Inferring Parameters for an Elementary Step Model of DNA Structure Kinetics with Locally Context-Dependent Arrhenius Rates

Sedigheh Zolaktaf[1], Frits Dannenberg[2], Xander Rudelis[2], Anne Condon[1], Joseph M Schaeffer[3], Mark Schmidt[1], Chris Thachuk[2], and Erik Winfree[2(✉)]

[1] University of British Columbia, Vancouver, BC, Canada
[2] California Institute of Technology, Pasadena, CA, USA
[3] Autodesk Research, San Francisco, CA, USA
winfree@caltech.edu

Abstract. Models of nucleic acid thermal stability are calibrated to a wide range of experimental observations, and typically predict equilibrium probabilities of nucleic acid secondary structures with reasonable accuracy. By comparison, a similar calibration and evaluation of nucleic acid kinetic models to a broad range of measurements has not been attempted so far. We introduce an Arrhenius model of interacting nucleic acid kinetics that relates the activation energy of a state transition with the immediate local environment of the affected base pair. Our model can be used in stochastic simulations to estimate kinetic properties and is consistent with existing thermodynamic models. We infer parameters for our model using an ensemble Markov chain Monte Carlo (MCMC) approach on a training dataset with 320 kinetic measurements of hairpin closing and opening, helix association and dissociation, bubble closing and toehold-mediated strand exchange. Our new model surpasses the performance of the previously established Metropolis model both on the training set and on a testing set of size 56 composed of toehold-mediated 3-way strand displacement with mismatches and hairpin opening and closing rates: reaction rates are predicted to within a factor of three for 93.4% and 78.5% of reactions for the training and testing sets, respectively.

1 Introduction

Although nucleic acids are commonly synthesized and applied in various settings, it remains difficult to predict the kinetics of their interaction and conformational change. Accurate models of nucleic acid kinetics are desirable for biological and biotechnological applications, such as understanding the various roles of RNA within the cell and the design of sensitive molecular probes. Within the field of molecular programming, hairpin motifs and toehold-mediated strand displacement are commonly used to implement autonomous devices such as DNA walkers and logic gates. Models of nucleic acid thermal stability have been extensively calibrated to experimental data [4, 16] and enable secondary

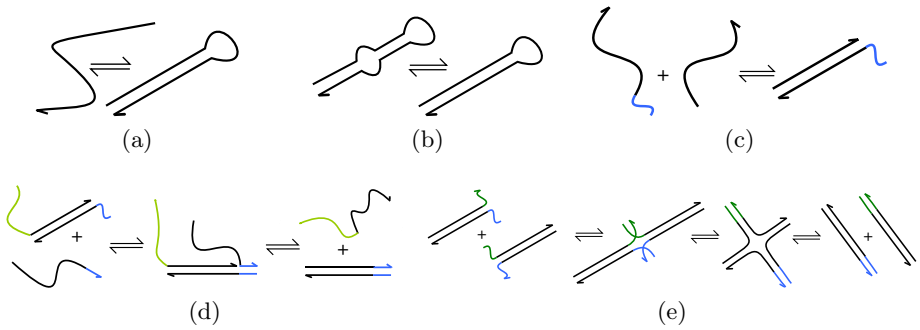

Fig. 1: Five types of reactions that we simulate and for which reaction rate constants have been measured. (a) Hairpin closing and opening. (b) Bubble closing. (c) Helix association and dissociation. (d) Toehold-mediated 3-way strand displacement. (e) Toehold-mediated 4-way strand exchange.

structure software such as RNAsoft, ViennaRNA, RNAstructure, NUPACK, and mfold [3, 12, 26, 27, 29] to efficiently predict the equilibrium probabilities of nucleic acid secondary structures. In comparison, a similar extensive calibration and evaluation of nucleic acid kinetic models has not been attempted so far, despite the development of kinetic models and simulation software such as Multistrand and Kinefold [7, 9, 21, 22, 25]. Of particular interest is a study by Srinivas et al., which demonstrates that the Metropolis model of Multistrand is incompatible with observations of toehold-mediated strand displacement [23].

We develop a nucleic acid kinetic model based on Arrhenius dynamics that surpasses the performance of the Metropolis model. States in our continuous-time Markov chain (CTMC) model correspond to non-pseudoknotted secondary structures and each transition in the model corresponds to either the opening or closing of a base pair. A key difference with the Metropolis model is the use of activation energy, which depends on the immediate local environment surrounding the affected base pair. To calibrate and evaluate the Arrhenius and the Metropolis models, we compile a dataset of 376 experimentally determined reaction rate constants that we source from existing publications and cover a wide range of reactions, including hairpin closing and opening, bubble closing, helix association and dissociation, toehold-mediated 3-way strand displacement, and toehold-mediated 4-way strand exchange (see Fig. 1). To efficiently infer parameters and to obtain posterior parameter distributions, we use an ensemble Markov chain Monte Carlo (MCMC) approach. Similar to the Metropolis model, our model is consistent with existing thermodynamic models and Gillespie's stochastic simulation algorithm can be used to estimate kinetic rate constants for a variety of reactions. However, obtaining precise predictions using explicit stochastic simulation is computationally expensive, making MCMC parameter inference difficult. Instead we employ a reduced state space approach, enabling reaction rate constants to be computed efficiently and exactly using a sparse

matrix representation. Our state space is based on 'zipper models' that were investigated previously to model DNA hybridization [11].

Our results are encouraging and suggest that the new Arrhenius model is applicable to a wide range of DNA dynamic interactions and can be efficiently trained with our framework. The rest of this paper is organized as follows. Section 2 describes preliminaries and the Metropolis kinetic model, Section 3 introduces our Arrhenius kinetic model, Section 4 introduces our kinetic dataset, Section 5 introduces our inference framework, Section 6 describes our results comparing the inferred parameters to the database of experimental measurements, Section 7 discusses the limitations of our approach and directions for future research, and in Section 8 we describe details of the methods we used.

2 Preliminaries

In this section, we briefly discuss the type of reactions we are interested in modeling, and we discuss the Metropolis kinetic model (Section 2.1).

