draft

Pete Rigas

October 28, 2020

1 Introduction

Families of CRISPR proteins have attracted significant attention in recent studies, a few of which have devoted attention towards the regulation of microbial metabolic rate [19], genome engineering applications in Cas9 & Cas12 [6,11,13,15,17,27,35], dynamic imaging of telomeres [7], theoretically driven predictions of protein folding and atomistic simulation [22,24,25,34], and kinetic models detailing the process by which different Cas proteins identify target sequences and impact vital cellular functions [8-9,12,14,20,21,28]. In [29], a thermodynamic approach to model binding determinants is introduced, which is primarily rooted in analyses of the individual stages of Fn Cas12a binding which is comprised of PAM & crRNA inspection, followed by a reconfiguration stage. In the discussion, the authors reflect upon potential generalizations of the thermodynamics approach, particularly in making use of the formalism to interpret binding activity of catalytically active Cas12a nuclease.

Despite other studies which have implemented machine learning techniques to simulate trajectories in the energy landscape in efforts to provide analyses of binding for different proteins [10,16,30], as well as first principled models quantifying the expression of genes through transcription factors [5,33], generalizing the thermodynamic approach of [29] is advantageous in providing more interpretations of individual stages of binding for different Cas proteins which are known to variably depend on the random walk motion that the protein undergoes throughout the PAM inspection phase [29,32], in addition to blunt versus staggered cuts that are characteristic of Cas9 & Cas12, respectively [17,31,35]. To systematically quantify the rates at which particles diffuse across subsequent base pairs to an absorbing boundary as the Cas protein inspects a target sequence for complementarity, a dimensionless ODE from a well posed IVP is solved to obtain exit times. With numerical approximations to the solution, numerical approximations of the exit time of variable absorbing boundary length are obtained through studies of Fokker Plank type equations, whose IVPs can be placed into correspondence with those of the Langevin equation [21].

Computations of mean exit, or passage, times have been previously applied under diverse geometrical and biological constraints, with one study detailing a procedure for the reconstruction of drift terms through a change of variables transformation of the backward Kolmogorov equation to the Schrodinger equation, in addition to a mapping into the Euler Lagrange equations which recovers potentials [2,8]. Numerical manipulations of the solution to the ODE for numerical approximations to the first passage time can be readily adapted to obtain passage times across other base pairs in the target sequence by numerically adjusting the upper limit of integration in the solution, which in the case of simple classes of potentials can be approximated by Gaussian functions, yielding estimations for the first passage time from Kramer's Result, with other studies of similarly posed diffusion processes obtained from solutions of the Smoluchowski equation in [18].

For CRISPR-Cas binding, an IVP corresponding to the first passage problem can be formulated by enforcing initial conditions which stipulate that the position of the particle undergoing diffusion is centered at the origin when target inspection is initiated, and that the particle subject to a unit initial velocity. As the protein inspection continues for remaining base pairs in the sequence, passage times can be computed by making use of numerical relations from the closed form of solutions to the ODE, primarily based in obtaining three variational formulas involving participation from several terms. To satisfactorily generate realistic binding energy landscapes that proteins encounter throughout inspection, we reflect upon separate approaches to determine mean exit times, from one approach which is capable of obtaining the exit time through numerical approximations of solutions to a stochastically driven oscillator, while another method raises an inverse problem, similar to that studied in [2], in which potential energy landscapes can be uniquely reconstructed from distributions of exit times. The inverse problem formulation presented here is focused towards the construction of the binding landscape potential from collections of exit times up to an absorbing membrane of variable length. Additional comparisons between probability measures, in which probability measures with another Hamiltonian, against the probability measure $p_i = \exp(\nabla U_i)/Z$ with potential U_i, will also be established.

2 Methodology

2.1 Description

The inverse problem poised towards reconstruction of the binding potential from exit time distributions relies on the following framework. To study the rates at which particles diffuse across base pairs in the binding process throughout crRNA inspection, solutions τ to the dimensionless, second order ODE of the form,

$$-\mathcal{A}\frac{\mathrm{d}^2\tau}{\mathrm{d}x^2} + \mathcal{U}'(x)\frac{\mathrm{d}\tau}{\mathrm{d}x} = 1 \ ,$$

are determined where the normalization introduced to obtain the dimensionless equation is proportional to the product of the Boltzmann constant and ambient temperature of bond melting, $\mathcal{A} \equiv \frac{k_b T}{\nu}$, $\mathcal{U}'(x) \equiv \frac{U'(x)}{\nu}$, and $U'(x)$ is the potential landscape before normalization by the driving force ν. To specify classes of binding potentials for which solutions are to be determined, we impose the criterion that candidate potentials from the admissible landscape space possess one degree of freedom for each base pair at which binding occurs. In numerical applications of potential landscape reconstruction for mean exit time distributions, enforcing straightforward conditions on the mean and variance of the exit distributions themselves can respectively be achieved through specifying the first sample that is drawn from the time distribution, in addition to the maximum and minimum sample that can be drawn afterwards to specify the variance of the distribution which is also related to the fatness of its tails. For such classes of potentials, solutions to the ODE take the form,

