overall additional searching time required to overcome the sequence dependent barrier that creates the additional decision point is smaller than the overall reduction in searching time due to mismatched pairings that are rejected at the added decision point without allowing more binding sites to interact.

V. CONCLUSION

Using RecA system as an example, we have shown that, consistent with previous work, the statistical distribution of energy mismatches within the sample being searched plays a crucial role in determining optimal self-assembly strategies [1], including the number of sequence dependent barriers that should be included and how many sites should be tested in order to pass each barrier.

The results also illustrate that in a system with only one bound state, if the contacts are required to be tested simultaneously, grouping the binding sites is advantageous as long as the entropic benefit due to the reduction in the number of possible conformations exceeds the enthalpic penalty. The scaling of the number of states and the enthalpic penalty is readily obtained, but the benefit in search time also depends on the energy distribution characteristic of the mismatches present in the sample.

We have also shown that allowing many contacts to simultaneously attempt binding is a poor strategy for recognition because all of those simultaneous contacts result in large entropy. For searching samples where the number of accidental mismatches decreases sufficiently fast when the number of binding sites considered increases, it is far better to perform an iterative search in which only a few contacts are initially allowed to interact. That initial test must then be followed by subsequent tests involving one or more sequence transitions between bound conformations that require passing through sequence dependent barriers. The tests involving only a few initial contacts have low entropic and enthalpic penalties, allowing the test of those contacts to be very rapid; however, the sequence dependent barriers greatly increase the time required to test sequences that pass through the barriers. Thus, providing additional decision points to filter out mismatches as early as possible is advantageous if the rapid rejection of most mismatches more than compensates for the longer testing times required for sequences that pass through the barriers; however, eventually adding more decision points is no longer useful and may even result in longer searching times.

Acknowledgments

We thank Y. Kafri for insightful feedback and B. Drauschke for proofreading. We thank D. Yang for preparing Figs. 1(a) and 1(b), which were rendered using VMD developed by the Theoretical and Computational Biophysics Group in the Beckman Institute for Advanced Science and Technology at the University of Illinois at Urbana-Champaign [39]. Research was supported by NIH Grant No. NIH/NIGMS-R01-GM025326 to M.P. and N. Kleckner. L.J. acknowledges financial support from Harvard Physics Department and Harvard College Research Program.

References

1. Bénichou O, Kafri Y, Sheinman M, Voituriez R. Phys Rev Lett. 2009; 103:138102. [PubMed: 19905543]
2. Sheinman M, Bénichou O, Kafri Y, Voituriez R. Rep Prog Phys. 2012; 75:026601. [PubMed: 22790348]
3. Slutsky M, Mirny L. Biophys J. 2004; 87:4021. [PubMed: 15465864]
4. Gerland U, Moroz J, Hwa T. Proc Natl Acad Sci USA. 2002; 99:12015. [PubMed: 12218191]
5. Veksler A, Kolomeisky AB. J Phys Chem B. 2013; 117:12695. [PubMed: 23316873]
6. Chen H, Meisburger SP, Pabst SA, Sutton JL, Webb WW, Pollack L. Proc Natl Acad Sci USA. 2011; 109:799. [PubMed: 22203973]
7. Kuznetsov SV, Ansari A. Biophys J. 2012; 102:101. [PubMed: 22225803]
8. Feinstein, E. PhD thesis. Harvard University; 2009.
9. England J. Structure. 2011; 19:967. [PubMed: 21742263]
10. Clark A, Margulies A. Proc Natl Acad Sci USA. 1965; 53:451. [PubMed: 14294081]
11. Weinstock GM, McEntee K, Lehman IR. Proc Natl Acad Sci USA. 1979; 76:126. [PubMed: 370822]
12. McEntee K, Weinstock GM, Lehman IR. Proc Natl Acad Sci USA. 1979; 76:2615. [PubMed: 379861]
13. Shibata T, Cunningham RP, DasGupta C, Radding CM. Proc Natl Acad Sci USA. 1979; 76:5100. [PubMed: 159453]
14. Cassuto E, West SC, Mursalin J, Conlon S, Howard-Flanders P. Proc Natl Acad Sci USA. 1980; 77:3962. [PubMed: 6449004]
15. Cox MM, Lehman IR. Proc Natl Acad Sci USA. 1981; 78:6018. [PubMed: 6273839]
16. Roca A, Cox M. Mol Biol. 1990; 25:415.
17. Kowalczykowski S, Eggleston A. Annu Rev Biochem. 1994; 63:991. [PubMed: 7979259]
18. Chen Z, Yang H, Pavletich N. Nature (London). 2008; 453:489. [PubMed: 18497818]
19. Cox MM. Annu Rev Microbiol. 2003; 57:551. [PubMed: 14527291]
20. Xiao J, Lee A, Singleton S. ChemBioChem. 2006; 7:1265. [PubMed: 16847846]
21. Bazemore L, Folta-Stogniew E, Takahashi M, Radding C. Proc Natl Acad Sci USA. 1997; 94:11863. [PubMed: 9342328]
22. Adzuma K. Genes Dev. 1992; 6:1679. [PubMed: 1516828]
23. Stasiak A. Mol Microbiol. 1992; 6:3267. [PubMed: 1484482]
24. Hopfield J. Proc Natl Acad Sci USA. 1974; 71:4135. [PubMed: 4530290]
25. Mazin A, Kowalczykowski S. Proc Natl Acad Sci USA. 1996; 93:10673. [PubMed: 8855238]
26. Menetski J, Bear D, Kowalczykowski S. Proc Natl Acad Sci USA. 1990; 87:21. [PubMed: 2404275]
27. Rosselli W, Stasiak A. J Mol Biol. 1990; 216:335. [PubMed: 2147722]
28. Hsieh P, Camerini-Otero C, Camerini-Otero R. Proc Natl Acad Sci USA. 1992; 89:6492. [PubMed: 1631148]
29. Raganathan K, Joo C, Ha T. Structure. 2011; 19:1064. [PubMed: 21827943]
30. Kates-Harbeck J, Tilloy A, Prentiss M. Phys Rev E. 2013; 88:012702.
31. van Loenhout MTJ, van der Heijden T, Kanaar R, Wyman C, Dekker C. Nucl Acids Res. 2009; 37:4089. [PubMed: 19429893]

