Targeting Selectivity in CRISPR CAS9

Han Yao Choong

August 14, 2017

Abstract

This report describes a project to investigate selectivity of sgRNA targeting sequences in the CRISPR CAS9 system. Using DNA thermodynamics, models were constructed which describe binding between the crisprRNA and the targeted genome in equilibrium state. Non-equilibrium behaviour has also been investigated using kinetic models. A significant result from the simulations is the existence of a minimum targeting sequence length of 16-18 nucleotides from equilibrium thermodynamic models.

1 Introduction

The CRISPR-Cas9 system is an immune mechanism found in the bacteria *streptococcus pyogenes* which targets invading viruses and plasmids [3]. The system consists of a Cas9 nuclease binded in a complex with a sgRNA, in turn consisting of a crRNA and tracrRNA (fig. 1).

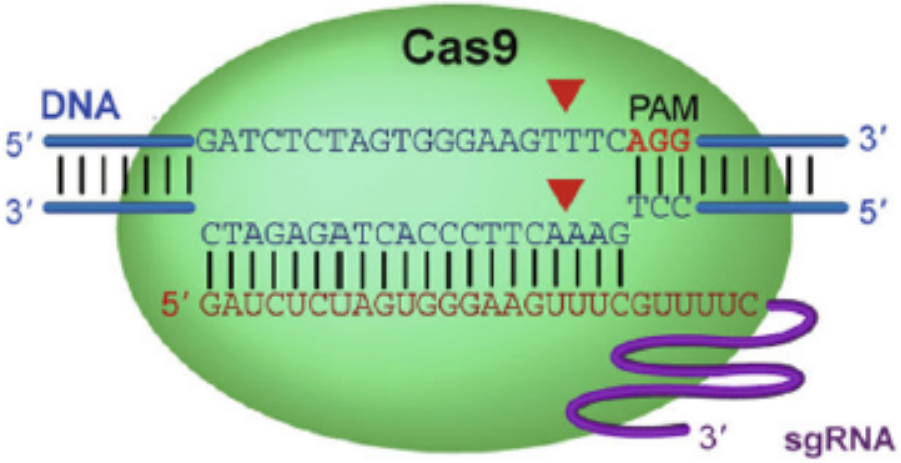


Figure 1: A diagram displaying the CRISPR Cas9 system, consisting of the Cas9 nuclease and sgRNA, which binds to the targeted DNA[3].

The system relies on the RNA sequence at the 5' end of the sgRNA to recognise the targeted site in the genome. Once the sgRNA is bound to the targeted site, two independent nuclease domains in the Cas9 each cleave one of the DNA strands at the position adjacent to the protospacer adjacent motif (PAM) sequence and result to a double stranded break (DSB) [3]. Using two sgRNA/Cas9 complexes to carry out DSBs at positions up from 10 to 10^6 bp apart, genome editing can then facilitated through the homology-directed repair mechanism (HDR) to introduce alternative DNA into the gap.

To explore questions surrounding off-target binding, the quantity of 'targeting specificity' S is herein introduced as a measure of the proportion of on-target to off-target binding events between a targeted genome and well-defined targeting sequence, under equilibrium conditions. Explicitly in the thermodynamic models to be considered, the selectivity S is defined as follows,

$$S = \frac{1}{\frac{Z}{q_{comp}} - 1} \tag{1}$$

where Z is the partition function and q_{comp} is the probability that the targeting sequence binds to the targeted (complementary) sequence in the genome which has the form $p(os = cs)e^{-\beta\Delta G}$, where $p(os = cs)$ is the probability that the opposite sequence is the complementary sequence and ΔG is the free energy of binding (hybridization) between the sequences.

Calculations of S in thermodynamic models in this project relies upon the DNA thermodynamic database of SantaLucia et al [1], [2] by first considering purely the thermodynamics of DNA hybridisation in equilibrium state. Several thermodynamic models were constructed with increasing complexity, culminating in an incorporation of the nearest neighbour (NN) model. To include features distinct to the CRISPR-Cas9 system, a kinetic model imposes requirements in binding time such that the time needed for Cas9 to carry out a DSB will only be satisfied for certain sequences.

Finally, while noting that the project's central aim is to investigate in particular the CRISPR CAS9 system, many of the methods and results in this investigation can also be applied to several other biological 'barcode recognition' problems.