$$\tau \approx \int_0^x \int_0^x \prod_{i=2}^{20} \exp\Big(-\frac{u^i}{i}\Big) \prod_{i=2}^{20} \exp\Big(\frac{u^i}{i}\Big) \mathrm{d}u \ \mathrm{d}u + \int_0^x \int_0^v \{\prod_{i=2}^{20} \exp\Big(-\frac{u^i}{i}\Big) \mathrm{d}u\} \prod_{i=2}^{20} \exp\Big(\frac{v^i}{i}\Big) \mathrm{d}v + \int_0^x \prod_{i=2}^{20} \exp\Big(-\frac{u^i}{i}\Big) \mathrm{d}u$$

indeed reflective of contributions from all nonzero polynomial terms specified in the candidate potential. Under simple rearrangements, variatonal formulae for the difference of exit times $\Delta_\tau = \tau_{x_2} - \tau_{x_1}$ are obtained through analyzing, on a case by case basis, the space of possible combinations of absorbing membrane lengths at arbitrary positions $x_1 < x_2$ of the target sequence. Once the formulae have been established, further discussion will be devoted towards the construction of admissible distributions from which potential landscapes can be reconstructed. Before derivations of the variational formulae, we characterize the dependence between solutions and the class of potentials that we have identified, with solutions for varying arrangements of absorbing membranes.

Within the potential space, the goal of numerically obtaining mean exit times is to generate small perturbations to the energy landscape corresponding to small perturbations in the exit time. Regardless of experimental constraints in experiments that have been carried through measurements of the rate at which reactants are consumed in Fn Cas12a binding [29], the inverse problem of recovering the landscape can be numerically realized readily in several ways. First, one method involves producing approximations of the mean exit time from a given potential through Gaussian approximations on the integral terms from τ, while another second closely related numerical approach involves numerically approximating the exit time after rearranging terms from τ through possible values on the innermost variable v of integration from the second term in τ. Third, another approach entails that we rearrange terms from τ depending on the position of the exit time of interest v, in which it is possible to make use of linearity of the integral to obtain variational formulae below. Before proceeding to the computations, solutions $\mathcal{S}$ for the variational formulae include,

$$\mathcal{S}_{v \neq x}(v,x) \equiv -\int_0^x \Big(\int_0^x \{\prod_{i=2}^{20} \exp\Big(\frac{u^i}{i}\Big) - 1\}\Big) \mathrm{d}u \prod_{i=2}^{20} \exp\Big(\frac{u^i}{i}\Big) \mathrm{d}u + \int_0^x \int_0^v \{\prod_{i=2}^{20} \exp\Big(-\frac{u^i}{i}\Big) \mathrm{d}u\} \prod_{i=2}^{20} \exp\Big(\frac{v^i}{i}\Big) \mathrm{d}v \ ,$$

while for $v \equiv x$,

$$\mathcal{S}_{v=x}(v,x) \equiv -\int_0^x \prod_{i=2}^{20} \exp\biggl(\frac{u^i}{i}\biggr)\biggl(\int_0^x \{2\prod_{i=2}^{20} \exp\biggl(\frac{u^i}{i}\biggr) - 1\}\mathrm{d}u\biggr)\mathrm{d}u \ .$$

From $\mathcal{S}_{v=x}$ and $\mathcal{S}_{v\neq x}$, terms from the numerical approximation of exit times are implemented in the following cases. In the formulae, the passage time up to the first position x_1, interactions over the passage time to x_2, and the intermediate interactions between the first and second passage times numerically contribute, from which variations in one exit time parameter generate classes of potential landscapes. For the inverse problem, at onset we require specification of one basis element in the potential space, and its corresponding exit time, in addition to the deviation from the exit time through specification of the second exit time. The potential corresponding to the second exit time can be recovered through numerical approximation of the variational formulae.

2.2 (Var1) equality for $\Delta_\tau, x_1, x_2 \neq 1$

In the first realization of the variational formula, solutions can be fashioned towards recovering potential landscapes for the passage times through a decrement of τ_{x_1} with τ_{x_2}, in which Δ_τ takes the form,

$$\Delta_\tau \equiv \tau_{x_2} - \tau_{x_1} = \mathcal{S}_{v\neq x}(v,x_2) - \mathcal{S}_{v\neq x}(v,x_1) \ ,$$