32. Cherstvy AG, Kolomeisky AB, Kornyshev AA. J Phys Chem B. 2008; 112:4741. [PubMed: 18358020]
33. Folta-Stogniew E, O'Malley S, Gupta R, Anderson K, Radding C. Mol Cell. 2004; 15:965. [PubMed: 15383285]
34. Peacock-Villada A, Yang D, Danilowicz C, Feinstein E, Pollack N, McShan S, Coljee V, Prentiss M. Nucl Acids Res. 2012; 40:10441. [PubMed: 22941658]
35. Savir Y, Tlusty T. Mol Cell. 2010; 40:388. [PubMed: 21070965]
36. Bouvier B, Zakrzewska K, Lavery R. Angew Chem Intl Ed. 2011; 50:6516.
37. Depken M, Galburt EA, Grill SW. Biophys J. 2009; 96:2189. [PubMed: 19289045]
38. Iwahara J, Levy Y. J Phys Chem B. 2013; 117:13005. [PubMed: 23668488]
39. Humphrey W, Dalke A, Schulten K. J Mol Graphics. 1996; 14:33.
40. The Wellcome Trust Sanger Institute; <http://www.sanger.ac.uk/resources/downloads/bacteria> [accessed Jun 2013]
41. [accessed Jun 2013] ecogene E. Coli Genome Sequence Download. www.ecogene.org/?q=ecodownload/sequence

APPENDIX A: GENOME STATISTICS

1. Repeated sequences

We computationally searched several sequenced bacterial genomes to determine the number of bp sequences that have a repeated segment somewhere in the genome (Fig. 7). In all bacteria studied, a sudden drop of repeated sequences at $l \approx 18$ was observed. After $l \approx 18$, the number of repeated sequences decreased slowly. This agrees well with probability calculation, assuming the bases are independent:

$$\text{number of repeated sequences} \approx N \left[1 - \left(\frac{4^l - 1}{4^l} \right)^N \right], \quad (\text{A1})$$

where N is the size of the genome in bp, and l is the length of the sequence.

As the probability of an accidental match of length l decreases strongly with l , increased stringency is achieved by testing over longer lengths until $l = 15-18$ bps. Beyond 18 bps, a BRW that tests the triplet sequentially until ~ 80 bps would offer a better search mechanism than continuing the simultaneous test over larger length scales.

2. Mismatch frequencies

We computationally searched bacterial genomes to obtain the frequencies of having j mismatches out of the first l bps, shown as stars in Fig. 8. The results agree well with the binomial distributions, $\text{Bin}(l, \frac{3}{4})$.

We will use this result extensively in our model. Fast searches may allow some mismatches to have very long unbinding times, as long as those mismatches rarely occur in the sample.

The search time depends not only on the number of mismatches, but also on the number of accidental matches $[\sim \text{Bin}(l, \frac{1}{4})]$. Accidental matches increase the dwell time in the strand

exchanged state. This increased dwell time can cause the system to become kinetically trapped in the strand exchanged state. Similar issues arise in any self-assembling system. It results in a substantial increase in search time as a function of the number of binding sites tested even in systems that do not begin initially paired and thus do not require initial bond breaking in order to establish pairing.