When DNA strands interact, base pairs form and break stochastically under the influence of thermal noise, resulting in a highly stochastic back-and-forth dynamic process. When two strands share a mutual base pair, we regard the strands as connected and we define a complex to be a set of connected strands. A single complex can have many different secondary structures. Similar to Kinfold [9] and Multistrand [20,21], we model the kinetics of interacting DNA strands as a CTMC, where the state space $\mathcal{S}$ is a set of non-pseudoknotted secondary structures. Transitions between states correspond to the forming or breaking of a single base pair, which may be called an elementary step. For example, in Fig. 2, state i can transition to states h and j. The rate at which a transition triggers is determined by a kinetic model, that is, the Metropolis or the Arrhenius model, and we distinguish between unimolecular and bimolecular transitions. Because all transitions in our model are reversible, we group transitions into pairs of forward and reverse reactions; a transition in the model is called bimolecular if a complex grows or shrinks by one strand, and is called unimolecular otherwise. As a result, successful helix association and helix dissociation both require at least one bimolecular transition to trigger, despite the latter reaction being strictly first order.

Experimentally observable reactions involve pathways of multiple elementary step transitions, are also inherently reversible, and thus can be classified similarly. We are interested in modeling both unimolecular and bimolecular reactions. In a unimolecular reaction, a complex of strands is altered through the formation or disruption of base pairs, but all strands in the complex remain connected. An example of a unimolecular reaction is hairpin closing (Fig. 1a), where a DNA strand hybridizes itself and forms a hairpin loop. Another example of a unimolecular reaction is bubble closing (Fig. 1b). Helix association (Fig. 1c) is a bimolecular reaction. Toehold-mediated 3-way strand displacement (Fig. 1d) is another example of a bimolecular reaction, where one of the strands in a duplex is replaced by the invader strand. The duplex consists of an incumbent

strand and a complementary strand. In addition to the hybridized domain, the incumbent strand also contains an unhybridized region called a toehold. The invading strand binds to the toehold region of the substrate and then displaces the incumbent strand via three-way branch migration. Another bimolecular example is toehold-mediated 4-way strand exchange (Fig. 1e), where two duplexes simultaneously exchange strands via four-way branch migration.

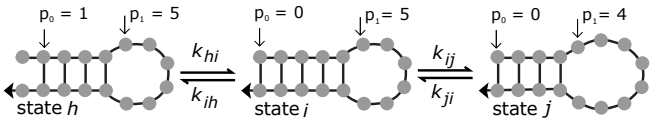

Fig. 2: State i can transition to states h and j. See Section 5.1 for definitions of the pointers p_0 and p_1.

2.1 The Metropolis Kinetic Model

The Metropolis model is one of the kinetic rate models implemented in Multistrand [20,21]. The Multistrand model considers a finite set of strands in a fixed volume ('the box') and defines the energy of a state as the sum of the standard free energy for each complex and a volume-dependent entropy term. To ensure that simulations converge to the Boltzmann distribution over the states at equilibrium, the transition rates between any two adjacent states i and j must satisfy detailed balance:

$$k_{ij}/k_{ji} = \exp\left\{-\left(\Delta G^0_{\text{box}}(j) - \Delta G^0_{\text{box}}(i)\right)/RT\right\} \tag{1}$$

where k_{ij} is the transition rate from state i to state j, $\Delta G^0_{\text{box}}(i)$ is the free energy of state i, R is the gas constant, and T is the temperature. For a state i containing $\mathcal{N}$ strands and $\mathcal{M}$ complexes, the free energy is

$$\Delta G^0_{\text{box}}(i) = \sum_{c=1}^{\mathcal{M}} \Delta G^0_{\text{complex}}(c) + (\mathcal{N} - \mathcal{M})\Delta G^0_{\text{volume}} \tag{2}$$

where $\Delta G^0_{\text{complex}}(c)$ is the difference in Gibbs free energy of complex c relative to the reference state and standard buffer conditions ($[\text{Na}^+] = 1$ M), and $\Delta G^0_{\text{volume}} = -RT \ln u$ is the loss of entropy resulting from fixing the position of a strand of concentration u relative to the standard concentration (1 M). Unimolecular transition rates are given by

$$k_{ij} = \begin{cases} k_{\text{uni}} & \text{if } \Delta G^0_{\text{box}}(j) < \Delta G^0_{\text{box}}(i) \\ k_{\text{uni}} \exp\left(\frac{\Delta G^0_{\text{box}}(i) - \Delta G^0_{\text{box}}(j)}{RT}\right) & \text{otherwise} \end{cases} \tag{3}$$

where $k_{\text{uni}} > 0$ is the unimolecular rate constant (units: s^{-1}). For bimolecular transitions $i \to j$ where two previously unconnected strands form a mutual base pair, the rate is given as

$$k_{ij} = k_{\text{bi}} u \tag{4}$$

and the rate of dissociation for the bimolecular transition $j \to i$ is given by

$$k_{ji} = k_{\text{bi}} e^{-\frac{\Delta G^0_{\text{box}}(i) - \Delta G^0_{\text{box}}(j) + \Delta G^0_{\text{volume}}}{RT}} \times \text{M} \tag{5}$$

where $k_{\text{bi}} > 0$ is the bimolecular rate constant (units: $\text{M}^{-1}\text{s}^{-1}$). We treat $\theta = \{\ln k_{\text{uni}}, \ln k_{\text{bi}}\}$ as 2 free parameters in the model that we calibrate to experimental measurements. We emphasize that the rate of dissociation, Eq. 5, is independent of concentration u and $\Delta G^0_{\text{volume}}$, which follows from the definition of the free energy in a state (Eq. 2).

3 The Arrhenius Kinetic Model

In our Arrhenius kinetic model, the activation energy of each transition depends on the immediate context of the closing or opening base pair. Our classification incorporates some, but not all, factors that may affect the activation energy of a transition. For example, the activation energy might depend on the strand sequence, but modeling this dependence would increase the number of free parameters, and we anticipate to have insufficient experimental evidence to accurately distinguish all relevant factors. However, we emphasize that transition rates in the model still depend on the nucleotide sequence via the nearest neighbor model of base pair stability that determines the free energy of a complex (see Eq. 3, 5).

