Chapter 2

System

We are interested in simulating nucleic acid molecules (DNA or RNA) in a stochastic regime; that is to say that we have a discrete number of molecules in a fixed volume. This regime is found in experimental systems that have a small volume with a fixed count of each molecule present, such as the interior of a cell. We can also apply this to experimental systems with a larger volume (such as a test tube) when the system is well mixed, as we can pick a fixed (small) volume and deal with the expected counts of each molecule within it, rather than the whole test tube.

To discuss the modeling and simulation of the system, we need to be very careful to define the components of the system, and what comprises a state of the system within the simulation.

2.1 Strands

Each DNA molecule to be simulated is represented by a *strand*. Our system then contains a set of strands Ψ^*, where each strand $s \in \Psi^*$ is defined by $s = (id, label, sequence)$. A strand's *id* uniquely identifies the strand within the system, while the *sequence* is the ordered list of nucleotides that compose the strand.

Two strands could be considered *identical* if they have the same sequence. However, in some cases it is convenient to make a distinction between strands with identical sequences. For example, if one strand were to be labelled with a fluorophore, it would no longer be physically identical to another with the same sequence but no fluorophore. Thus, the *label* is used to designate whether two strands are identical. We define two strands as being *identical* if they have the same labels and sequences. In most cases this distinction between the label and the sequence is not used, so it will be explicitly noted when it is important.

2.2 Complex Microstate

A *complex* is a set of strands connected by base pairing (secondary structure). We define the state of a complex by $c = (ST, \pi^*, BP)$, called the "complex microstate". The components are a nonempty set of strands $ST \subseteq \Psi^*$, an ordering π^* on the strands ST, and a list of base pairings $BP = \{(i_j \cdot k_l) \mid$ base i on strand j is paired to base k on strand l, and $j \leq l$, with $i < k$ if $j = l\}$, where we note that "strand l" refers to the strand occuring in position l in the ordering π^*. Note that we require a complex to be "connected": there is no proper subset of strands in the complex for which the base pairings involving those strands do not involve at least one base outside that subset. Given a complex microstate c, we will use $ST(c), \pi^*(c), BP(c)$ to refer to the individual components.

While this definition defines the full space of complex microstates, it is common to disallow some secondary structures due to physical or computational constraints. For example, we disallow the pairing of a base with any other within three bases on the same strand, as this would correspond to an impossible physical configuration. Another class of disallowed structures are called the *pseudoknotted* secondary structures, which require computationally difficult energy model calculations, and are fully defined and discussed further in Section D.

2.3 System Microstate

A system microstate represents the configuration of the strands in the volume we are simulating (the "box"). Since we allow complexes to be formed of one or more strands, every unique strand in the system must be present in a single complex and thus we can represent the system microstate by a set of those complexes.

We define a *system microstate i* as a set of complex microstates, such that each strand in the system is in exactly one complex within the system. This is formally stated in the following equation:

$$\bigcup_{c \,\in\, i} ST(c) = \Psi^* \text{ and } \forall c, c' \in i \text{ with } c \neq c', ST(c) \cap ST(c') = \emptyset \tag{2.1}$$

This definition leads to the natural use of $|i|$ to indicate the number of complexes present in system microstate i, and $i \setminus j$ to indicate the complex microstates present in system microstate i that are not in j.

Chapter 3

Energy

The conformation of a nucleic acid strand at equilibrium can be predicted by a well studied model, called the nearest neighbor energy model [16, 15, 17]. Recent work has extended this model to cover systems with multiple interacting nucleic acid strands [5].

The distribution of system microstates at equilibrium is a Boltzmann distribution, where the probability of observing a microstate i is given by

$$Pr(i) = \frac{1}{Q} * e^{-\Delta G_{box}(i)/RT} \tag{3.1}$$

where $\Delta G_{box}(i)$ is the free energy of the system microstate i, and is the key quantity determined by these energy models. $Q = \sum_i e^{-\Delta G_{box}(i)/RT}$ is the partition function of the system, R is the gas constant, and T is the temperature of the system.