2 Notation

Item	Quantity
N_S^X	Number of X nucleotides in targeting sequence
l	Number of nucleotides in targeting sequence
$N_{G_S}^X$	Number of X nucleotides in part of genome bound to targeting sequence
N_G^X	Number of X nucleotides in genome
N_G	Number of nucleotides in genome
p_X	Frequency of X nucleotides in genome
$p(X\|W)$	Frequency of X nucleotides in genome, given that the previous nucleotide is W
N_{cp}	The number of complementary pairs bound between genome and targeting sequence
X^*	Complementary to X
$\overline{X}$	Not Complementary to X
$(-)\epsilon$	Binding Energy (Attractive interaction denoted by -)

Table 1: Definition of variables used

3 Model 1

3.1 Model Description & Definitions

In this initial model, as is the case with all subsequent thermodynamic models, S is calculated via partition function (Z) calculations. These calculations are based on considering energetic states as strictly binary. That is, they correspond to whether the targeting sequence is completely hybridized with the opposite sequence or completely unhybridized. When hybridized, the energetic state is given by the sum of all ΔG contributions from the sequence, relative to a zero defined as the completely unhybridized state. In the thermodynamic models, the targeting sequence is a perfectly ('user') defined sequence, serving as the input variable. The opposite sequence can be considered as a random variable described by model parameters such as p_X and $p(X|W)$. The partition function then expresses the expectation $E(p(os)e^{-\beta\Delta G}) = E(p_{X_1}\prod_{i=1}^{l-1} p(X_{i+1}|X_i)e^{-\beta\Delta G})$ given the mean field model parameters and the targeting sequence.

For model 1, two key assumptions are in place, combining to form a highly simplistic initial model. The first assumption is that the occurrence of specific nucleotides in the targeted genome are not dependent on the identity of neighbours, i.e. $p(X|W) = p_X$ for all 16 W, X combinations. The second assumption is that for all hybridizations involving a Watson Crick base pair, a contribution of ϵ=-0.054eV is added to the free energy of hybridization ΔG, while for non-WC base pairs no contribution is added (0eV). Hence model 1 is not a NN model of the form presented in SantaLucia et al. (1998) [1] and only considers binding of opposite nucleotides in the sequences, without regard to neighbours. Model 1 is thus a parameterisation of the binding selectivity $\boldsymbol{S}$ in terms of the nucleotide composition of the targeting sequence (N_S^X), targeting sequence length (l), relative nucleotide frequencies in the targeted genome (p_X), temperature and a degenerate defect energy ϵ which is identical for all of the 8 possible pairwise defects.

Explicitly outlining the model, once again we note that a general definition of S can be given as

$$S = \frac{1}{\frac{Z}{q_{comp}} - 1} \tag{2}$$

while noting that, as to be seen in subsequent discussions and the appendix, crucial distinctions are to be made between S_{ideal} and several S_{real} which represent corrections to S_{ideal}. Nevertheless, in the general definition of S, q_{comp} is given by

$$\begin{aligned} q_{comp} &= p_A^{N_S^T} p_T^{N_S^A} p_C^{N_S^G} p_G^{N_S^C} e^{-\beta l \epsilon} \\ &(= p_c e^{-\beta \Delta G_c}) \end{aligned} \tag{3}$$

and Z for model 1 is given by

$$\begin{aligned} Z &= \sum_{N_{cp}, N_{G_S}^A, N_{G_S}^T, N_{G_S}^C, N_{G_S}^G}^{4^l seqs} q_{(N_{cp}, N_{G_S}^A, N_{G_S}^T, N_{G_S}^C, N_{G_S}^G)} \\ &= \sum_{N_{cp}, N_{G_S}^A, N_{G_S}^T, N_{G_S}^C, N_{G_S}^G}^{4^l seqs} p_A^{N_{G_S}^A} p_T^{N_{G_S}^T} p_C^{N_{G_S}^C} p_G^{N_{G_S}^G} e^{-\beta N_{cp} \epsilon} \\ &(= Z_{ideal}) \end{aligned} \tag{4}$$

To simplify this summation expression, the approach of partition function factorization can be used to write Z in the form

$$\begin{aligned} Z &= \prod_{i=1, X_i \in \{ts\}}^{l} [1 + p_{X_i^*}(e^{-\beta\epsilon} - 1)] \\ &(= Z_{ideal}) \end{aligned} \tag{5}$$

In addition to calculating the partition function Z and the selectivity S, a useful quantity '$A_{N_{cp}}$' can be defined. Noting from equation 4, Z can be written as $Z = \sum^{4^l seqs} p_A^{N_{G_S}^A}, p_T^{N_{G_S}^T}, p_C^{N_{G_S}^C}, p_G^{N_{G_S}^G}, e^{-\beta N_{cp} \epsilon}$. This summation can be expressed more succinctly as

$$Z = \sum_{N_{cp}=0}^{l} A_{N_{cp}} e^{-\beta N_{cp} \epsilon} \tag{6}$$

i.e. grouping together the summation components with common factors of $e^{-\beta N_{cp} \epsilon}$ such that the summation is carried out over the number of complementary pairs between the sequences (N_{cp}). Here $A_{N_{cp}}$ is a number which encapsulates the probability and degeneracy information of the possible states of the genome binding region with N_{cp} fixed. $A_{N_{cp}}$ in turn is given by:

$$A_{N_{cp}} = \sum_{(N_{cp}=const), N_{G_S}^A, N_{G_S}^T, N_{G_S}^C, N_{G_S}^G} D_{(N_{cp}, N_{G_S}^A, N_{G_S}^T, N_{G_S}^C, N_{G_S}^G)} p_A^{N_{G_S}^A} p_T^{N_{G_S}^T} p_C^{N_{G_S}^C} p_G^{N_{G_S}^G} \tag{7}$$

where D is the degeneracy (the number of states) of the binding region of the genome having a fixed number of A,T,C,G nucleotides, and complementary pairs as constraints. (D can be algorithmically calculated, its expression is too long to be displayed here).