Consider a reaction where a base pair is formed or broken, and denote by $l, r \in \mathcal{C}$ one half of the local context on either side of the base pair. Our model differentiates between seven different half contexts

$$\mathcal{C} = \{\text{stack}, \text{loop}, \text{end}, \text{stack+loop}, \text{stack+end}, \text{loop+end}, \text{stack+stack}\} \tag{6}$$

so that the set of local contexts is given by $\mathcal{C} \times \mathcal{C}$. The different half contexts are shown in Fig. 3. The Arrhenius model is equal to the Metropolis model (Eq. 3, 4, 5), except that we now re-define $k_{\text{uni}} : \mathcal{C} \times \mathcal{C} \to \mathbb{R}_{>0}$ and $k_{\text{bi}} : \mathcal{C} \times \mathcal{C} \to \mathbb{R}_{>0}$ by setting

$$k_{\text{uni}}(l, r) = k_l k_r \qquad k_l = A_l \exp\left(-E_l/RT\right) \qquad k_r = A_r \exp\left(-E_r/RT\right) \tag{7}$$

$$k_{\text{bi}}(l, r) = \alpha k_{\text{uni}}(l, r) \tag{8}$$

where A_l, A_r are Arrhenius rate constants, E_l, E_r are activation energies, and α is a bimolecular scaling constant. We treat $\theta = \{\ln A_l,\ E_l \mid \forall l \in \mathcal{C}\} \cup \{\alpha\}$ as 15 free parameters that we fit to data.

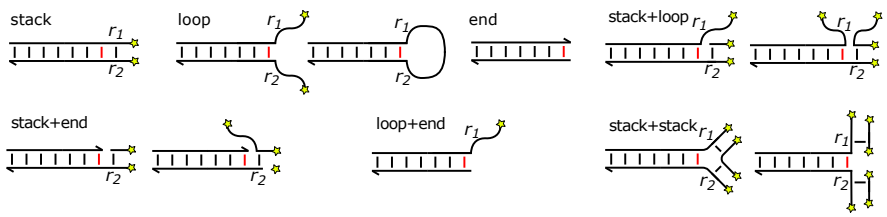

Fig. 3: The right side of the red base pair forms one half of the local context. The classification of the half context depends on the pairing status of the two bases r_1 and r_2 (if they exist) immediately to the right side of the base pair: stack means r_1 and r_2 form a base pair with each other, loop means that neither r_1 nor r_2 forms a base pair, end means that neither r_1 nor r_2 exists, stack+loop means that both r_1 and r_2 exist and one of the bases forms a base pair with another base while the other does not, stack+end means that only one of r_1 or r_2 exists and forms a base pair, loop+end means that only one of r_1 or r_2 exists and it does not form a base pair, and stack+stack means that both r_1 and r_2 exist and they both form base pairs with other bases. Stars indicate the possible continuation of the strands, which may be connected to other starred strands, provided the resulting complex is non-pseudoknotted.

4 Dataset

We compiled a dataset of experimentally determined reaction rate constants, extracting 376 reaction rate constants published in the literature. Each data point in our dataset is annotated with a reaction temperature and the concentration of Na^+ and Mg^{2+} cations in the buffer. The dataset is partitioned into a training set of size 320, which we call $\mathcal{D}_{\text{train}}$, and a testing set with size 56, which we call $\mathcal{D}_{\text{test}}$. The training set covers a wide range of observations, in terms of both reaction types and half contexts. The testing set includes both unimolecular and bimolecular reactions. An overview of our dataset is given in Table 1.

5 Modeling Framework

We augmented the Multistrand software [20,21] to implement the new Arrhenius model using the full state space of all non-pseudoknotted secondary structures. Given values for the 15 free parameters, a sufficient number of stochastic simulations could be run to estimate the models prediction for an experimental reaction of interest. Unfortunately, obtaining low error bars on this estimate is prohibitively slow, and thus is not feasible within the inner loop of parameter inference procedures. To address this limitation, we developed a computational framework in which we obtain fast, exact predictions for a feasible approximation of the full Multistrand state space. Specifically, we use a reduced state space that is a strict subset of the full state space, enabling sparse matrix computations of mean first passage times, from which reaction rate constants are predicted. With this computation in the inner loop, we used two methods for training the model. The first is a maximum a priori (MAP) approach that optimizes a single set of

Table 1: Dataset of experimentally measured reaction rate constants. The † sign indicates that the experiment was performed without Na^+ in the buffer, in which case our model computes the free energy as if 50 mM $[Na^+]$ is present (in addition to Mg^{2+}).

$\mathcal{D}_{\text{train}}$	$[Na^+]$ /M	$[Mg^{2+}]$ /mM	T / °C	Source
Hairpin closing and opening	0.1		10–49	Fig. 4 of Bonnet et al. [6]
	0.1–0.5		10–49	Fig. 6 of Bonnet et al. [6]
	0.25		18–49	Fig. 3.28 of Bonnet [5]
	0.137		20	Fig. 3 of Kim et al. [14]
Bubble closing	0.1		25–45	Fig. 4 of Altan-Bonnet et al. [2]
Association and dissociation	1.0		4–68	Fig. 6 of Morrison and Stols [17]
	0.05†	4	30–55	Fig. 6a of Reynaldo et al. [19]
Toehold-mediated 3-way strand displacement	0.05†	4	30–55	Fig. 6b of Reynaldo et al. [19]
	0.05†	12.5	25	Fig. 3b of Zhang and Winfree [28]
Toehold-mediated 4-way strand exchange	0.05†	12.5	25	Table 5.2 of Dabby [8]
$\mathcal{D}_{\text{test}}$				
Hairpin closing and opening	0.137		10–60	Fig. 5a, b of Kim et al. [14]
Toehold-mediated 3-way strand displacement with mismatches	0.05†	10	23	Fig. 2d of Machinek et al. [15]

parameters, and the second is based on MCMC that produces an ensemble of parameter sets. In the latter case, a posterior parameter probability density is computed.