3.1 Energy of a System Microstate

We now consider the energy of the system microstate i, and break it down into components. The system consists of many complex microstates c, each with their own energy. We also must account for the entropy of the system (the number of configurations of the complexes spatially within the "box") in the energy, and thus must define these two terms.

Let us first consider the entropy term. We consider the "zero" energy system microstate to be the one in which all strands are in separate complexes, thus our entropy term is in terms of the reduction of available states caused by having strands join together. We assume that the number of complexes in the system, N, is much smaller than the number of solvent molecules within our box, M_s. We can then approximate the standard statistical entropy of the system as $N * RT \log M_s$. Letting K be the total number of strands in the system, our zero state is then $K * RT \log M_s$. Defining $\Delta G_{volume} = RT \log M_s$, the contribution to the energy of the system microstate i from the entropy of the box is then:

$$(K - N) * \Delta G_{volume}$$

And thus in terms of $N, K, \Delta G_{volume}$ and $\overline{\Delta G}(c)$ (the energy of complex microstate c, defined in the next section), we define $\Delta G_{box}(i)$, the energy of the system microstate i, as follows:

$$\Delta G_{box}(i) = (K - N) * \Delta G_{volume} + \sum_{c \in i} \overline{\Delta G}(c)$$

Before we turn our attention to the energy of a complex microstate, let us examine a situation where we are modeling a fixed volume of a larger solution, and how that relates to the quantity M_s. If we assume that the volume studied is V (units of liters), we can easily compute the number of solvent molecules in this volume using the density d of water (the solvent, in g/L) , the molar mass M of water (in grams per mol), and Avogadro's number, as follows: $M_s = \frac{V*d}{M} * N_A$. In practice, we may wish to choose a simulation volume V based on other physical quantities, such as the concentration of a single molecule within the volume[1].

The energy formulas derived here, suitable for our stochastic model, differ from those in [5] in two main ways: the lack of symmetry terms, and the addition of the ΔG_{volume} term. We compare this stochastic model to the mass action model in much more detail in Section C.

[1] Calculating M_s from a concentration u in mol/L of a single molecule in the volume is straightforward. We assume that our concentration u implies the volume V is chosen corresponding to exactly one molecule being present in that volume, as follows: $V = \frac{1}{u*N_A}$ and thus $M_s = \frac{d}{M*u} = \frac{\rho_{H_2O}}{u}$ where ρ_{H_2O} is the molarity of water (55.14 mol/L at 37 °C) and the other quantities are as defined above.

3.2 Energy of a Complex Microstate

We previously defined a complex microstate in terms of the list of base pairings present within it. However, the well studied models are based upon nearest neighbor interactions between the nucleic acid bases. These interactions divide the secondary structure of the system into local components which we refer to as *loops*, shown in figure 3.1.

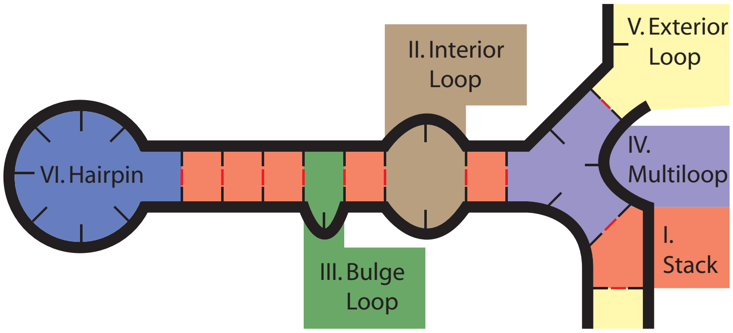

Figure 3.1: Secondary structure divided into loops.

These loops can be broken down into different categories, and parameter tables for each category have been determined from experimental data [17]. Each loop l has an energy, $\Delta G(l)$ which can be retrieved from the appropriate parameter table for its category, which is discussed in more detail in section B.3. Each complex also has an energy contribution associated with the entropic initiation cost [3] (e.g. rotational) of bringing two strands together, ΔG_{assoc}, and the total contribution is proportional to the number of strands L within the complex, as follows [2]: $(L-1) * \Delta G_{assoc}$.