As the most likely number of mismatches in a nine-bp sequence is 7 (Fig. 8), the most probable mismatch must overcome a 14-kT barrier to progress beyond the first decision point. Thus, almost all of the most probable pairings will unbind without moving forward. In contrast, even under the assumption that the strand exchange of matched sequences involves no free energy change, perfect matches can have a high probability of moving forward if we choose suitable values of R_u and R_j . For most of this work, we chose R_u and R_j values that allow 90% of the correct matches to progress forward.

APPENDIX B: EXPLANATION FOR THE SEQUENCE-INDEPENDENT λ

In the model, we assume λ is sequence-independent. This is a quite valid assumption since for all heterolog sequences, the homology after a decision point is independent of that before the decision point. The number of sequences passing the j th BRW step scales off as $(\frac{1}{64})^j$. Thus the time spent in BRW is proportional to the number of sequences passing the decision point multiplied by the time each sequences spends in BRW. In fact, as our previous work has shown, most heterologs will be rejected in the first step of the BRW because of the presence of mismatches in the first triplet to be tested in BRW [30].

The only exception to the independence of homology before and after the decision point is when the bases before the decision point are perfectly matched. Because the genome is a finite sample, and the target ssDNA has a matching partner in the complementary strand with certainty, if mutation is not taken into account, when sequences are completely matched before the decision point, the average time spent in BRW for all such sequences is likely to be much longer because of the existence of the homolog. However, since the model is only concerned with the time spent on heterologs, a sequence-independent λ is still a valid assumption.

APPENDIX C: MORE ON THE CALCULATION OF THE PARAMETERS

1. τ

Note that t_H is expressed in terms of the time for the nine bases to reach equilibrium, yet systems with different energy settings reach equilibrium on different time scales. τ is a time scaling factor to account for the difference. For a given set of energy parameters, τ is given by the number of Monte Carlo steps required to reach equilibrium for the slowest type of sequences, which is the homolog.

The differential equation approach is only accurate if the time required for the flipping bases to reach equilibrium is much faster than the time required to change the number of bound triplets either by unbinding or by binding. To calculate the scaling factor τ , we assume all bases are in the i to begin with. We record the number of sequences in the initial

conformation, $P_i(t)$, at each time step t . We define τ as the time to reach equilibrium, by the first time $P_i(t)$ begins to consistently fall within 95–105% of the equilibrium P_i . τ is different for different flipping systems. We use simulation to obtain τ 's for all the mechanisms we discuss here.

2. P_i and P_t

P_i is evaluated by

$$P_i = \frac{\exp(-E_i/k_B T)}{\sum_s \exp(-E_s/k_B T)}. \quad (C1)$$

As an example, let us consider a system where each Watson-Crick pairing is 2 kT and bases flip individually and do not stack with each other (case A1). For a perfectly matched sequence, no matter how many bases have been flipped, any conformation will have the same Boltzmann factor, $\exp(-E_i/k_B T) = \exp(0)$. As there are $2^9 = 512$ possible conformations in total, $P_i = P_t = \frac{1}{512}$.

For a less trivial example, let us consider a nine-bp-long sequence with two mismatches. All states have one of the three energies: $\exp(0)$ if neither of the mismatched bps is flipped, $\exp(-2)$ if one of them is flipped, and $\exp(-4)$ if both are flipped. To calculate among the

$\sum_{l=0}^9 \binom{9}{l}$ states how many have energy $\exp(-2)$, we need to find among the 512 conformations how many have one and only one of the two mismatches flipped. Among all

$\binom{9}{1}$ cases where one base is flipped in total, $\binom{2}{1}$ of them have the mismatched one

flipped; among all $\binom{9}{2}$ cases where two bases are flipped in total, $\binom{2}{1} \binom{7}{1}$ of them have one of the two mismatched ones flipped and one of the seven matched ones flipped, etc.

Thus adding them together, the total number of states that are of energy $\exp(-2)$ is

$\binom{2}{1} \sum_{l=0}^7 \binom{7}{l} = \binom{2}{1} 2^7$. Similarly, the total number of states that are of energy $\exp(-4)$

are $\binom{2}{2} \sum_{l=0}^7 \binom{7}{l} = \binom{2}{2} 2^7$. The rest of the states are all of energy $\exp(0)$, and the number of such states can be obtained by subtracting the above two cases from the total number of states. Therefore the canonical partition function Z for $j = 2$ is

$$Z = 2^9 + \left[(\exp(-2) - 1) \binom{2}{1} + (\exp(-4) - 1) \binom{2}{2} \right] 2^7$$

$$\Rightarrow P_i = \frac{1}{Z} = 0.0061, \quad P_t = \frac{\exp(-4)}{Z} = 0.0001. \quad (C2)$$

In general, Z is

$$Z = 2^L + \sum_{l'=1}^m (\exp(-2l' - 1)) \binom{m}{l'} 2^{L-m}, \quad (C3)$$

where L is the sequence testing length of the search stage. Calculation of P_j and P_l for the other cases is similar.