5.1 State Space

In this section, we describe our reduced state space. In the future, our aim is to train the model using a larger set of non-pseudonotted secondary structures. In either case, the number of states in the model directly affects the computational cost of inference through the set of linear equations (Eq. 10 in Section 5.2) that is solved for each reaction at each iteration of the parameter search. In this study, the largest state space in the training data is toehold-mediated 4-way strand exchange and contains 14,438 states.

In our reduced state space, base pairs are permitted to form if and only if they occur in either the initial or final state of our simulation. For example, during the simulation of duplex hybridization, only base pairs that are consistent with the perfect alignment of the two strands are permitted to form. We further

prune the state space by only allowing base pairs to form or break at the edge of a hybridized domain.

A separate state space $\mathcal{S}_r$ is constructed for each reaction r that we wish to model (Fig. 1). Each state corresponds to a set of indices $\langle p_0, p_1, ... \rangle \in \mathcal{S}_r$, where the indices indicate the begin and end points of the hybridized domains. The maximum number of continuously hybridized domains is precisely defined for each reaction r. For example, the state space for hairpin closing and opening (Fig. 1a) and hybridization (Fig. 1c) only contain one hybridized domain. In such cases, the state description requires only two indices, and the length of the hybridized domain is given by $p_1 - p_0$. In Fig. 2, we show the pointers for the states h, i, and j in the state space for hairpin closing and opening. In each transition, one of the pointers is incremented or decremented. Specifically, state i can transition to state h by incrementing p_0 and it can transition to state j by decrementing p_1. We restrict $0 \leq p_0 \leq p_1 \leq m$, where m is the length of the stem in the closed state. If $p_0 = p_1$, then the domain is absent in the given state. A full description of the state space is given in the online appendix.

5.2 Estimating Mean First Passage Times with Exact Solvers

Given a parametrized kinetic model, we describe how to compute the mean first passage time of a CTMC with state space $\mathcal{S}$ using a sparse matrix representation. Let the mean first passage time t be the average time it takes to reach one of a set of final states $\mathcal{S}_{\text{final}}$ from an initial state i_0. For a first order reaction r, the reaction rate constant is found as $\hat{k}_r = \frac{1}{t}$. For a second order reaction, the reaction rate constant is computed as $\hat{k}_r = \frac{1}{t}\frac{1}{u}$ where u is the initial concentration of the reactants in the simulation [20]. A bimolecular reaction may be effectively first order or second order under the given conditions, depending on the time scale of the unimolecular portion of the reaction pathway relative to the overall reaction time. In our reaction kinetics dataset, all bimolecular reactions are second order in the forward direction.

Let the random variable T_i^{final} represent the time required to reach any state in $\mathcal{S}_{\text{final}}$ starting in state $i \in \mathcal{S}$, where $T_i^{\text{final}} = 0$ for $i \in \mathcal{S}_{\text{final}}$. The time required to reach $\mathcal{S}_{\text{final}}$ starting in i is equal to the initial holding time in state i, which we call h_i, plus the time required to hit $\mathcal{S}_{\text{final}}$ starting in the next visited state. h_i is distributed exponentially with exit rate $k_i = \sum_{j \in S} k_{ij}$. The probability to move to state j is directly proportional to the transition rate, so that $\mathrm{P}(i \to j) = \frac{k_{ij}}{k_i}$. Therefore, the mean first passage time is found as [24]

$$\mathbb{E}[T_i^{\text{final}}] = \frac{1}{k_i} + \sum_{j \in S} \frac{k_{ij}}{k_i} \mathbb{E}[T_j^{\text{final}}]. \tag{9}$$

Multiplying Eq. 9 by the exit rate k_i and applying $k_i = \sum_{j \in S} k_{ij}$ then yields

$$\sum_{j \in S} k_{ij} (\mathbb{E}[T_j^{\text{final}}] - \mathbb{E}[T_i^{\text{final}}]) = -1. \tag{10}$$

Eq. 10 permits a sparse matrix representation $\boldsymbol{K}\boldsymbol{t} = -\mathbf{1}$ for a rate matrix $\boldsymbol{K}$ and solution vector $\boldsymbol{t}$, where $\boldsymbol{K}_{ij} = k_{ij}$ for $i \neq j$, $\boldsymbol{K}_{ii} = -\sum_{j \in S} k_{ij}$, and $\boldsymbol{t}_i = \mathbb{E}[T_i^{\text{final}}]$. To compute first passage times for a distribution over initial states $\mathcal{S}_{\text{init}}$ rather than an individual state, the weighted average of the first passage time is computed.

5.3 Estimating the Unnormalized Posterior Distribution of the Parameters

Let θ be the set of parameters in a kinetic model. For a given experimentally observable reaction r, the predicted reaction rate constant $\hat{k}_r$ will deviate from the experimental measurement k_r. We define the error of the prediction to be the $\log_{10}$ difference, $\epsilon_r = \log_{10} k_r - \log_{10} \hat{k}_r$. To produce a measure of likelihood for our parameter valuation, we assume ϵ_r is normally distributed with an unbiased mean and variance σ^2, so that $\epsilon_r \sim N(0, \sigma^2)$. We treat σ as a nuisance parameter. For reaction r the likelihood function is given as

$$\mathrm{P}(r|\theta, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\left(\log_{10} k_r - \log_{10} \hat{k}_r\right)^2 \Big/ 2\sigma^2 \right\} \tag{11}$$

and the likelihood function over the set of training data is given as

$$\begin{aligned} \mathrm{P}(\mathcal{D}_{\text{train}}|\theta, \sigma) &= \prod_{r \in \mathcal{D}_{\text{train}}} \mathrm{P}(r|\theta, \sigma) \\ &= \exp\left\{ -\frac{\sum_{r \in \mathcal{D}_{\text{train}}} \left(\log_{10} k_r - \log_{10} \hat{k}_r\right)^2}{2\sigma^2} - \frac{n}{2} \log 2\pi\sigma^2 \right\} \end{aligned} \tag{12}$$