[2] The free energy ΔG° for a reaction $A + B \rightleftharpoons C$ is usually expressed in terms of the equilibrium constant K_{eq} and the concentrations $[A],[B],[C]$ (in mol/L) of the molecules involved, as follows: $e^{\Delta G^\circ/RT} = K_{eq} = \frac{[A][B]}{[C]}$. We can also express the free energy $\Delta G'$ in terms of the dimensionless mole fractions x_A, x_B, x_C, where $x_i = [i]/\rho_{H_2O}$ (for dilute solutions), and ρ_{H_2O} is the molarity of water. In this case, we have $e^{\Delta G'/RT} = K'_{eq} = \frac{x_A * x_B}{x_C}$, and relating it to the previous equation, we see that $e^{\Delta G'/RT} = \frac{([A]/\rho_{H_2O})*([B]/\rho_{H_2O})}{[C]*\rho_{H_2O}} = \frac{[A][B]}{[C]} * \frac{1}{\rho_{H_2O}} = e^{\Delta G^\circ/RT} * e^{-\log \rho_{H_2O}}$. Thus if we have an energy ΔG° which was for concentration units and we wish to use mole fraction units, we must adjust it by $-RT \log \rho_{H_2O}$ to obtain the correct quantity. In general, if we have a complex of N molecules, the conversion to mole fractions will require an adjustment of $-(N-1) * RT \log \rho_{H_2O}$. To be consistent with [5], we wish to always use free energies which are based on the mole fraction units, and thus must include this factor since the reference free energies are for concentration units. In [5], the factor is included in the ΔG_{assoc} term, and thus we include it in the same place, as follows: $\Delta G_{assoc} = \Delta G^{pub}_{assoc} - RT \log \rho_{H_2O}$, where ΔG^{pub}_{assoc} is found in [3]. Thus our ΔG_{assoc} is the same as the ΔG^{assoc} found in [5] (footnote 13).

The energy of a complex microstate c is then the sum of these two types of contributions. We can also divide any free energy ΔG into the enthalpic and entropic components, ΔH and ΔS related by $\Delta G = \Delta H + T * \Delta S$, where T is the temperature of the system. For a complex microstate, each loop can have both enthalpic and entropic components, but ΔG_{assoc} is usually assumed to be purely entropic [16]. This becomes important when determining the kinetic rates, in section 4.

We use $\overline{\Delta G}(c)$ to refer to the energy of a complex microstate to be consistent with the nomenclature in [5], where $\overline{\Delta G}(c)$ refers to the energy of a complex when all strands within it are consider unique (as is the case in our system), and $\Delta G(c)$ is the energy of the complex, without assuming that all strands are unique (and thus it must account for rotational symmetries). This is discussed more in Section C.

In summary, the standard free energy of a complex microstate c, containing $L = |ST(c)|$ strands:

$$\overline{\Delta G}(c) = \left(\sum_{\text{loop } l \in c} \Delta G(l) \right) + (L-1)\Delta G_{assoc}$$

3.3 Computational Considerations

While the simulator could use the system microstate energy in the form given in the previous sections, it is convenient for us to group terms such that the computation need only take place per complex. Thus we wish to include the $(K - N)\Delta G_{volume}$ term in the energy computation for the complex microstates. Recall that K is the number of strands in the system, and N is the number of complexes in the system microstate. Assuming that we are computing the energy ΔG_{box} of system microstate i, we can rewrite K and N as follows:

$$\begin{aligned} K &= \sum_{c\,\in\, i} |ST(c)| \\ N &= \sum_{c\,\in\, i} 1 \end{aligned}$$

And thus arrive at:

$$\Delta G_{box}(i) = \sum_{c \in i} \left(\overline{\Delta G}(c) + (|ST(c)| - 1) * \Delta G_{volume} \right)$$

