

## Supplemental Information

### RecA-Mediated Homology Search

#### as a Nearly Optimal Signal Detection System

Yonatan Savir and Tsvi Tlusty

## Supplemental Experimental Procedures

### Bayesian Detection Cost

The average Bayesian cost,  $B$ , is (Helstrom, 1995),

$$B = \sum_{i,j} C_{ij} p_{ij} = \sum_{i,j} C_{ij} p_h(j) p_d(i|j), \quad (1)$$

where  $p_{ij}$  is the probability of receiving an input,  $j$ , and making a decision,  $i$ , and  $C_{ij}$  is the weight assigned to such a decision, which measures its consequences on the system. The value  $p_h(j)$  is the probability of receiving input,  $j$ , and  $p_d(i|j)$  is the conditional probability of making a decision,  $i$ , given an input,  $j$ . A Bayesian decision rule is obtained by minimizing the average cost (1) with respect to the conditional decision probabilities,  $p_d$ . Consider a system that has two possible inputs (+/-) and two possible decisions (+/-). Since  $\sum_i p_d(i|j) = 1$ , we find that the average cost can be expressed as

$$\begin{aligned} B &= C_{+,+} \cdot p_h(+ ) \cdot p_d(+ | +) + C_{-,+} \cdot p_h(+ ) \cdot (1 - p_d(+ | +)) + C_{+,-} \cdot p_h(- ) \cdot p_d(+ | -) + C_{-,-} \cdot p_h(- ) \cdot (1 - p_d(+ | -)) \\ &= -c_+ \cdot p_h(+ ) \cdot p_d(+ | +) + c_- \cdot p_h(- ) \cdot p_d(+ | -) + \alpha, \end{aligned} \quad (2)$$

where  $c_+ = C_{-,+} - C_{+,+}$ ,  $c_- = C_{+,-} - C_{-,-}$ , and  $\alpha = C_{-,+} \cdot p_h(+ ) + C_{+,-} \cdot p_h(- )$ . It is clear that  $\alpha$  does not depend on the decision probabilities,  $p_d$ , or, in signal detection terminology, it is independent of how we assign points in the observation space (Helstrom, 1995). Thus,  $\alpha$  is irrelevant for the optimization problem and minimizing (2) amounts to minimizing,

$$C_b = -c_+ \cdot p_h(+ ) \cdot p_d(+ | +) + c_- \cdot p_h(- ) \cdot p_d(+ | -). \quad (3)$$

$p_h(+)$  and  $p_h(-)$  are the probabilities that the NPF encounters as input, a complementary or a non-complementary dsDNA segment, respectively.  $p_d(+,+)$  and  $p_d(+,-)$  are the probabilities to proceed with HR given that the

encountered dsDNA segment is complementary or non-complementary, respectively. Both  $c_+$  and  $c_-$  are positive, and, therefore, increasing  $p_d(+,+)$  reduces the cost while increasing  $p_d(+,-)$  increases the cost.

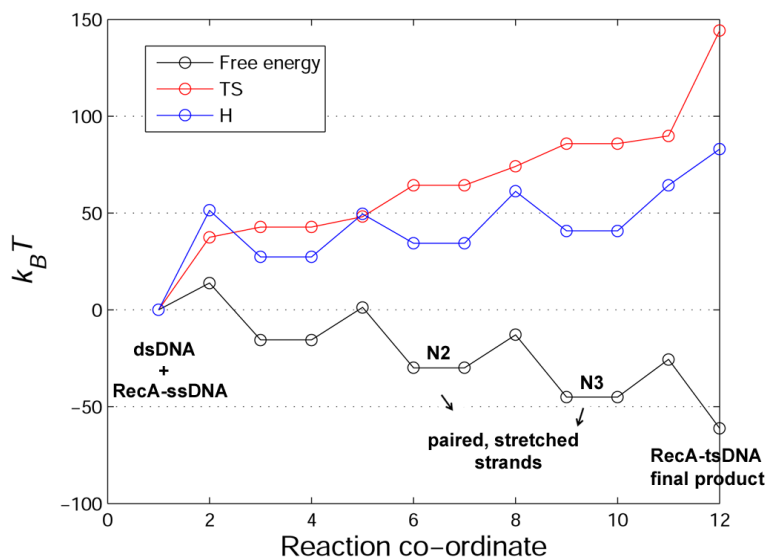
At the molecular level, the decision step involves the binding of a NPF segment to the corresponding dsDNA segment at a probability  $p_b$ , whereas  $p_f$  is the probability that the formed complex is functional and that the HR process follows. Thus, the conditional decision probability,  $p_d$  is the product of the binding probability and the functionality probability,  $p_d = p_b \cdot p_f$ , and the cost function takes the form of Eq.1 in the MS,