As it can be seen, a homologous sequence will have higher P_l and a complete heterolog (all first nine bases are mismatched) will have lower P_j . If we consider tension and sequence-dependent stacking, they also depend on the location of the mismatches.

3. $\varepsilon_k(j', j)$

$\varepsilon_k(j', j)$ relates $jprime$, the number of mismatches in the first l_{k-1} bases, to j , the number of mismatches in the first l_k bases based on probability distribution. For example, there are $N_{pass, k=2, j=2}$ sequences which passed the second decision point at nine bases ($k=2$) with two mismatches ($j=2$). $\varepsilon_3(2, 3)$ gives the probability that these sequences will have three mismatches in total in the first 18 bases, which is equivalent to the probability of having one mismatch in nine bases, assuming statistical independence. $\varepsilon_3(2, 3)$ would simply be

$\binom{9}{1} \frac{3}{4} \left(\frac{1}{4}\right)^8$. It is also evident to see that for $j' > j$, $\varepsilon_k(j', j) = 0$. Thus generally, $\varepsilon_k(j', j)$ is given by

$$\varepsilon_k(j', j) = \begin{cases} \binom{l_k - l_{k-1}}{j - j'} \left(\frac{3}{4}\right)^{l_k - l_{k-1} - (j - j')} \left(\frac{1}{4}\right)^{j - j'}, & \text{if } j > j' \\ 0, & \text{otherwise.} \end{cases} \quad (C4)$$

Therefore, $\sum_{j'=0}^9 N_{pass, k=2, j'} \varepsilon_3(j', j)$, for example, gives the total number of sequences passed from stage 2 that have j mismatches in the first 18 bases. This is proportional to the decision time spent in stage 3.

FIG. 1.

(Color online) (a) VMD rendering of the outgoing, complementary, and incoming strands in cyan, purple, and orange; (b) the three strands bound to successive RecA proteins shown in white and green; (c) one-dimensional illustration of the strand exchange process. Short lateral black lines connecting two strands represent base pairing. Ovals represent RecA protein; orange (left) and green (right) part of the oval represent binding site II and I.

FIG. 2.

Cartoon representation of the homology recognition process for one single attempted pairing divided according to homology. The times shown, $T_{\text{pre-BRW}}$, T_{BRW} , and T_1 , are the cumulative time taken up by all near-homologs and complete heterologs respectively.

FIG. 3.

(Color online) Flipping conformations during the initial sequence recognition stage. (a) Unbound conformation: dsDNA unbinds from the RecA protein (not shown) at rate R_u . (b) Nine bases in conformation i . (c) Intermediate states: complementary strand bases flip back and forth between pairing with outgoing and incoming strand bases. Note that the diagram only shows one possible mixed conformation out of many. (d) Nine bases in conformation t . (e) First nine bases are in conformation f and the nucleoprotein filament binds a fourth triplet in the dsDNA in conformation i . Sequences in (d) move forward at rate R_f . Panels (a) and (e) are two absorbing boundaries.

FIG. 4.

(Color online) Total search time for various bp pairing energy, flipping group size, and initial sequence length, as a function of λ , the BRW penalty.

FIG. 5.

The decision tree of a three-stage homology recognition process. Each decision point passes on sufficiently matched I_k -base-long sequences to the next stage. Later stages take longer time to make decisions.

FIG. 6.

Comparison of homology recognition of 18 bps in one, two, and three stages. (a) Number of sequences entering BRW which starts at the 18th bp. (b) Total search time. x axis: 1 represents only one decision at 18; 2 represents two decision points at 9 and 18 bp; 3 represents three decision points at 3, 9, and 18 bp.

FIG. 7.

Number of sequences that are repeated somewhere in the bacterial genome as a function of sequence length. (a) *E. coli* [40]; (b) *E. coli* [41]; (c) *Pseudomonas* [41]; (d) *Neisseria* [41].

FIG. 8.

(Color online) (a) Frequency of L -bp long sequences with j mismatches. Magenta circle line represents the calculated frequency for $L=3$: given a target three-bp long sequence, the probabilities of a random three-bp sequence that has 0, 1, 2, and 3 mismatched base(s) from the target. Green is for $L=6$; red is for $L=9$; black is for $L=12$; cyan is for $L=15$. The circles show the calculated frequencies, whereas the blue stars show the actual frequencies from a sequenced *E. coli* genome [40]. (b) Frequency of L -bp long sequences with $L-j$ accidental matches.