We then define $\Delta G^*(c) = \overline{\Delta G}(c) + (|ST(c)| - 1) * \Delta G_{volume}$, and $L_c = |ST(c)|$ and thus have the following forms for the energy of a system microstate and the energy of a complex microstate:

$$\begin{aligned} \Delta G_{box}(i) &= \sum_{c \in i} \Delta G^*(c) \\ \Delta G^*(c) &= \left(\sum_{\text{loop } l \in c} \Delta G(l) \right) + (L_c - 1) * (\Delta G_{assoc} + \Delta G_{volume}) \end{aligned}$$

Since we expect the probability of observing a particular complex microstate to remain the same no matter what reference units we use for the free energy (see footnote 2), this implies that if we wanted to express our $\Delta G^*(c)$ for concentration units, we would use $\Delta G_{assoc} = \Delta G_{assoc}^{pub}$ and $\Delta G_{volume} = RT \log \frac{M_s}{\rho_{H_2O}} = RT \log \frac{1}{u} = RT \log \frac{V}{V_0}$, where u is the molar concentration of a single molecule in the box volume V, and V_0 is the volume for 1 molecule at the standard concentration of 1 M.

Chapter 4

Kinetics

4.1 Basics

Thermodynamic predictions have only limited use for some systems of interest, if the key information to be gathered is the reaction rates and not the equilibrium states. Many systems have well defined ending states that can be found by thermodynamic prediction, but predicting whether it will reach the end state in a reasonable amount of time requires modeling the kinetics. Kinetic analysis can also help uncover poor sequence designs, such as those with alternate reactions leading to the same states, or kinetic traps which prevent an intended reaction from occurring quickly.

The kinetics are modeled as a continuous time Markov process over secondary structure space. System microstates i, j are considered adjacent if they differ by a single base pair (Figure 4.1), and we choose the transition rates k_{ij} (the transition from state i to state j) and k_{ji} such that they obey detailed balance:

$$\frac{k_{ij}}{k_{ji}} = e^{-\frac{\Delta G_{box}(j) - \Delta G_{box}(i)}{RT}} \tag{4.1}$$

This property ensures that given sufficient time we will arrive at the same equilibrium state distribution as the thermodynamic prediction, (i.e. the Boltzmann distribution on system microstates, equation 3.1) but it does not fully define the kinetics as this only constrains the ratio $\frac{k_{ij}}{k_{ji}}$. We discuss how to choose these transition rates in the following sections, but regardless of this choice, we can still determine how the next state is chosen and the time at which that transition occurs.