after which substitution for solutions $\mathcal{S}$ gives,

$$\begin{gathered} -\int_0^{x_2} \prod_{i=2}^{20} \exp\biggl(\frac{u^i}{i}\biggr)\biggl(\int_0^{x_2} \{\prod_{i=2}^{20} \exp\biggl(\frac{u^i}{i}\biggr) - 1\}\mathrm{d}u\biggr)\mathrm{d}u + \int_0^{x_2}\int_0^{v} \{\prod_{i=2}^{20} \exp\biggl(-\frac{u^i}{i}\biggr)\mathrm{d}u\} \prod_{i=2}^{20} \exp\biggl(\frac{v^i}{i}\biggr)\mathrm{d}v \ + \\ \int_0^{x_1} \prod_{i=2}^{20} \exp\biggl(\frac{u^i}{i}\biggr)\biggl(\int_0^{x_1} \{\prod_{i=2}^{20} \exp\biggl(\frac{u^i}{i}\biggr) - 1\}\mathrm{d}u\biggr)\mathrm{d}u + \int_0^{x_1}\int_0^{v} \{\prod_{i=2}^{20} \exp\biggl(-\frac{u^i}{i}\biggr)\mathrm{d}u\} \prod_{i=2}^{20} \exp\biggl(\frac{v^i}{i}\biggr)\mathrm{d}v \ , \end{gathered}$$

from which we consider subcases of possible v.

2.2.1 $v \equiv x_2$

Rearrangements yield,

$$\begin{gathered} \boxed{\int_0^{x_2}\int_0^{x_2} \{\prod_{i=2}^{20} \exp\biggl(-\frac{u^i}{i}\biggr) - \exp\biggl(\frac{u^i}{i}\biggr) + 1\}\mathrm{d}u \prod_{i=2}^{20} \exp\biggl(\frac{v^i}{i}\biggr)\mathrm{d}v \ + \cdots} \\ \boxed{\int_0^{x_1}\int_0^{x_1} \{\prod_{i=2}^{20} \exp\biggl(\frac{u^i}{i}\biggr) - 1\}\mathrm{d}u \prod_{i=2}^{20} \exp\biggl(\frac{u^i}{i}\biggr)\mathrm{d}u + \int_0^{x_1}\int_0^{x_2} \{\prod_{i=2}^{20} \exp\biggl(-\frac{u^i}{i}\biggr)\mathrm{d}u\} \prod_{i=2}^{20} \exp\biggl(\frac{v^i}{i}\biggr)\mathrm{d}v} \ . \end{gathered} \qquad (\textbf{Var1A})$$

2.2.2 $v \equiv x_1$

Rearrangements yield,

$$\boxed{\int_0^{x_1}\int_0^{x_1}\{\prod_{i=2}^{20}\exp\Big(-\frac{u^i}{i}\Big)-\exp\Big(\frac{u^i}{i}\Big)-1\}\mathrm{d}u\ \prod_{i=2}^{20}\exp\Big(\frac{v^i}{i}\Big)\ \mathrm{d}v\ +\cdots}$$
$$\boxed{\int_0^{x_2}\Big(\int_{x_2}^{x_1}\{\prod_{i=2}^{20}\exp\Big(\frac{u^i}{i}\Big)-1\}\mathrm{d}u\Big)\prod_{i=2}^{20}\exp\Big(\frac{v^i}{i}\Big)\mathrm{d}v}\ . \qquad \textbf{(Var1B)}$$

For exit times in which the length of the absorbing boundaries at exit times τ_{x_1} and τ_{x_2} are positions along the genome, neither of which are positioned at the unit boundary, the first variational formula **(Var1)** permits for solutions to the inverse exit time problem through specification of the exit time free parameters.

2.3 (Var2) equality for $\Delta_\tau, x_2 > 1, x_1 \equiv 1$

In the second realization of the variational formula, solutions can be fashioned towards recovering potential landscapes for the passage times through a similarly defined decrement of τ_{x_2}, instead taking into account another possible arrangement of the absorbing boundary length, in which $\Delta_\tau \equiv \tau_{x_2} - \tau_{x_1} = \mathcal{S}_{v\neq x}(v, x_2) - \mathcal{S}_{v=x}(v, x_1)$ implies,

$$-\int_0^{x_2}\Big(\int_0^{x_2}\{\prod_{i=2}^{20}\exp\Big(\frac{u^i}{i}\Big)-1\}\mathrm{d}u\Big)\prod_{i=2}^{20}\exp\Big(\frac{u^i}{i}\Big)\mathrm{d}u+\int_0^{x_2}\int_0^{v}\{\prod_{i=2}^{20}\exp\Big(-\frac{u^i}{i}\Big)\mathrm{d}u\}\prod_{i=2}^{20}\exp\Big(\frac{v^i}{i}\Big)\mathrm{d}v\ +$$
$$\int_0^{x_1}\prod_{i=2}^{20}\exp\Big(\frac{u^i}{i}\Big)\Big(\int_0^{x_1}\{2\prod_{i=2}^{20}\exp\Big(\frac{u^i}{i}\Big)-1\}\mathrm{d}u\Big)\mathrm{d}u\ , \qquad \textbf{(Lin1)}$$

now requiring that we make use of linearity of one order of integration, **(Lin1)**, while holding the other order constant. By hypothesis, $x_1 < x_2$, allowing for rearrangements in the second term, **(Lin2)**,