However if the probabilities p_X are degenerate ($pX{=}p{=}0.25$ for all X), then $A_{N_{cp}}$ can be written in a simpler form (which is illustrative to consider).

$$A_{N_{cp}} = 3^{(l - N_{cp})} \frac{l!}{N_{cp}!(l - N_{cp})!} p^l \tag{8}$$

Developing on the meaning and purpose of $A_{N_{cp}}$, it is useful to note that the system has two distinct temperature regimes that give rise to S with different characteristics. Note that in the high temperature limit as the exponential argument associated with the energetic term tends to zero, the selectivity is solely described by the relative frequencies of nucleotides in the targeted genome. Hence $Z = \sum_{N_{cp}=0}^{l} A_{N_{cp}}$ and as expected, S becomes extremely low ($S \to \frac{p_c}{\sum_{N_{cp}=0}^{l} A_{N_{cp}}}$, or could approach $\frac{1}{N_G}$). In the low temperature limit, it is then clear to see that the energetic contribution associated with the exponential term dominates in comparison to the frequency contribution.

3.2 Results & Discussion

The following figure shows the output of a computational calculation of the partition function components done for a particular targeting sequence.

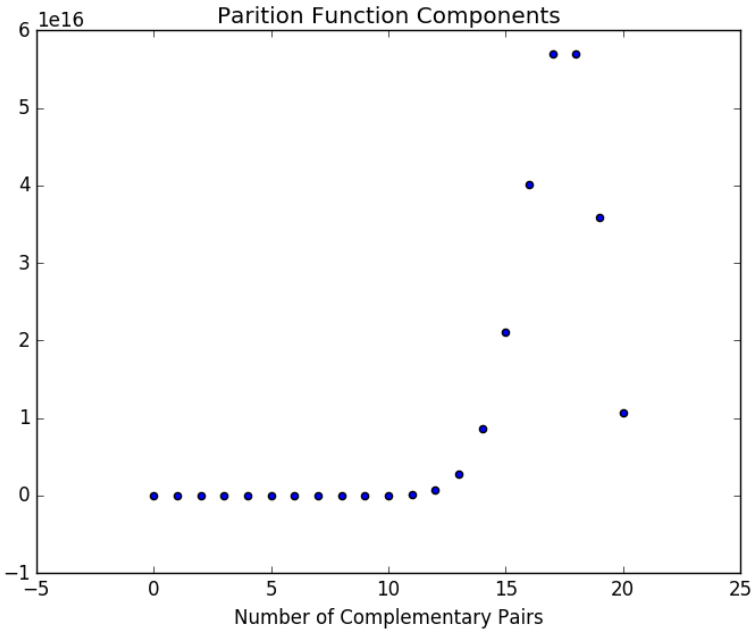


Figure 2: A plot of the 'constant n_{cp}' components ($A_{n_{cp}} e^{-\beta n_{cp} \epsilon}$) of the partition function against the number of complementary pairs in the system, N_{cp}. There are 21 data points corresponding to each possible value of N_{cp}

In this simulation, the value usedpair binding energy. for ϵ is 0.1eV, while T=300K. The trial targeting sequence used was TTTATATACTTTTTGTTTTG (or equivalently, any sequence with n_T=14, n_A=3, n_G=2, n_C=1), with targeted genome nucleotide frequencies of 0.1, 0.2, 0.3 and 0.4 for A,T,C,G respectively. (Note that these values were chosen quite arbitrarily, while a sequence with a large number of a particular nucleotide was chosen to shorten the running time, which grows substantially for more 'mixed' sequences with $N_S^X \to 5$ for all X).

The calculation of the partition function components can be quite insightful because it demonstrates the proportion of binding events that occur for sequences of all possible number of defects (0-20). In the figure above, the peak away from the n_{cp}=20 shows that actually the majority of binding events would occur for sequences with 2 and 3 defects. From this it can be seen that the energetic penalty associated with the defects, combined with their sufficiently high probabilistic occurrence leads to a large favourability associated with binding to sequences with 2-3 defects. It is only until 6-7 defects with the proportion of binding events becomes negligible.