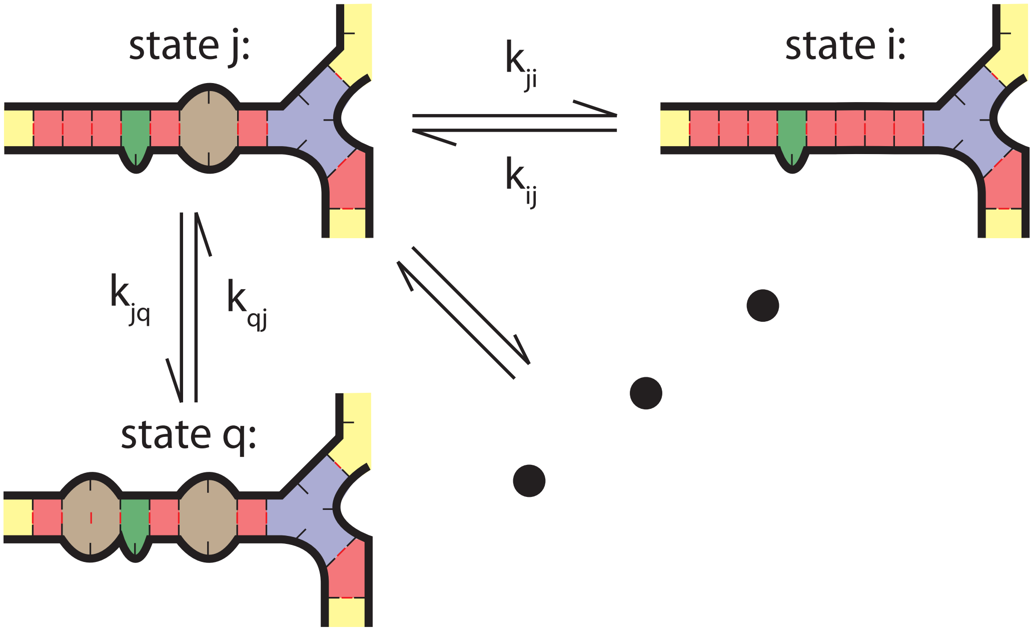

Figure 4.1: System microstates i, q adjacent to current state j, with many others not shown.

Given that we are currently in state i, the next state m in a simulated trajectory is chosen randomly among the adjacent states j, weighted by the rate of transition to each.

$$Pr(m) = \frac{k_{im}}{\Sigma_j k_{ij}} \tag{4.2}$$

Similarly, the time taken to transition to the next state is chosen randomly from an exponential distribution with rate parameter λ, where λ is the total rate out of the current state, $\Sigma_j k_{ij}$.

$$Pr(\Delta t) = \lambda \exp(-\lambda \Delta t) \tag{4.3}$$

We will now classify transitions into two exclusive types: those that change the number of complexes present in the system, called *bimolecular transitions*, and those where changes are within a single complex, called *unimolecular transitions*.

4.2 Unimolecular Transitions

Because unimolecular transitions involve only a single complex, it is natural to define these transitions in terms of the complex microstate which changed, rather than the full system microstate. Like Figure 4.1 implies, we define a complex microstate d as being adjacent to a complex microstate c if it differs by exactly one base pair. We call a transition from c to d that adds a base pair a *creation* move, and a transition from c to d that removes a base pair a *deletion* move. The exclusion of pseudoknotted structures is not inherent in this definition of adjacent states, but rather arises from our disallowing pseudoknotted complex microstates.

In more formal terms we now define the adjacent states to a system microstate, rather than those adjacent to a complex microstate as in the simple definition above. Recall from section 2.3 that $|i|$ is the number of complexes present in system microstate i, and $i \setminus j$ is the set of complex microstates in i that are not also in system microstate j.

Two system microstates i, j are adjacent by a unimolecular transition iff $\exists c \in i, d \in j$ such that:

$$|i| = |j| \text{ and } i \setminus j = \{c\} \text{ and } j \setminus i = \{d\} \tag{4.4}$$

and one of these two holds:

$$BP(c) \subset BP(d) \text{ and } |BP(d)| = |BP(c)| + 1 \tag{4.5}$$

$$BP(d) \subset BP(c) \text{ and } |BP(c)| = |BP(d)| + 1 \tag{4.6}$$

In other words, the only differences between i and j are in c and d, and they differ by exactly one base pair. If equation 4.5 is true, we call the transition from i to j a *base pair creation move*, and if equation 4.6 is true, we call the transition from i to j a *base pair deletion move*. Note that if i to j is a creation move, j to i must be a deletion move, and vice versa. Similarly, if there is no transition from i to j, there cannot be a transition from j to i, which implies that every unimolecular move in this system is reversible.

4.3 Bimolecular Transitions

A bimolecular transition from system microstate i to system microstate j is one where the single base pair difference between them leads to a differing number of complexes within each system microstate. This differing number of complexes could be due to a base pair joining two complexes in i to form a single complex in j, which we will call a *join move*. Conversely, the removal of this base pair from i could cause one complex in i to break into two complexes within j, which we will call a *break move*. Note that if i to j is a join move, then j to i must be a break move, and vice versa. As we saw before, this also implies that every bimolecular move is reversible.