$$\int_0^{x_1}\prod_{i=2}^{20}\exp\Big(\frac{u^i}{i}\Big)\Big(\int_0^{x_2}\{\prod_{i=2}^{20}\exp\Big(\frac{u^i}{i}\Big)-1\}\mathrm{d}u\Big)\mathrm{d}u+\int_{x_1}^{x_2}\prod_{i=2}^{20}\exp\Big(\frac{u^i}{i}\Big)\Big(\int_0^{x_2}\{\prod_{i=2}^{20}\exp\Big(\frac{u^i}{i}\Big)-1\}\mathrm{d}u\Big)\mathrm{d}u\ , \qquad \textbf{(Lin2)}$$

in addition to the first. Without loss of generality, for the second term in Δ_τ suppose that $v < x_1$; we envoke linearity of the integral for the innermost variable,

$$\int_0^{x_2}\Big(\int_0^{x_1}\prod_{i=2}^{20}\exp\Big(-\frac{u^i}{i}\Big)\mathrm{d}u+\int_{x_1}^{v}\prod_{i=2}^{20}\exp\Big(-\frac{u^i}{i}\Big)\mathrm{d}u\Big)\prod_{i=2}^{20}\exp\Big(\frac{v^i}{i}\Big)\mathrm{d}v\ ,$$

while for $v \equiv x_1$ and $v > x_1$, the second term corresponding to each possibility in Δ_τ is of the form,

$$\int_0^{x_2}\int_0^{x_1}\{\prod_{i=2}^{20}\exp\Big(-\frac{u^i}{i}\Big)\mathrm{d}u\}\prod_{i=2}^{20}\exp\Big(\frac{v^i}{i}\Big)\mathrm{d}v\ ,$$

and

$$\int_0^{x_2}\Big(\int_0^{v}\prod_{i=2}^{20}\exp\Big(-\frac{u^i}{i}\Big)\mathrm{d}u+\int_v^{x_2}\prod_{i=2}^{20}\exp\Big(-\frac{u^i}{i}\Big)\mathrm{d}u\Big)\prod_{i=2}^{20}\exp\Big(\frac{v^i}{i}\Big)\mathrm{d}v\ ,$$

respectively. We collect like terms which are taken over the same region of integration. This requires that we identify rearrangements of terms in each subcase below.

2.3.1 $v \equiv x_2$

Under this assumption the variational formula takes the form,

$$\boxed{\int_0^{x_2}\int_0^{x_2}\{\prod_{i=2}^{20}\exp\Big(-\frac{u^i}{i}\Big)-\exp\Big(\frac{u^i}{i}\Big)+1\}\mathrm{d}u\prod_{i=2}^{20}\exp\Big(\frac{v^i}{i}\Big)\mathrm{d}v+\int_0^{x_1}\prod_{i=2}^{20}\exp\Big(\frac{u^i}{i}\Big)\Big(\int_0^{x_1}\{2\prod_{i=2}^{20}\exp\Big(\frac{u^i}{i}\Big)-1\}\mathrm{d}u\Big)\mathrm{d}u} \tag{\textbf{Var2A}}$$

2.3.2 $v \equiv x_1$

Under this assumption we invoke **(Lin2)**, from which the variational formula takes the form,

$$-\int_0^{x_2}\prod_{i=2}^{20}\exp\Big(\frac{u^i}{i}\Big)\Big(\int_0^{x_2}\{\prod_{i=2}^{20}\exp\Big(\frac{u^i}{i}\Big)-1\}\mathrm{d}u\Big)\mathrm{d}u+\int_0^{x_2}\int_0^{x_1}\{\prod_{i=2}^{20}\exp\Big(-\frac{u^i}{i}\Big)\mathrm{d}u\}\prod_{i=2}^{20}\exp\Big(\frac{v^i}{i}\Big)\mathrm{d}v\ +$$
$$\int_0^{x_1}\prod_{i=2}^{20}\exp\Big(\frac{u^i}{i}\Big)\Big(\int_0^{x_1}\{2\prod_{i=2}^{20}\exp\Big(\frac{u^i}{i}\Big)-1\}\mathrm{d}u\Big)\mathrm{d}u\ ,$$

from which an application of **(Lin2)** to the second term yields,

$$\int_0^{x_1}\int_0^{x_1}\{\prod_{i=2}^{20}\exp\Big(-\frac{u^i}{i}\Big)\mathrm{d}u\}\prod_{i=2}^{20}\exp\Big(\frac{v^i}{i}\Big)\mathrm{d}v+\int_{x_1}^{x_2}\int_0^{x_1}\{\prod_{i=2}^{20}\exp\Big(-\frac{u^i}{i}\Big)\mathrm{d}u\}\prod_{i=2}^{20}\exp\Big(\frac{v^i}{i}\Big)\mathrm{d}v\ ,$$

implying that

$$\boxed{-\int_0^{x_2}\int_0^{x_2}\prod_{i=2}^{20}\exp\Big(\frac{u^i}{i}\Big)\Big(\int_0^{x_2}\{\prod_{i=2}^{20}\exp\Big(\frac{u^i}{i}\Big)-1\}\mathrm{d}u\Big)\mathrm{d}u+\int_0^{x_1}\Big(\int_0^{x_1}\{\prod_{i=2}^{20}\exp\Big(\frac{u^i}{i}\Big)-1\}\mathrm{d}u\Big)\mathrm{d}u\ +\cdots}$$
$$\boxed{\int_{x_1}^{x_2}\int_0^{x_1}\{\prod_{i=2}^{20}\exp\Big(-\frac{u^i}{i}\Big)\mathrm{d}u\}\prod_{i=2}^{20}\exp\Big(\frac{v^i}{i}\Big)\mathrm{d}v}\ . \tag{\textbf{Var2B}}$$