Mathematically, a formulation brining together the defect energetic penalties and probabilities then becomes useful to help understand figure 2. In this formulation the partition function can be expressed as a sum over 21 components corresponding to each of the allowed N_{cp} values:

$$Z = \sum_{N_{cp}=0}^{l} A_{N_{cp}} e^{-\beta N_{cp} \epsilon} = \sum_{N_{cp}=0}^{l} e^{-\beta N_{cp} \epsilon + ln A_{N_{cp}}} \tag{9}$$

where $-\beta N_{cp}\epsilon + lnA_{N_{cp}}$ can now be defined as an effective free energy $\Delta G'_{N_{cp}}$, incorporating both combinatorial and energetic information for each N_{cp} component of the partition function.

To investigate the variation of selectivity S as a function of targeting sequence length l, model 1 was set to a targeting sequence consisting of only T nucleotides with equal genome nucleotide frequencies of p_A=p_T=p_C=p_G=0.25. An important component of the full investigation of $S(l)$ also consists of investigating differences between S_{real} and S_{ideal}, as introduced in the previous section **??**. To motivate the need for a distinction between S_{ideal} and S_{real}, it should be kept in mind that a particular 20nt sequence appearing in a chromosome is generally likely to be unique in the entire chromosome. To appreciate this fact, if the relative nucleotide frequencies are considered to be equal while their appearance in a point in the chromosome is purely random, then it would take up to 10000 chromosomes before the same 20nt sequence appears again, or alternatively three hundred 3×10^9 bp genomes. In other words, if the genome were truly random, it is quite likely that the complementary sequence would not appear in the first place. Therefore the distinction made between S_{ideal} and S_{real} is that the genome is respectively treated as being completely random and semi-random, where the existence of a complementary sequence is guaranteed by mathematical corrections. Hence S_{real} correspond to more realistic scenarios where the existence of a complementary sequence is known. See appendix for derivation and more details.

With both S_{ideal} and S_{real}. In both cases the variation was found to be consistently monotonic (exponential), with no extrema at finite l. The plots below show the result for these targeting sequences from length of 1 to 40 nucleotides. In this case the targeting sequence was set to only consist of T. ϵ in this model was chosen to be -0.054eV (see appendix for more details). All defect binding energies are set to be 0eV. The temperature was set as 310K. For the calculation of the S_{ideal}, a value for the genome length of 3Gbp was used, assuming that the targeting sequence has (complete and uniform) access to the entire human genome in a cell.

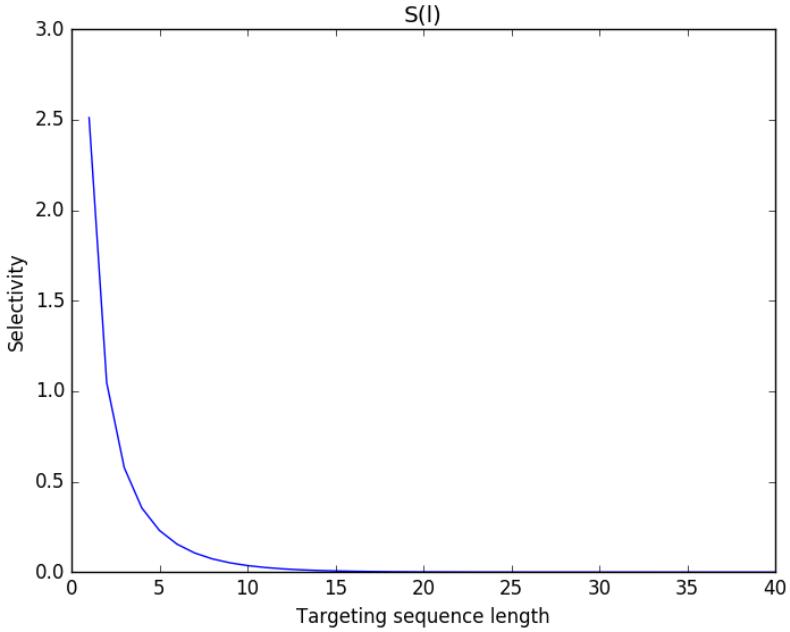


Figure 3: A plot of S_{ideal} against the targeting sequence length l.

The figure above shows a monotonic decrease of S against l. This is merely as a consequence of the simple definition of the selectivity in this case as $S_{ideal} = \frac{(pe^{-\beta\epsilon})^l}{[1+p(e^{-\beta\epsilon}-1)]^l-(pe^{-\beta\epsilon})^l}$. By directly substituting in values used in the simulation, one arrives at the expressions $S_{ideal} \approx \frac{1.88^l}{2.63^l-1.88^l} = \frac{1}{1.40^l-1}$.