Formally, a transition from system microstate i to system microstate j is a join move if $|i| = |j| + 1$ and we can find complex microstates $c, c' \in i$ and $d \in j$, with $c \neq c'$ such that the following equations hold:

$$i \setminus \{c, c'\} = j \setminus \{d\} \tag{4.7}$$

$$\exists x \in BP(d) \text{ s.t. } BP(d) \setminus \{x\} = BP(c) \cup BP(c') \tag{4.8}$$

Similarly, a transition from system microstate i to system microstate j is a break move if $|i| + 1 = |j|$ and we can find complex microstates $c \in i$ and $d, d' \in j$ with $d \neq d'$ such that the following equations hold:

$$i \setminus \{c\} = j \setminus \{d, d'\} \tag{4.9}$$

$$\exists x \in BP(c) \text{ s.t. } BP(c) \setminus \{x\} = BP(d) \cup BP(d') \tag{4.10}$$

While arbitrary bimolecular transitions are not inherently prevented from forming pseudoknots in this model, we again implicitly prevent them by using only complex microstates that are not pseudoknotted.

4.4 Transition Rates

Now that we have defined all of the possible transitions between system microstates, we must decide how to assign rates to each transition. We know that if there is a transition from system microstate i to system microstate j with rate k_{ij} there must be a transition from j to i with rate k_{ji} which are related by:

$$\frac{k_{ij}}{k_{ji}} = e^{-\frac{\Delta G_{box}(j) - \Delta G_{box}(i)}{RT}} \tag{4.11}$$

This condition is known as *detailed balance*, and does not completely define the rates k_{ij}, k_{ji}. Thus a key part of our model is the choice of *rate method*, the way we set the rates of a pair of reactions so that they obey detailed balance.

While our simulator can use any arbitrary rate method we can describe, we would like our choice to be physically realistic (i.e. accurate and predictive for experimental systems). There are several rate methods found in the literature [10, 11, 26] which have been used for kinetics models for single-stranded nucleic acids [7, 26] with various energy models. As a start, we have implemented three of these simple rate methods which were previously used in single base pair elementary step kinetics models for single stranded systems. In addition we present a rate method for use in bimolecular transitions that is physically consistent with both mass action and stochastic chemical kinetics. We verify that the kinetics model (and thus our choice of rate method) have been correctly implemented by verifying that the detailed balance condition holds (Section 6.1.2).

In order to maintain consistency with known thermodynamic models, each pair of k_{ij} and k_{ji} must satisfy detailed balance and thus their ratio is determined by the thermodynamic model, but in principle each pair could be independently scaled by some arbitrary prefactor, perhaps chosen to optimize agreement with experimental results on nucleic acid kinetics. However, since the number of microstates is exponential, this leads to far more model parameters (the prefactors) than is warranted by available experimental data. For the time being, we limit ourselves to using only two scaling factors: k_{uni} for use with unimolecular transitions, and k_{bi} for bimolecular transitions.

4.5 Unimolecular Rate Models

The first rate model we will examine is the Kawasaki method [10]. This model has the property that both "downhill" (energetically favorable) and uphill transitions scale directly with the steepness of their slopes.

$$\begin{aligned} k_{ij} &= k_{uni} * e^{-\frac{\Delta G_{box}(j) - \Delta G_{box}(i)}{2RT}} && (4.12) \\ k_{ji} &= k_{uni} * e^{-\frac{\Delta G_{box}(i) - \Delta G_{box}(j)}{2RT}} && (4.13) \end{aligned}$$

The second rate model under consideration is the Metropolis method [11]. In this model, all downhill moves occur at the same fixed rate, and only the uphill moves scale with the slope. This means that the maximum rate for any move is bounded, and in fact all downhill moves occur at this rate. This is in direct contrast to the Kawasaki method, where there is no bound on the maximum rate.

$$\begin{aligned} \text{if } \Delta G_{box}(i) > \Delta G_{box}(j) \text{ then} \quad k_{ij} &= 1 * k_{uni} && (4.14) \\ k_{ji} &= k_{uni} * e^{-\frac{\Delta G_{box}(i) - \Delta G_{box}(j)}{RT}} && (4.15) \\ \text{otherwise,} \quad k_{ij} &= k_{uni} * e^{-\frac{\Delta G_{box}(j) - \Delta G_{box}(i)}{RT}} && (4.16) \\ k_{ji} &= 1 * k_{uni} && (4.17) \end{aligned}$$