2.3.3 $v > x_1$

Under this assumption, we invoke **(Lin2)** in the second term, after which invoking **(Lin1)** in the second term, and subsequent rearrangements with the third term yields

$$\overset{\textbf{(Lin1)}}{\Rightarrow} \int_0^{x_2} \Big(\int_0^{x_1} \prod_{i=2}^{20} \exp\Big(- \frac{u^i}{i} \Big) \mathrm{d}u + \int_{x_1}^{x_2} \prod_{i=2}^{20} \exp\Big(- \frac{u^i}{i} \Big) \mathrm{d}u \Big) \prod_{i=2}^{20} \exp\Big(\frac{v^i}{i} \Big) \mathrm{d}v \ ,$$

$$\overset{\textbf{(Lin2)}}{\Rightarrow} \int_0^{x_1} \Big(\int_0^{x_1} \prod_{i=2}^{20} \exp\Big(- \frac{u^i}{i} \Big) \mathrm{d}u + \int_{x_1}^{x_2} \prod_{i=2}^{20} \exp\Big(- \frac{u^i}{i} \Big) \mathrm{d}u \Big) \prod_{i=2}^{20} \exp\Big(\frac{v^i}{i} \Big) \mathrm{d}v \ +$$

$$\int_{x_1}^{x_2} \Big(\int_0^{x_1} \prod_{i=2}^{20} \exp\Big(- \frac{u^i}{i} \Big) \mathrm{d}u + \int_{x_1}^{x_2} \prod_{i=2}^{20} \exp\Big(- \frac{u^i}{i} \Big) \mathrm{d}u \Big) \prod_{i=2}^{20} \exp\Big(\frac{v^i}{i} \Big) \mathrm{d}v \ .$$

We obtain intermediate terms of the relation,

$$- \int_0^{x_2} \int_0^{x_2} \{ \prod_{i=2}^{20} \exp\Big(\frac{u^i}{i} \Big) - 1 \} \mathrm{d}u \mathrm{d}u + \int_0^{x_1} \int_0^{x_1} \{ 3 \prod_{i=2}^{20} \exp\Big(\frac{u^i}{i} \Big) - 1 \} \mathrm{d}u \ \mathrm{d}v \ +$$

$$\int_{x_1}^{x_2} \int_{x_1}^{x_2} \{ \prod_{i=2}^{20} \exp\Big(- \frac{u^i}{i} \Big) \mathrm{d}u \} \prod_{i=2}^{20} \exp\Big(\frac{v^i}{i} \Big) \mathrm{d}v \ ,$$

for integral terms that are uniform in the boundaries of integration on each order, while the remaining terms are of the form,

$$\int_0^{x_1} \int_{x_1}^{x_2} \{ \prod_{i=2}^{20} \exp\Big(- \frac{u^i}{i} \Big) \mathrm{d}u \} \prod_{i=2}^{20} \exp\Big(\frac{v^i}{i} \Big) \mathrm{d}v + \int_{x_1}^{x_2} \int_0^{x_1} \{ \prod_{i=2}^{20} \exp\Big(- \frac{u^i}{i} \Big) \mathrm{d}u \} \prod_{i=2}^{20} \exp\Big(\frac{v^i}{i} \Big) \mathrm{d}v \ ,$$

are composed of nonuniform bounds on each order order of integration. Applying Fubini to either integral above permits for consolidation of the mixed ordered terms. Putting together all rearrangements implies,

$$\boxed{- \int_0^{x_2} \int_0^{x_2} \{ \prod_{i=2}^{20} \exp\Big(\frac{u^i}{i} \Big) - 1 \} \mathrm{d}u \mathrm{d}u + \int_0^{x_1} \int_0^{x_1} \{ 3 \prod_{i=2}^{20} \exp\Big(\frac{u^i}{i} \Big) - 1 \} \mathrm{d}u \ \mathrm{d}v \ + \cdots}$$

$$\boxed{\int_{x_1}^{x_2} \int_{x_1}^{x_2} \{ \prod_{i=2}^{20} \exp\Big(- \frac{u^i}{i} \Big) \mathrm{d}u \} \prod_{i=2}^{20} \exp\Big(\frac{v^i}{i} \Big) \mathrm{d}v + 2 \int_0^{x_1} \int_{x_1}^{x_2} \{ \prod_{i=2}^{20} \exp\Big(- \frac{u^i}{i} \Big) \mathrm{d}u \} \prod_{i=2}^{20} \exp\Big(\frac{v^i}{i} \Big) \mathrm{d}v} \quad \textbf{(Var2C)}$$