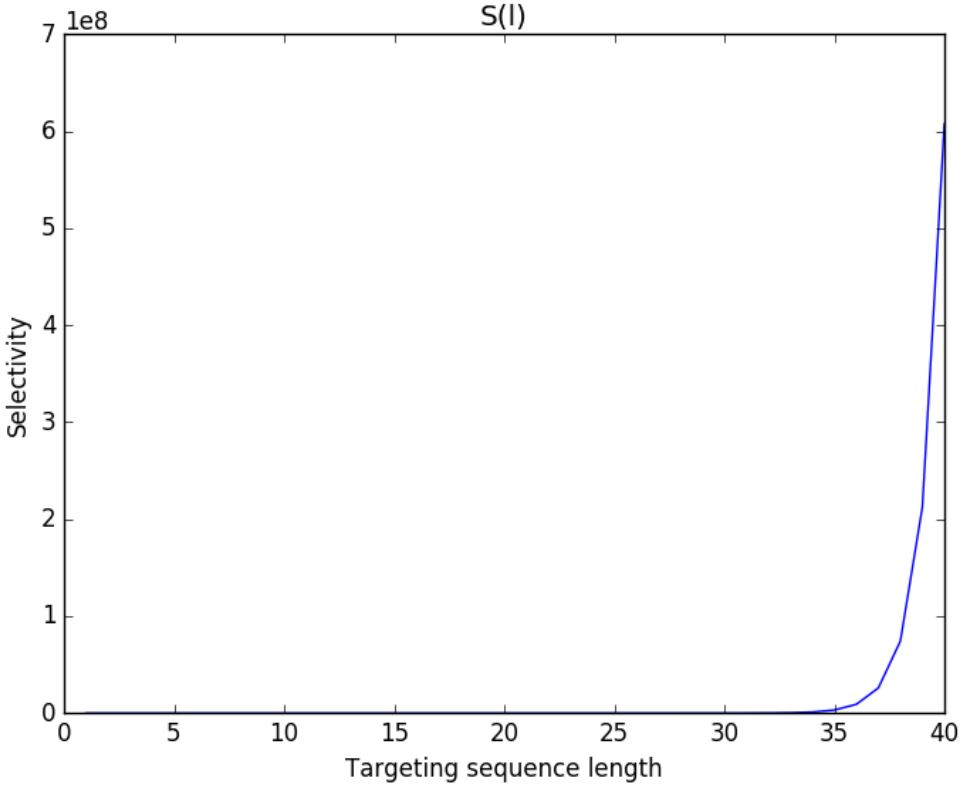


Figure 4: A plot of S_{real} against the targeting sequence length l.

A simplified expression for the S_{real} is $\frac{1}{N_G - l}\frac{(e^{-\beta\epsilon})^l}{[1+p(e^{-\beta\epsilon}-1)]^l-(pe^{-\beta\epsilon})^l}$. Similarly directly substituting in values used, $S_{real} = \frac{1}{3\times10^9}\frac{7.54^l}{2.63^l-1.88^l}$.

For S_{real}, it can be seen that the $e^{-\beta l\epsilon}$ term in q_{comp} becomes dominant for large l.

Due to the fact that a random 20nt sequence requires several hundred genomes to re-occur, the Z_{real} correction does indeed result to more accurate selectivity compared to the uncorrected mean field model which assigns a very low probability to the occurrence of the complementary sequence in the targeted genome hence resulting to the decreasing selectivity seen as a function of l.

However now, two questions are needed to be addressed for the S_{real} result. Firstly, is the increase in selectivity over l realistic and expected physically? Secondly, can this result for S_{real} be said to be valid for small S? It appears that if that is the case, it may indeed provide some insight into the optimality of the 20nt targeting sequence as it may be consistent with the rapid increase in selectivity around $l = 20$ as seen.

It may be that this model is satisfactory after all and that its main shortcoming is that it assumes that binding energy can increase infinitely, which leads to the dominance of the energetic term. If the S_{real} is incorporated with a more accurate energetic model describing energetic penalties at large sequence lengths, an optimal value for l may well emerge in the expected regime (given N_G=3Bbp, of which l_{opt} is a function). The fact that a rapid increase occurs at around 20nt for l may seem a promising indication.

(However could it just be a coincidence? (+))

To address this question the following expression for S_{real} can be considered:

$$S_{real} = \frac{1}{\frac{\frac{1}{N_G-l+1}e^{-\beta\Delta G_c}+\frac{N_G-l}{N_G-l+1-l}Z_{ideal}}{\frac{1}{N_G-l+1}e^{-\beta\Delta G_c}+\frac{N_G-l}{N_G-l+1}p_c e^{-\beta\Delta G_c}}-1} \tag{10}$$

$$S_{real} \approx \frac{1}{\frac{\frac{1}{N_G}e^{-\beta\Delta G_c}+Z_{ideal}}{\frac{1}{N_G}e^{-\beta\Delta G_c}+p_c e^{-\beta\Delta G_c}}-1} \tag{11}$$

As seen in the equation above, the point at which the selectivity rapidly increases can be considered to be approximately when the condition $\frac{1}{N_G}e^{-\beta\Delta G_c} = Z_{ideal}$ is satisfied. Approximating Z_{ideal} as $[1 + p(e^{-\beta\epsilon} - 1)]^l$, l can be solved as $l = \frac{-\beta\Delta G_c - lnN_G}{ln[1+p(e^{-\beta\epsilon}-1)]}$ using considerations from model 1. From values used, a solution of $l = 19.2$ was found. Evidently, it is important to note again that this simplistic model and calculation has genome nucleotide frequencies equal at $p = 0.25$.