where n is the number of observations in $\mathcal{D}_{\text{train}}$. To define the probability of the parameters given the data we need to assume prior distributions for θ and σ. During preliminary fitting, a number of parameter values were found to be divergent, which we explain as follows. For a fixed temperature T and a fixed local context (l, r), there are many assignments of A_l, E_l and A_r, E_r that result in nearly equal transition rates $k_{\text{uni}}(l, r) = A_l A_r \exp\{-(E_l + E_r)/RT\}$ (we expand Eq. 7) that result in similar model predictions $\hat{k}_r$. This allows dissimilar valuations for E and A to have nearly equal (log)likelihood scores (Eq. 12). The problem becomes even more apparent when we consider the intrinsic measurement error on k_r (for example, a standard deviation of 22% was reported by Machinek et al. [15]), the limited range of temperatures (see Table 1) inherent to our observations, and the relative frequency of the different half contexts appearing in each simulation (see the online appendix). In practice, $k_{\text{uni}}(l, r)$ is well constrained for many different $l, r \in \mathcal{C}$. As is common in data-fitting applications, we assume a regularization prior that improves the stability of the estimation. We assume that all parameters in θ are independent and identically Gaussian distributed with mean 0 and variance $\frac{1}{\lambda}$. In our inference, we use $\lambda = 0.02$, and the predictive quality of the model does not change for minor adjustments to

λ. For the nuisance parameter σ, we use a non-informative Jeffreys prior [13]. Under these assumptions, the posterior distribution is proportional to:

$$\begin{aligned} \mathrm{P}(\theta, \sigma|\mathcal{D}_{\text{train}}) &= \frac{\mathrm{P}(\mathcal{D}_{\text{train}}|\theta, \sigma)\mathrm{P}(\theta)\mathrm{P}(\sigma)}{\mathrm{P}(\mathcal{D}_{\text{train}})} \propto \mathrm{P}(\mathcal{D}_{\text{train}}|\theta, \sigma)\mathrm{P}(\theta)\mathrm{P}(\sigma) \\ &= \mathrm{P}(\mathcal{D}_{\text{train}}|\theta, \sigma) \left(\frac{2\pi}{\lambda}\right)^{-\frac{|\theta|}{2}} \exp\left\{-\frac{\lambda\|\theta\|_2^2}{2}\right\} \frac{1}{\sigma}. \end{aligned} \tag{13}$$

In conclusion, the log of the posterior distribution is equal to the following equation, up to an additive constant not depending on the parameters

$$\begin{aligned} &\log P(\theta, \sigma|\mathcal{D}_{\text{train}}) \approx \\ &\quad -(n+1)\log\sigma - \frac{1}{2\sigma^2} \sum_{r \in \mathcal{D}_{\text{train}}} \left(\log_{10} k_r - \log_{10} \hat{k}_r\right)^2 - \frac{\lambda}{2}\|\theta\|_2^2 \end{aligned} \tag{14}$$

where the squared $L2$ norm in Eq. 14 is computed as $\|\theta\|_2^2 = \alpha^2 + |\ln k_{\text{uni}}|^2 + |\ln k_{\text{bi}}|^2$ for the Metropolis model and as $\|\theta\|_2^2 = \alpha^2 + \sum_{l \in \mathcal{C}} |\ln A_l|^2 + \sum_{l \in \mathcal{C}} |E_l|^2$ for the Arrhenius model. Note that $|\theta| = 2$ for the Metropolis model and $|\theta| = 15$ for the Arrhenius model.

Our MAP approach seeks a unique parameter set that maximizes the normalized log posterior of the dataset (Eq. 14). We use the Nelder-Mead optimization method [18], a gradient-free local optimizer. For MCMC, we use the *emcee* software package [10], that implements an affine invariant ensemble sampling algorithm.

Table 2: Performance of the Metropolis and the Arrhenius models on the training and testing sets. The Mean Squared Error (MSE) is the mean of $|\log_{10} k_r - \log_{10} \hat{k}_r|^2$ over $r \in \mathcal{D}$. The Within Factor of Three metric shows the percentage of reactions for which $|\log_{10} k_r - \log_{10} \hat{k}_r| \leq \log_{10} 3$. Initial is the initial parameter set of the MAP approach (Section 8). MAP is the MAP inference method. Mode is the parameter set from the MCMC ensemble that has the highest posterior on $\mathcal{D}_{\text{train}}$. Ensemble is the MCMC ensemble method where the reaction rate constant $\hat{k}_r$ is averaged over all parameter sets.

		Mean Squared Error		Within Factor of Three	
		$\mathcal{D}_{\text{train}}$	$\mathcal{D}_{\text{test}}$	$\mathcal{D}_{\text{train}}$	$\mathcal{D}_{\text{test}}$
Metropolis	Initial	.55	1.3	69.3%	33.9%
	MAP	.33	.94	79.0%	41.0%
	Mode	.33	.95	79.0%	41.0%
	Ensemble	.33	.99	79.6%	37.5%
Arrhenius	Initial	.59	1.3	71.2%	33.9%
	MAP	.14	.47	92.1%	73.2%
	Mode	.12	.40	92.8%	78.5%
	Ensemble	.12	.42	93.4%	78.5%

6 Results

Table 2 shows the performance of the Metropolis and the Arrhenius models with the MAP and MCMC approaches. For details on computational settings for the approaches see Section 8. The Arrhenius model fits the training data better than the Metropolis model (for details see the online appendix, Figs. S3-S14), which is unsurprising when considering the increase of adjustable parameters in the Arrhenius model (2 vs 15). However, the Arrhenius model also has better predictive qualities for the testing set, as evidenced by the MCMC ensemble mean standard deviation of $\sqrt{0.99} = 0.99$ for the Metropolis model and $\sqrt{0.42} = 0.64$ for the Arrhenius model. The improvement in the prediction of the testing set is apparent in Fig. 4, where both models predict the Machinek et al. study of toehold-mediated 3-way strand displacement with mismatches, and in predictions of opening and closing rates for hairpin with short stems (1-2 nt) (Figs. S15 and S16 in the online appendix). It is impressive that the models, when trained on a comprehensive training dataset, can predict the results of experiments not seen during training.