2.4 (Var3) equality for $\Delta_\tau, x_2 \equiv 1, x_1 < 1$

In the third realization of the variational formula, solutions can be fashioned towards recovering potential landscapes for the passage times through the increments of τ_x, through the previously used formulation

$$\Delta_\tau \equiv \tau_{x_2} - \tau_{x_1} = \mathcal{S}_{v=x}(v, x_2) - \mathcal{S}_{v \neq x}(v, x_1)$$

$$= - \int_0^{x_2} \Big(\int_0^{x_2} \{ 2 \prod_{i=2}^{20} \exp\Big(\frac{u^i}{i} \Big) - 1 \} \mathrm{d}u \Big) \prod_{i=2}^{20} \exp\Big(\frac{u^i}{i} \Big) \mathrm{d}u \ -$$

$$\int_0^{x_1} \Big(\int_0^{x_1} \{ \prod_{i=2}^{20} \exp\Big(\frac{u^i}{i} \Big) - 1 \} \Big) \mathrm{d}u \prod_{i=2}^{20} \exp\Big(\frac{u^i}{i} \Big) \mathrm{d}u + \int_0^{x_1} \int_0^{v} \{ \prod_{i=2}^{20} \exp\Big(- \frac{u^i}{i} \Big) \mathrm{d}u \} \prod_{i=2}^{20} \exp\Big(\frac{v^i}{i} \Big) \mathrm{d}v \ ,$$

2.4.1 $v \equiv x_1$

Under this assumption, the relation takes the form,

$$\boxed{\int_0^{x_1}\int_0^{x_1}\{\prod_{i=2}^{20}\{\exp\Big(\frac{u^i}{i}\Big)+\exp\Big(-\frac{u^i}{i}\Big)\}-1\}\prod_{i=2}^{20}\exp\Big(\frac{v^i}{i}\Big)\mathrm{d}u\ \mathrm{d}v-\int_0^{x_2}\int_0^{x_2}\{2\prod_{i=2}^{20}\exp\Big(\frac{u^i}{i}\Big)-1\}\mathrm{d}u\prod_{i=2}^{20}\exp\Big(\frac{u^i}{i}\Big)\mathrm{d}u}\ . \tag{\textbf{Var3A}}$$

2.4.2 $v \equiv x_2$

Under this assumption, the relation takes the form,

$$\boxed{\int_0^{x_1}\int_0^{x_1}\prod_{i=2}^{20}\{\exp\Big(-\frac{u^i}{i}\Big)-\exp\Big(\frac{u^i}{i}\Big)+1\}\prod_{i=2}^{20}\exp\Big(\frac{u^i}{i}\Big)\mathrm{d}u\mathrm{d}u-\int_0^{x_2}\Big(\int_0^{x_2}\{2\prod_{i=2}^{20}\exp\Big(\frac{u^i}{i}\Big)-1\}\mathrm{d}u\Big)\cdots}$$
$$\boxed{\prod_{i=2}^{20}\exp\Big(\frac{u^i}{i}\Big)\mathrm{d}u+\int_0^{x_1}\int_{x_1}^{x_2}\{\prod_{i=2}^{20}\exp\Big(-\frac{u^i}{i}\Big)\mathrm{d}u\}\prod_{i=2}^{20}\exp\Big(\frac{u^i}{i}\Big)\mathrm{d}u}\ , \tag{\textbf{Var3B}}$$

resulting from an application of **(Lin2)** to the third term which yields

$$\int_0^{x_1}\Big(\int_0^{x_1}\prod_{i=2}^{20}\exp\Big(-\frac{u^i}{i}\Big)\mathrm{d}u+\int_{x_1}^{x_2}\prod_{i=2}^{20}\exp\Big(-\frac{u^i}{i}\Big)\mathrm{d}u\Big)\prod_{i=2}^{20}\exp\Big(\frac{u^i}{i}\Big)\mathrm{d}u\ .$$

We exclude the case pertaining to $v \geq x_1$.

2.5 Numerical test cases with the variational formulae

We now briefly review the composition of numerical simulations from each variational formula.

- **(Var1A)**: The first variational formula encapsulates straightforward numerical behaviors of the fluctuations in the potential landscape from corresponding fluctuations in the exit time. Only one term in the relation is dependent on both positions of exit. For vanishing x_1 (x_2), only one term from the relation is recovered.

- **(Var1B)**: For exit times at x_1 & x_2, the second variational formula encapsulates more intermediate contributions from the exit times over (x_1, x_2), from which vanishing x_1 or x_2 would cause the last term in the relation to vanish altogether. On the other hand, for vanishing x_1, it is possible to only recover the second term in the relation.