Intuitively, this condition simply corresponds to the scenario where the variation of l causes binding to the 'imposed' complementary sequence and other non-complementary sequences to be equally likely, due to the increasing dominance associated with the complementary binding energy over the partition function Z_{ideal} representing the combined probabilistic and energetic likelihood of non-complementary binding, which increases more slowly as l increases (at least in model 1).

Even if there may be energetic corrections missing in this model (corrections from biological considerations and Santa Lucia rules may result to differing results), by equating $\frac{1}{N_G}e^{-\beta\Delta G_c} = Z_{ideal}(l)$ and solving for l, an estimate for the optimal l may be yielded. The general approach of equating $q_c omp$ and Z may therefore be a good first estimate for an optimal value for l.

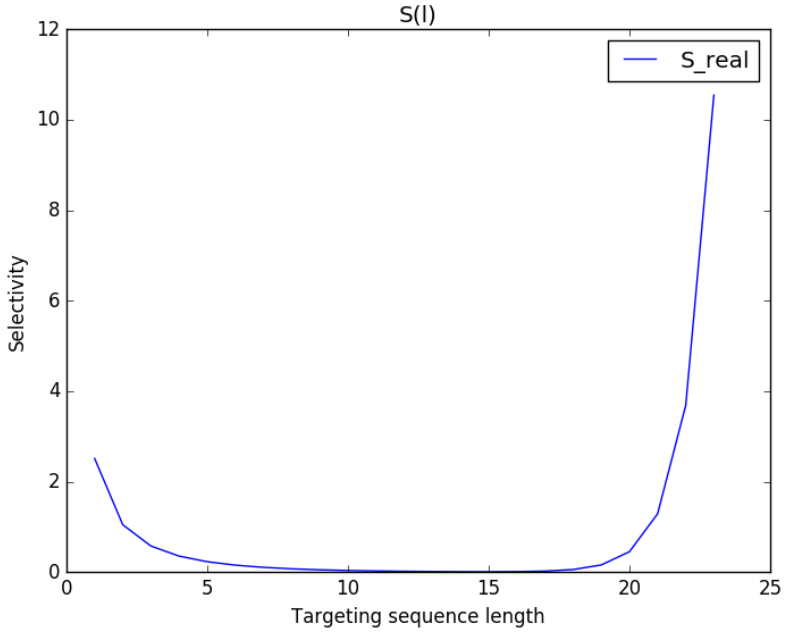


Figure 5: A plot of the modified S_{real} against l. Here the modified S_{real} is that which allows 'correct binding' to be at more than one site in the genome, should the complementary sequence occur 'randomly' via the mean field.

The figure above shows an alternative version of S_{real} where all complementary sites are counted as being 'correct binding' sites rather than just one particular complementary site in the entire genome. The expression for this S_{real} is given by $\frac{[1+(N_G-l)p_c]}{(N_G-l)p_c}S_{ideal}$. The extra factor arises from the fact that at low l regimes, 'correct bindings' can be dominated bindings with complementary sequences away from the 'imposed' complementary sequence. (Taking the extreme example of a 1 nt targeting sequence, it is obvious there will be complementary sites in the genome other than a particular site, the existence of which is 'insisted') Hence, ultimately it is important to note that the choice of either models lies in the choice of the definition of what a 'correct binding' constitutes- i.e. whether it is binding to any complementary sequence which happens to exist in the genome or one particular complementary sequence found in a fixed location in the genome.

Also comparing this with the next figure, it can be seen that the modified S_{real} here is essentially the sum of S_{ideal} and the unmodified S_{real}, dominating in the low and high l regime respectively. (Note that the visual difference between the two graphs is because in figure 4 the maximum x limit is one nt greater than that in figure 5, the figures are correspond to each other mathematically)