$$C_b = -c_+ \cdot p_h(+)^+ \cdot p_b(+)^+ \cdot p_f(+)^+ + c_- \cdot p_h(-)^- \cdot p_b(-)^- \cdot p_f(-)^- \quad (4)$$

### General step size

The relation between  $\Delta G_t$ ,  $\Delta G_{ext}$  and  $\Delta G_b$  does depend on the number of *inter-triplet* extensions, which may vary for different  $N$  values. For example, if the step size is 5, there could be either one or two extensions in one step. In general, the number of extensions is between  $\text{int}(N/3)$  and  $\text{int}(N/3) + 1$ , where  $\text{int}()$  denotes the integral part.

### Experimental value of the binding free energy.



**Figure S1. The Experimental Value of the Binding Free Energy, Related to the Experimental Procedures.** The free energy changes during HR in the presence of ATP $\gamma$ S, a non-hydrolyzable ATP analogue. The data were

reproduced from Table 3 in (Xiao et al., 2006). In this experiment, a 30-nt ssDNA was pre-incubated with RecA to form a nucleo-protein filament (NPF). This NPF was mixed with homologous dsDNA. The enthalpy  $H$ , the entropy  $S$ , and the resulting free energy are plotted.  $N_2$  and  $N_3$  are intermediates in which all three strands are in an extended, stretched conformation where the incoming ssDNA is paired with its complementary counterpart. The free energy difference between the initial state of dsDNA and NPF in solution and these intermediates is  $30 k_B T$  and  $45 k_B T$ , respectively, which correspond to a  $\Delta G_t$  value of about  $1 - 1.5 k_B T$ . The free energy of the final product, which involves strand displacement, is  $60 k_B T$  per segment, or  $2 k_B T$  per bp, which sets an upper bound on  $\Delta G_t$ .

### Estimate of the tolerance parameter for HR

The tolerance parameter is defined as  $t = [p_h(+)\cdot p_f(+)\cdot c_+]/[p_h(-)\cdot p_f(-)\cdot c_-]$ , where  $p_h(-)$  and  $p_h(+)$  are the probabilities that the NPF encounters a complementary or a non-complementary dsDNA segment, respectively,  $p_f$  is the probability that the formed complex is functional, and  $c_+$  and  $c_-$  are the costs assigned with correct and incorrect decisions, respectively.

*Evaluating  $p_h$ :* If we assume that base pairs are distributed randomly, then the ratio between the probability of finding a segment with  $m > 0$  mismatches and finding a segment with  $m = 0$  mismatches is  $3^m \binom{N}{m}$ . For  $N = 3$ ,  $p_h(+)/p_h(-) = 1/9, 1/27, 1/27$  for  $m = 1, 2, 3$ , respectively.

*Evaluating  $p_f$ :* In general, it is unreasonable that a non-complementary complex would be more functional than a complementary one. Hence,  $p_f(+)/p_f(-)$  is always larger than 1. The value of  $p_f$  can be evaluated from the rate at which non-homologues recombination takes place. For example, it was shown that introducing 6 mismatches into of 83 nucleotide long substrate results in 3-fold decrease in the rate whereas introducing 12 mismatches results in a 6-fold reduction (Bazemore et al., 1997). So a rough estimate for  $p_f(+)/p_f(-)$  is 3-10.