- **(Var2A)**: For exit times at x_1 & x_2, the potential corresponding to the x_1 exit time can be recovered through numerical approximations of the **(Var2A)** equality. With apriori knowledge of the potential corresponding to the exit time at x_1, the potential corresponding to the exit time at x_2 is obtained through manipulation of integral terms which are separately dependent on the positions of passage to x_1 and to x_2. In the limit as $x_2 \rightarrow 0^+$ ($x_1 \rightarrow 0^+$), the second (first) terms vanish.

- **(Var2B)**: For exit times at x_1 & x_2, it is possible to recover the potential associated with the exit time at x_1, with apriori knowledge of all other terms in the relation. In contrast to other variational relations, we observe that from **(Var2A)** that the relation is composed of intermediate interactions between x_1 and x_2, in turn influencing the variation in the landscape potential associated with x_1. As $x_1 \rightarrow x_2^-$, the intermediate term vanishes and **(Var2B)** is closely representative of **(Var2A)**, with the exception of different contributions from the numerical approximation of the exit time up to x_2.

- **(Var2C)**: For exit times at x_1 & x_2, the variational formula **(Var2C)** involves simultaneous contributions from the exit times up to x_1 & x_2, in addition to contributions from the exit times in intermediate positions. As $x_1 \to 0^+$, half of the contributions from the equality vanish, while as $x_2 \to 0^+$, the first term which has a strong dependence on x_2 vanishes.

- **(Var3A)**: For exit times at x_1 & x_2, **(Var3A)** involves contributions from the exit times up to those positions, with no contributions from intermediate positions. The contributions for the exit time up to x_i, as $x_i \to 0^+$, vanish for $i \in \{1, 2\}$.

- **(Var3B)**: For exit times at x_1 & x_2, the final variational relation is comprised of a mixture of contributions similar to that of **(Var2C)**, in which contributions persist across intermediate positions between x_1 and x_2. In the vanishing x_1 limit from above, two of the three terms in the relation disappear, while in the vanishing x_2 limit from above, two terms in the relation survive.

2.6 Generating exit time distributions

We readily generate exit time distributions from which potential landscapes will be recovered. In a suitable free parameter generalization for different Cas proteins, the exit time formalism must be capable of determining the perturbation in the landscape given an initial potential choice with degrees of freedom at each base pair. The three variational formulae that have been presented are capable of exectuting the potential reconstruction, in addition to other quantities pertaining to the drift terms associated with well characterized properties of protein kinetics [2].

This final subsection is devoted towards not only a description of how the visit distributions for the exit times can be generated, but also how apriori choicess of the exit times at both positions, in addition to the potential corresponding to the first exit time, can be enforced. The implementation of our approach is capable of recovering potentials associated with the free exit time, permitting for numerical simulations of energy landscapes for different Cas proteins. To numerically implement the variational formulae for landscape reconstruction, we make use of individual instances of the relation described extensively above. Primarily, we are interested in determining the order of difference in the exit times across subsequences of the target sequence that are at least one base pair long, despite being able to compute

For a suitable parameter free generalization of the thermodynamic model of dCas binding,

3 References

[1] Allison, D. & Wang, G. R-loops: formation, function, and relevance to cell stress. *Cell Stress* **3**(2) (2019).

[2] Bal, G. & Chou, T. On the reconstruction of diffusions from first-exit time distributions. *Inverse Problems* **20**(4) (2004).

[3] Bintu, L., Buchler, N. E., Garcia, H. G., Gerland, U., Hwa, T., Kondev, J., Kuhlman, T., & Phillips, R. Transcriptional regulation by the numbers: applications. *Current opinion in genetics & development*, **15**(2), 125–135 (2015).

[4] Borys, P. & Grzywna, Z. The Fokker-Planck Equation for Chaotic Maps. *Acta Physica Polonica B* **37**(2) (2006).

[5] Brewester, R., Weinert, F., Garcia, H., Song, D., Rydenfelt, M. & Phillips, R. The Transcription Factor Titration Effect Dictates Level of Gene Expression. *Cell* **6**, 1312-1323 (2014).

[6] Chen, J., Dagdas, Y., Kleinstiver, B., Welch, M., Sousa, A., Harrington, L., Sternberg, S., Joung, J., Yildiz, A. & Doudna, J. Enhanced proofreading governs CRISPR-Cas9 targeting accuracy. *Nature*, **550**, 407-410 (2017).

[7] Chen, B., Gilbert, L., et al. Dynamic Imaging of Genomic Loci in Living Human Cells by an Optimized CRISPR/Cas System. *Cell*, **155**(7) 1479-1491 (2013).

[8] Chou, T. & D'Orsogna, M. First Passage Problems in Biology. *Arxiv* (2014).

[9] D'Orsogna, M., Lakatos, G. & Chou, T. Stochastic self-assembly of incommensurate clusters. *J Chem Phys* **136**(8) (2012).

[10] Eitzinger, S., Asif, A., Watters, K. E., Iavarone, A. T., Knott, G. J., Doudna, J. A., & Minhas, F. Machine learning predicts new anti-CRISPR proteins. *Nucleic acids research*, **48**(9), 4698–4708 (2020).