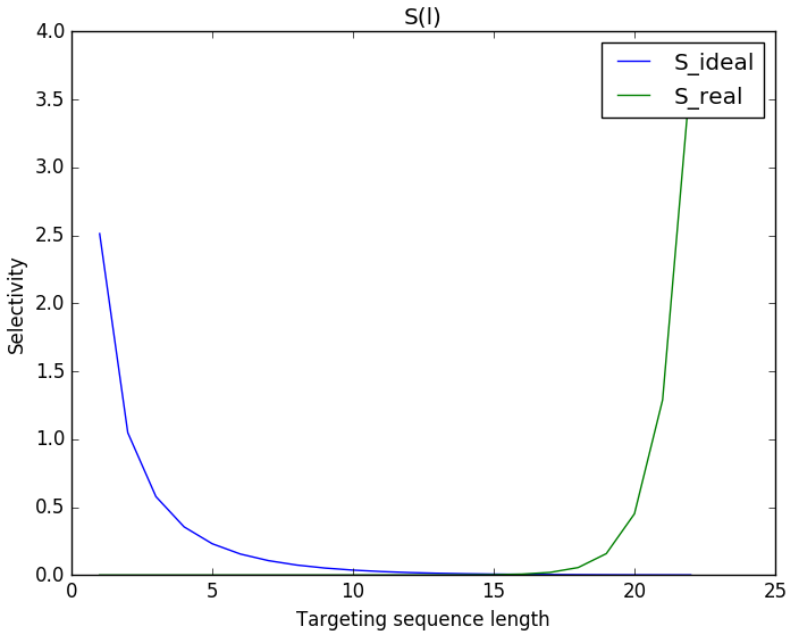


Figure 6: The blue line indicates S_{ideal} while the green line indicates S_{real} against the targeting sequence length l.

While noting that the variation of the real selectivity is a function of the genome length, it is perhaps interesting to note that with a value of 3×10^9 used for the genome length, an 'minimum acceptable' selectivity of 1.0-4.0 is found at around l=21 to 23. While noting that the number 3×10^9 is particular to the human genome, and hence the optimal targeting sequence length should depend on the length of the genome accessible to the targeting sequence, associated with the organism under question.

As a general comment, it seems clear that the imposition of whether or not a complementary sequence 'must' exist in the genome makes a profound and fundamental difference in the behaviour of the selectivity as a function of l.

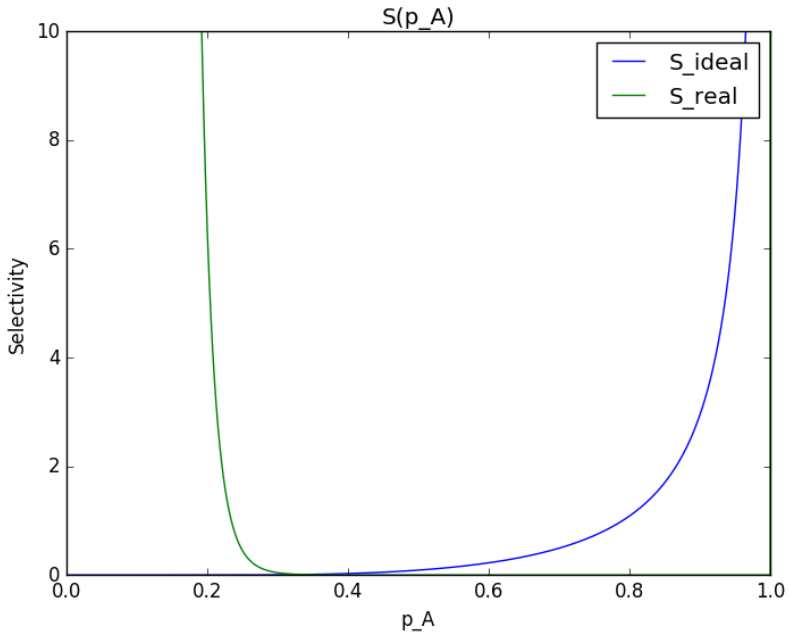


Figure 7: The blue line indicates S_{ideal} while the green line indicates S_{real} against the A nucleotide frequency in the genome. The targeting sequence in this case is a 20nt sequence consisting only of T.

As seen in the figure, the ideal selectivity exhibits a more expected behaviour of an asymptotic increase as p_A approaches 1. Indeed, the same also occurs for the real selectivity, but the asymptotic increase is heavily suppressed by the N_G factor appearing in the denominator. **Hence the expected asymptotic increase in S_{real} as $p_A \to 1$ does exist but just cannot be discerned visually on the graph**. (It is only around $p_A \approx 1 - 10^{-4}$ when the rapid asymptotic change becomes noticeable upon magnification)

The divergence of S_{real} can be explained by the fact that the S_{real} correction involves the insisting that a complementary sequence exists in the targeted genome. Due to the values used in the model, when p_A is small, the exponential energetic term dominates with an exponential argument $\beta l \epsilon \approx 40$. As such the partition function is heavily dominated by the energetic term associated with the complementary sequence. Clearly on the other hand for the case of S_{ideal}, this energetic dominance is suppressed by the probability term as $p_A \to 0$.

Interpreting the results of the model physically, the plausibility of the rapid increase of selectivity as $p_A \to 0$ at first seems questionable questionable. But it is important to keep in mind here that this apparently odd result is simply due to the very extreme and unlikely situation that for a genome of length N_G, a proportionately small region of l is filled with the complementary nucleotide while the rest of the genome is filled with non-complementary nucleotides. This, compounded with the nature of model 1, which assumes binding energy between non-complementary pairs to be zero causes the value of Z to be greatly dominated by the q_{comp} contribution, leading to a $\frac{Z}{q_{comp}}$ ratio of nearly 1, recalling that a l value of 20 is used.