There are two reasons for the superior performance of the Arrhenius model. First, the presence of the temperature dependent activation energy allows the Arrhenius model to better calibrate to measurements at varying temperatures. On average, the reaction rate constants $k_{\text{uni}}(l, r)$ double in the Arrhenius model

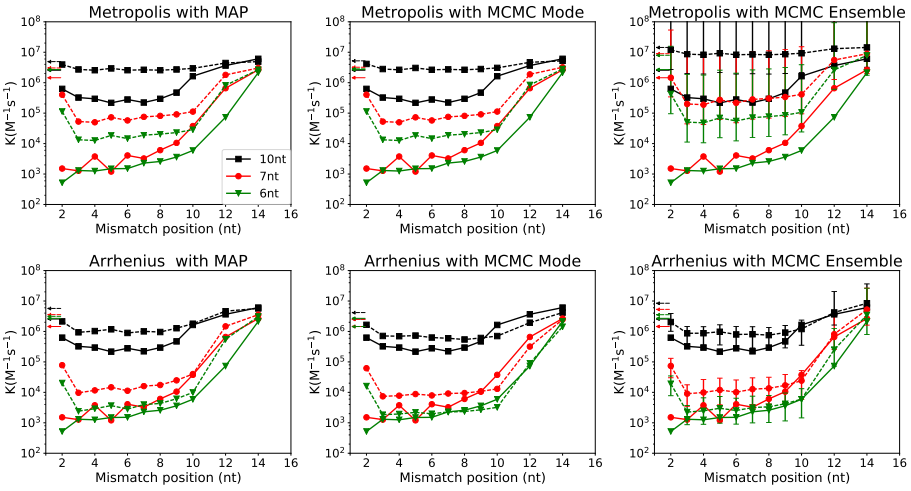

Fig. 4: Model predictions (dashed lines) of reaction rate constants (y axis) for toehold-mediated 3-way strand displacement with mismatches, experimental data (solid lines) from Fig. 2d of Machinek et al. [15]. For the MCMC ensemble method, error bars indicate the range (minimum to maximum) of 100 predictions (see Section 8). Arrows indicate no mismatch. The mismatch in the invading strand affects the reaction rate. The length of the toehold domain is ten, seven, and six nucleotides long for ■, ●, and ▼, respectively.

between $T = 25\,^{\circ}\mathrm{C}$ and $T = 60\,^{\circ}\mathrm{C}$ (this follows from the parameter values in which $\mathbb{E}[E_l + E_r] = 3.32$ kcal mol^{-1}). A second factor is the relation between the activation energy of a transition and the local context. In Fig. 5, the inferred distribution of $k_{\mathrm{uni}}(l, r)$ is given for all local contexts that occur in the model. Strikingly, for many local contexts, the $k_{\mathrm{uni}}(l, r)$ are narrowly distributed and often mutually exclusive, indicating that our model captures intrinsic qualitative differences in activation energy.

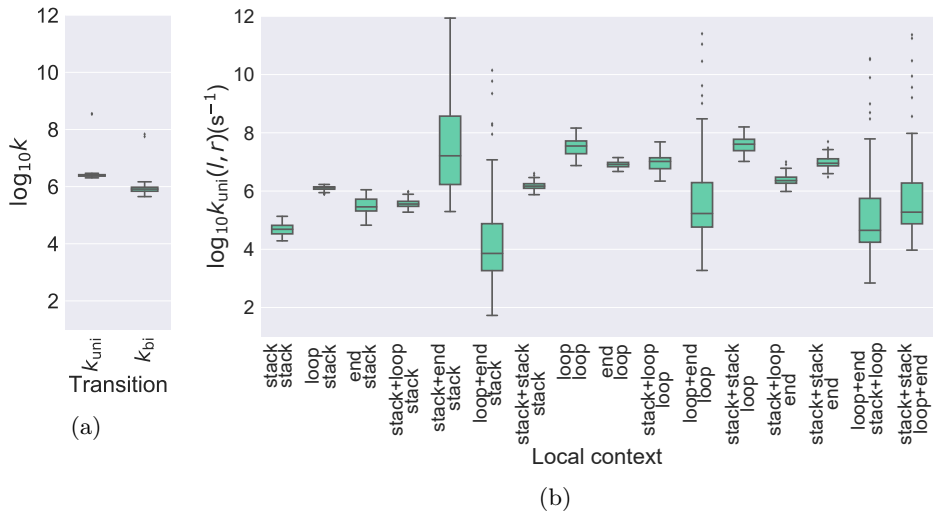

Fig. 5: Box plots of model features inferred by the MCMC ensemble method, using a sample of 100 parameter sets. Edges of the box correspond to the first and third quartile of the distribution. The whisker length is set to cover all parameter values in the sample, or is limited to at most 1.5 times the box height with the outliers plotted separately. a) k_{uni} and k_{bi} for the Metropolis model. b) $k_{\mathrm{uni}}(l, r)$ at $25\,^{\circ}\mathrm{C}$ for the Arrhenius model. Combinations that do not occur in the model are not shown.

7 Discussion

A common problem for Arrhenius models in biophysics is that the limited range of temperatures in experimental data can result in ambiguous parameter inference, and this is indeed the case for our model with the current data set. Despite the generally narrow bands for the transition rates (Fig. 5b), the inferred A and E parameters are poorly constrained, as is evident from the wide range in the parameter posterior probability distribution and correlation matrix (Fig. 6). Mathematically, measurements at a single temperature only restrict $\ln A_l + \frac{-E_l}{RT}$ rather than A_l and E_l independently, and a significant fraction of the measurements were performed at constant temperature. If further mining of the existing