[11] Esvelt, K. M., Mali, P., Braff, J. L., Moosburner, M., Yaung, S. J., & Church, G. M. Orthogonal Cas9 proteins for RNA-guided gene regulation and editing. Nature methods, **10**(11), 1116–1121 (2013).

[12] Eslami-Mossallam, B., Klein, M., Smagt, C., Sanden, K., Jones Jr, S., Hawkins, J., A kinetic model improves off-target predictions and reveals the physical basis of SpCas9 fidelity. *Preprint.*

[13] Garneau, J. et al. the CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468** 67-71 (2010).

[14] Horlbeck, M. A., Witkowsky, L. B., Guglielmi, B., Replogle, J. M., Gilbert, L. A., Villalta, J. E., Torigoe, S. E., Tjian, R., & Weissman, J. S. Nucleosomes impede Cas9 access to DNA in vivo and in vitro. eLife, **5**, e12677 (2016).

[15] Hultquist, J. F., Hiatt, J., Schumann, K., McGregor, M. J., Roth, T. L., Haas, P., Doudna, J. A., Marson, A., & Krogan, N. J. CRISPR-Cas9 genome engineering of primary CD4+ T cells for the interrogation of HIV-host factor interactions. *Nature protocols*, **14**(1), 1–27 (2019).

[16] Jackson, S., Suma, A. & Micheletti, C. How to fold intricately: using theory and experiments to unravel the properties of knotted proteins. *Current Opinion in Structural Biology* **42**, 6-14 (2017).

[17] Jeon, Y. et al. Direct observation of DNA target searching and cleavage by CRISPR-Cas12a. *Nature Communications*, **9**, 2777 (2018).

[18] Keener, J. Mathematical Biology Course Lecture Notes.

[19] Kim, B., Kim, H. & Lee, S. Regulation of Microbial Metabolic Rates Using CRISPR Interference with Expanded PAM Sequences. *Frontiers in Microbiology* **11**(282), (2020).

[20] Kinney, J., Tkacik, G. & Callan, G. Precise physical models of protein-DNA interaction from high-throughput data. *PNAS*, **104**(2), 501-506 (2007).

[21] Krapivsky, P., Redner, S. & Ben-Naim, E. A Kinetic View of Statistical Physics. *Cambridge University Press* 978-0-521-85103-9 (2010).

[22] Mallamace, F. et al. Energy landscape in protein folding and unfolding. *PNAS*, **113**(12), 3159-3163 (2016).

[23] Mekler, V., Minakhin, L., Semenova, E., Kuznedelov, K., & Severinov, K. Kinetics of the CRISPR-Cas9 effector complex assembly and the role of 3'-terminal segment of guide RNA. *Nucleic acids research*, **44**(6), 2837–2845 (2016).

[24] Mirny, L. & Shakhnovich, E. Protein Folding Theory: From Lattice to All-Atom Models. *Annual Reviews Biophysics*, **30** 261-96 (2001).

[25] Morra, G., Genoni, A. & Colombo, G. Protein Dynamics and Drug Design: The Role of Molecular Simulations. *Protein-Protein Complexes*, 340-385 (2010).

[26] Peled, R. Topics in Statistical Physics and Probability Theory. *Lecture Notes.*

[27] Ran, F., Hsu, P., Wright, J. Agarwala, V., Scott, D. & Zhang, F. Genome engineering using the CRISPR-Cas9 system. *Nature Protocols* **8** 2281-2308 (2013).

[28] Satori, P. & Leibler, S. Lessons from equilibrium statistical physics regarding the assembly of protein complexes. *PNAS* **117**(1) 114-120 (2020).

[29] Spetcht, D., Xu, Y. & Lambert, G. Massively parallel CRISPRi assays reveal concealed thermodynamic determinants of dCas12a binding. *PNAS* **117**(21) 11274-11282 (2020).

[30] Slutsky, M. & Mirny, L. Kinetics of Protein-DNA Interaction: Facilitated Target Location in Sequence-Dependent Potential. *Biophysical Journal*, **87** 4021-4035 (2004).

[31] Stella, S. et al. Conformational Activation Promotes CRISPR-Cas12a Catalysis and Resetting of the Endonuclease Activity. *Cell* **175** 1856-1871 (2018).

[32] Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C., & Doudna, J. A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507**(7490), 62–67 (2014).

[33] Tkacik, G., Callan, C. & Bialek, W. Information flow and optimization in transcriptional regulation. *PNAS*, **105**(34), 12265-12270 (2008).

[34] Wolf, S., Lickert, B., Bray, S. & Stock, G. Multisecond ligand dissociation dynamics from atomistic simulations. *Nature Communications*, **11**, 2918 (2020).

[35] Xu, X., Duan, D. & Chen, S. CRISPR-Cas9 cleavage efficiency correlates strongly with target-sgRNA folding stability: from physical mechanism to off-target assessment. *Sci Rep* **7**, 143 (2017).