*Evaluating  $c_+$  and  $c_-$ :*  $c_+$  could be evaluated from the fitness cost of mutant strains that lack RecA. For example, upon exposure to UV radiation pulse, the fraction of bacteria population without RecA that died was 5-fold larger than the dying fraction in wild type (Khidhir et al., 1985). However, the fitness cost due to the absence of RecA depends on the variety of stresses (UV radiation, High temperature, etc.) the organism has to cope with, on their intensity and on their abundance. Still, since eventually, any damage to the DNA results in mutations we may perform a general estimation.

Let us assume that due to a certain stress (e.g., UV radiation), a segment of the DNA,  $B$  nucleotides in length, is damaged. In the absence of RecA, this damage is not recovered, and therefore  $B$  mutations are introduced into the DNA. If the size of the damaged segment  $B$  is larger than the step size  $N$ , the fitness cost *per step* of not performing a recombination step is  $c_+ = N \times$  (fitness cost of a mutation). If on the other hand, an incorrect recombination was performed  $m$  mismatches are introduced and the fitness cost of that mistake is  $c_- = m \times$  (fitness cost of a mutation). In this case,  $c_+/c_- = N/m$ . In general,  $c_+/c_- = \min(B,N)/m$ . Thus,  $c_+/c_-$  is expected to be larger than one and for  $N = 3$ ,  $c_+/c_- = 3, 3/2, 1$ , for  $m = 1, 2, 3$ , respectively.

Taking all this into account we find that the lower bound on  $t$  is  $t > 1/3, 1/18, 1/27$  for  $m = 1, 2, 3$ , respectively. Taking into account  $p_f(+)/p_f(-)$  which is of the order of 3-10 we get that  $t$  further increases to be of around  $t \sim 1-3$  for  $m = 1$  (and  $1/10-1/2$  for the other cases). From this rather rough estimate it appears that the relevant values of  $t$  are higher than the critical value and qualitatively behave as the symmetrical case.

### The prominent competitor

We assume that the competitor sequences are distributed randomly. The probability of finding a random dsDNA of length  $N$  and of  $m$  mismatches is therefore,  $p_i(N,m) = (3^m/4^N) \binom{N}{m}$ . Clearly, competitors with more mismatches are more abundant than those with less mismatches (as long as  $m < (3/4)N$ ). By direct calculation of  $C$ , we found that, in the experimental range of  $3 < N < 10$ , the prominent competitor is such that the possible optimal free energy extension lies within the range of measured values. It is also interesting to consider the limit of large  $N$  values in which the prominent competitor is the one with  $m = (3/4)N$ . In this case, we found from this calculation that the optimal free energy  $\Delta G_t$  is no larger than  $(3/4)\Delta G_{mis} \approx 1.8 k_B T$ .

## Supplemental References

Bazemore, L. R., Folta-Stogniew, E., Takahashi, M., and Radding, C. M. (1997). RecA tests homology at both pairing and strand exchange. *Proc Natl Acad Sci U S A* *94*, 11863-11868.

Helstrom, C. W. (1995). *Elements of Signal Detection and Estimation*, Prentice-Hall).

Khidhir, M. A., Casaregola, S., and Holland, I. B. (1985). Mechanism of transient inhibition of DNA synthesis in ultraviolet-irradiated *E. coli*: inhibition is independent of *recA* whilst recovery requires RecA protein itself and an additional, inducible SOS function. *Mol Gen Genet* *199*, 133-140.

Xiao, J., Lee, A. M., and Singleton, S. F. (2006). Direct evaluation of a kinetic model for RecA-mediated DNA-strand exchange: the importance of nucleic acid dynamics and entropy during homologous genetic recombination. *ChemBiochem* *7*, 1265-1278.