Finally, the entropy/enthalpy method [26] uses the division of free energies into entropic and enthalpic components to assign the transition rates in an intuitive manner: base pair creation moves must overcome the entropic energy barrier to bring the bases into contact, and base pair deletion moves must overcome the enthalpic energy barrier in order to break them apart.

$$\begin{aligned}
\text{if } i \text{ to } j \text{ is a creation:} \quad k_{ij} &= k_{uni} * e^{\frac{\Delta S_{box}(j) - \Delta S_{box}(i)}{R}} && (4.18)\\
k_{ji} &= k_{uni} * e^{-\frac{\Delta H_{box}(i) - \Delta H_{box}(j)}{RT}} && (4.19)\\
\text{otherwise,} \quad k_{ij} &= k_{uni} * e^{-\frac{\Delta H_{box}(j) - \Delta H_{box}(i)}{RT}} && (4.20)\\
k_{ji} &= k_{uni} * e^{\frac{\Delta S_{box}(i) - \Delta S_{box}(j)}{R}} && (4.21)
\end{aligned}$$

We note that the value of k_{uni} that best fits experimental data is likely to be different for all three models.

4.6 Bimolecular Rate Model

When dealing with moves that join or break complexes, we must consider the choice of how to assign rates for each transition in a new light. In the particular situation of the join move, where two molecules in a stochastic regime collide and form a base pair, this rate is expected to be modeled by stochastic chemical kinetics.

Stochastic chemical kinetics theory [8] tells us that there should be a rate constant k such that the propensity of a particular bimolecular reaction between two species X and Y should be $k * \#X * \#Y/V$, where $\#X$ and $\#Y$ are the number of copies of X and Y in the volume V. Since our simulation considers each strand to be unique, $\#X = \#Y = 1$, and thus we see the propensity should scale as $1/V$. Recalling that $\Delta G_{volume} = RT \log(V * y)$, where y is a collection of constant terms (discussed in Section 3.1) and V is the simulated volume, we see that we can obtain the $1/V$ scaling by letting the join rate be proportional to $e^{-\Delta G_{volume}/RT}$.

Thus, we arrive at the following rate method, and note that the choice of k (above) or our scalar term k_{bi} can be found by comparison to experiments measuring the hybridization rate of oligonucleotides [21].

$$\begin{aligned}
\text{if } i \text{ to } j \text{ is a complex join move:} \quad k_{ij} = &\quad k_{bi} * e^{\frac{-\Delta G_{volume}}{RT}} && (4.22)\\
k_{ji} = &\quad k_{bi} * e^{-\frac{\Delta G_{box}(i) - \Delta G_{box}(j) + \Delta G_{volume}}{RT}} && (4.23)\\
\text{otherwise,} \quad k_{ij} = &\quad k_{bi} * e^{-\frac{\Delta G_{box}(j) - \Delta G_{box}(i) + \Delta G_{volume}}{RT}} && (4.24)\\
k_{ji} = &\quad k_{bi} * e^{-\frac{\Delta G_{volume}}{RT}} && (4.25)
\end{aligned}$$

This formulation is convenient for simulation, as the join rates are then independent of the resulting secondary structure. We could use the other choices for assigning rates from 4.4, but they would require much more computation time. While the above model is of course an approximation to the physical reality (albeit one which we believe at least intuitively agrees with what we expect from stochastic chemical kinetics), if we later determine there is a better approximation we could use that instead, even if it cost us a bit in computation time. One issue in the above model that we wish to revisit in the future is that due to the rate being determined for **every** possible first base pair between two complexes, the overall rate for two complexes to bind (by a single base pair) is proportional roughly to the square of the number of exposed nucleotides, in addition to the $\frac{1}{V}$ dependence noted earlier.