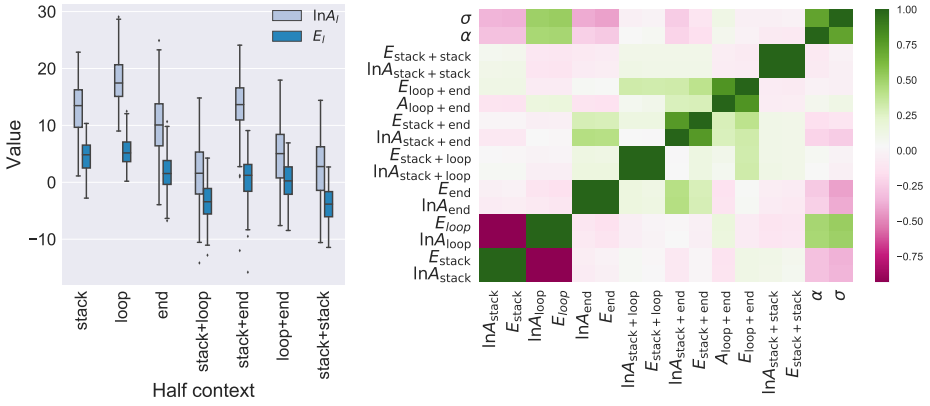

Fig. 6: The Arrhenius model parameters inferred by the MCMC ensemble method. a) Box plots of the half context parameters. Edges of the box correspond to the first and third quartile of the distribution. The whisker length is set to cover all parameter values in the sample, or is limited to at most 1.5 times the box height with the outliers plotted separately. b) The Pearson correlation coefficients $R_{ij} = \frac{\mathrm{cov}(\theta_i, \theta_j)}{\sigma_{\theta_i} \sigma_{\theta_j}}$, where $\mathrm{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$ and $\sigma_X = \sqrt{\mathbb{E}[(X - \mathbb{E}[X])^2]}$. (Color figure available online.)

experimental literature does not resolve the issue, one solution would be to develop customized experiments to calibrate the model further. Interestingly, the relative lack of correlation between the parameters for different half contexts suggests that there could be benefit in subdividing the half context categories further.

We envision further improvements to the model by adjusting the state space and the thermodynamic energy model. For the state space, the requirement for hybridizing strands to only engage in perfectly aligned base pairing is not realistic, and we plan on using a state space generated directly from stochastic Multistrand simulations to avoid these problems. Our simulation depends on the model of thermal stability implemented in the NUPACK software [27] and adjustments to the thermodynamic model also could improve the quality of our predictions. For example, hairpin closing rates are known to depend on the loop sequence, as open poly(A) loops are more rigid than poly(T) loops [1]. The current thermodynamic model does not incorporate this effect, and we avoid comparing the model to measurements on poly(A) loop hairpins. Similarly, the initiation of branch migration is known to have a significant thermodynamic cost, with one study measuring a cost of 2.0 kcal mol^{-1} at room temperature [23]. This initialization cost is not yet incorporated in NUPACK.

We have reported the initial results of our effort to develop accurate kinetic models for nucleic acids. Our Arrhenius model surpasses the performance of the Metropolis model, trained and evaluated on a wide range of experimental

DNA reaction rate constants. Although our current analysis focuses on DNA, we believe our approach would also apply to RNA reaction kinetics.

8 Methods

We fit the Metropolis and Arrhenius kinetic models using the MAP approach to a learn parameter set that maximizes Eq. 14. Using the MCMC approach, we maximize the same equation, but instead obtain an ensemble of parameter sets.

The MAP method is sensitive to the initial parameters, and for the Metropolis model, we use $k_{\text{uni}} = 8.2 \times 10^6\ \text{s}^{-1}$ and $k_{\text{bi}} = 3.3 \times 10^5\ \text{M}^{-1}\text{s}^{-1}$, following known estimates for a one dimensional model of toehold-mediated strand displacement [23]. For the Arrhenius model, we initialize $E_r = 3\ \text{kcal mol}^{-1}$ for all $r \in \mathcal{C}$ and we initialize α and A_r such that, at $T = 23\,^\circ\text{C}$, equally $k_{\text{uni}}(l, r) = 8.2 \times 10^6\ \text{s}^{-1}$ and $k_{\text{bi}}(l, r) = 3.3 \times 10^5\ \text{M}^{-1}\text{s}^{-1}$ for all local contexts $l, r \in \mathcal{C}$. For both models, we initialize $\sigma = 1$.

Results for the MCMC should generally depend less on the initial value of the sets in the ensemble. To initialize the parameter assignment for each parameter set in the MCMC ensemble, we realize random variables

$$\begin{aligned} E_r &\sim U(0, 6) \times \text{ kcal mol}^{-1} & A_r &\sim U(0, 10^4) \times s^{-1/2} & \forall r \in \mathcal{C} \\ k_{\text{uni}} &\sim U(0, 10^8) \times \text{ s}^{-1} & k_{\text{bi}} &\sim U(0, 10^8) \times \text{ M}^{-1}\text{s}^{-1} \\ \alpha &\sim U(0, 10) \times \text{M}^{-1} & \sigma &\sim U(0, 1) \end{aligned} \tag{15}$$

where $U(a, b)$ is the uniform distribution over (a, b). During the inference, the parameters are not restricted to initialization bounds, and instead we only require $k_{\text{uni}}, k_{\text{bi}}, A_l, \alpha$ and σ to be positive.

In the emcee software [10], an ensemble of walkers each represents a set of parameters, which are updated through *stretch moves*. Given two walkers θ_1 and θ_2, a new parameter assignment θ_1' for the first walker is generated as

$$\theta_1' = Z\theta_1 + (1 - Z)\theta_2 \qquad g(Z = z) \propto \begin{cases} \frac{1}{\sqrt{z}} & \text{if } z \in \left[\frac{1}{a}, a\right] \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

where $g(z)$ is the probability density of Z. We use $a = 2$ (default value) and an ensemble of 100 walkers. We only use the last step of each walker to make predictions, which results in an ensemble of 100 parameter sets for each model.

For the MAP approach, we continue the inference until an absolute tolerance of 10^{-4} is reached. For the MCMC approach, we continue the inference until 750 iterations are performed per walker.

We implemented our framework in Python. All experiments were run on a system with 16 2.93GHz Intel Xeon processors and 64GB RAM, running openSUSE 42.1. On this system, each iteration takes less than 6 s.

Our framework and dataset, as well as an online appendix that has a full description of the state space, more experimental plots and analysis, and algorithms that underlie our framework, are available at https://github.com/DNA-and-Natural-Algorithms-Group/ArrheniusInference.

Acknowledgements. We thank the U.S. National Science Foundation (awards 0832824, 1213127, 1317694, 1643606), the Gordon and Betty Moore Foundation's Programmable Molecular Technology Initiative, and the Natural Sciences and Engineering Research Council of Canada for support. We also thank the anonymous reviewers for their helpful comments and suggestions. XR's current address is Descartes Labs, Los Alamos, NM, USA.

References

1. Aalberts, D.P., Parman, J.M., Goddard, N.L.: Single-strand stacking free energy from DNA beacon kinetics. Biophysical Journal 84, 3212–3217 (2003)
2. Altan-Bonnet, G., Libchaber, A., Krichevsky, O.: Bubble dynamics in double-stranded DNA. Physical Review Letters 90, 138101 (2003)
3. Andronescu, M., Aguirre-Hernandez, R., Condon, A., Hoos, H.H.: RNAsoft: a suite of RNA secondary structure prediction and design software tools. Nucleic Acids Research 31, 3416–3422 (2003)
4. Andronescu, M., Condon, A., Hoos, H.H., Mathews, D.H., Murphy, K.P.: Computational approaches for RNA energy parameter estimation. RNA 16(12), 2304–2318 (2010)
5. Bonnet, G.: Dynamics of DNA breathing and folding for molecular recognition and computation. Ph.D. thesis, Rockefeller University (2000)
6. Bonnet, G., Krichevsky, O., Libchaber, A.: Kinetics of conformational fluctuations in DNA hairpin-loops. Proceedings of the National Academy of Sciences 95(15), 8602–8606 (1998)
7. Chen, S.J.: RNA folding: conformational statistics, folding kinetics, and ion electrostatics. Annu. Rev. Biophys. 37, 197–214 (2008)
8. Dabby, N.L.: Synthetic molecular machines for active self-assembly: prototype algorithms, designs, and experimental study. Ph.D. thesis, California Institute of Technology (2013)
9. Flamm, C., Fontana, W., Hofacker, I.L., Schuster, P.: RNA folding at elementary step resolution. RNA 6, 325–338 (2000)
10. Foreman-Mackey, D., Hogg, D.W., Lang, D., Goodman, J.: emcee: The MCMC hammer. Publications of the Astronomical Society of the Pacific 125, 306 (2013)
11. Gibbs, J., DiMarzio, E.: Statistical mechanics of helix-coil transitions in biological macromolecules. The Journal of Chemical Physics 30, 271–282 (1959)
12. Hofacker, I.L.: Vienna RNA secondary structure server. Nucleic Acids Research 31, 3429–3431 (2003)
13. Jeffreys, H.: An invariant form for the prior probability in estimation problems. In: Proceedings of the Royal Society of London a: mathematical, physical and engineering sciences. vol. 186, pp. 453–461. The Royal Society (1946)
14. Kim, J., Doose, S., Neuweiler, H., Sauer, M.: The initial step of DNA hairpin folding: a kinetic analysis using fluorescence correlation spectroscopy. Nucleic Acids Research 34, 2516–2527 (2006)
15. Machinek, R.R., Ouldridge, T.E., Haley, N.E., Bath, J., Turberfield, A.J.: Programmable energy landscapes for kinetic control of DNA strand displacement. Nature Communications 5 (2014)
16. Mathews, D.H., Sabina, J., Zuker, M., Turner, D.H.: Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. Journal of Molecular Biology 288(5), 911–940 (1999)

17. Morrison, L.E., Stols, L.M.: Sensitive fluorescence-based thermodynamic and kinetic measurements of DNA hybridization in solution. Biochemistry 32, 3095–3104 (1993)
18. Nelder, J.A., Mead, R.: A simplex method for function minimization. The Computer Journal 7, 308–313 (1965)
19. Reynaldo, L.P., Vologodskii, A.V., Neri, B.P., Lyamichev, V.I.: The kinetics of oligonucleotide replacements. Journal of Molecular Biology 297, 511–520 (2000)
20. Schaeffer, J.M.: Stochastic simulation of the kinetics of multiple interacting nucleic acid strands. Ph.D. thesis, California Institute of Technology (2012)
21. Schaeffer, J.M., Thachuk, C., Winfree, E.: Stochastic simulation of the kinetics of multiple interacting nucleic acid strands. In: Proceedings of the 21st International Conference on DNA Computing and Molecular Programming-Volume 9211 (2015)
22. Schreck, J.S., Ouldridge, T.E., Romano, F., Šulc, P., Shaw, L.P., Louis, A.A., Doye, J.P.: DNA hairpins destabilize duplexes primarily by promoting melting rather than by inhibiting hybridization. Nucleic acids research 43(13), 6181–6190 (2015)
23. Srinivas, N., Ouldridge, T.E., Šulc, P., Schaeffer, J.M., Yurke, B., Louis, A.A., Doye, J.P., Winfree, E.: On the biophysics and kinetics of toehold-mediated DNA strand displacement. Nucleic Acids Research 41, 10641–10658 (2013)
24. Suhov, Y., Kelbert, M.: Probability and Statistics by Example: Volume 2, Markov Chains: A Primer in Random Processes and Their Applications, vol. 2. Cambridge University Press (2008)
25. Xayaphoummine, A., Bucher, T., Isambert, H.: Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots. Nucleic Acids Research 33, W605–W610 (2005)
26. Xu, Z.Z., Mathews, D.H.: Experiment-assisted secondary structure prediction with RNAstructure. RNA Structure Determination: Methods and Protocols pp. 163–176 (2016)
27. Zadeh, J.N., Steenberg, C.D., Bois, J.S., Wolfe, B.R., Pierce, M.B., Khan, A.R., Dirks, R.M., Pierce, N.A.: NUPACK: analysis and design of nucleic acid systems. Journal of Computational Chemistry 32, 170–173 (2011)
28. Zhang, D.Y., Winfree, E.: Control of DNA strand displacement kinetics using toehold exchange. Journal of the American Chemical Society 131, 17303–17314 (2009)
29. Zuker, M.: Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Research 31, 3406–3415 (2003)