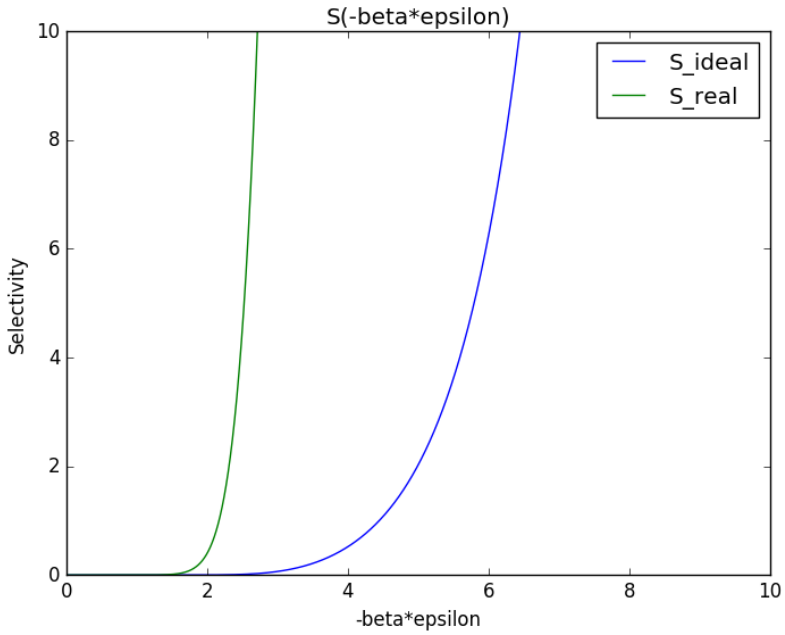


Figure 8: Ideal and real selectivities as a function of the exponential argument $(-\beta\epsilon)$ of the energy term. In this scenario again the targeting sequence is 20nt, consisting only of T. p_A is set as 0.25.

No surprises here, both S_{real} and S_{ideal} increase as the binding energy is increased. The increase in S_{ideal} is likely more prominent because of the higher probabilistic occurrence of the complementary sequence where it is imposed that one complementary sequence exists in the genome rather than the existence of one in several hundred 3×10^9bp genomes using the more naive mean field model.

For a comparison with the translation binding simulation constructed using model 3 considerations (Santa Lucia rules), S_{real} in model 1 ostensibly produces a different result than that of the translation binding simulation. Using model 1, a value of 19.1 was derived for S_{real}, while the value obtained from the translation binding simulation was 67300. This is likely due to the fact that model 1 treats all defect energies as being equal at '0', while from thermodynamic data provided by Santa Lucia et al (2004), it is notable that some single-defect dimer duplex energies (stacking energies) (those containing G) have free energies comparable to those of duplexes composed only of Watson Crick pairs. For example $\Delta G(GG/CG) = -1.11 kcal/mol$, while $\Delta G(TA/AT) = -0.58 kcal/mol$.

4 Model 2

I mainly focused and spent time on evaluating models 1 and 3 and building their associated simulations. For the time being I didn't think it was 'as' insightful to investigate model 2. Therefore so far I felt it was less of a priority.

Model 2 involves the incorporation of correlated probabilities of nucleotide occurrence and non-degenerate defect energies. First, these individual phenomena are modeled separately and then subsequently combined.

4.1 Non-identical defect energy

The non-identical energies simulation is carried out using a similar algorithm as model 1. The algorithm loops over all of the possible combinations of the number of A,T,C,G nucleotides and then the number of each of these nucleotides attached to A,T,C,G nucleotides on the targeting sequence (m_{AA},m_{TA},m_{CA},..., m_{AT},m_{TT}, ..., m_{AC}, ...).

The defect energy matrix is given by

$$E_D = \begin{pmatrix} E_{AA} & E_{AT} & E_{AC} & E_{AG} \\ E_{AT} & E_{TT} & E_{TC} & E_{TG} \\ E_{AC} & E_{TC} & E_{CC} & E_{CG} \\ E_{AG} & E_{TG} & E_{CG} & E_{GG} \end{pmatrix}$$

Note that this matrix is symmetric.

4.2 Nucleotide probability correlation

The correlation matrix is given by

$$M_c = \begin{pmatrix} P(A|A) & P(T|A) & P(C|A) & P(G|A) \\ P(A|T) & P(T|T) & P(C|T) & P(G|T) \\ P(A|C) & P(T|C) & P(C|C) & P(G|C) \\ P(A|G) & P(T|G) & P(C|G) & P(G|G) \end{pmatrix}$$

where $P(X_{n+1}|X_n)$ is defined as the probability that a subsequent nucleotide in a sequence is X_{n+1}, given that the previous nucleotide is X_n. The meaning of precedence in this case should be considered in terms of the direction towards a 3' or 5' end of the DNA strand.

4.3 Combined Model

Once separately created, the models can easily be combined to create a complete model that incorporates both nucleotide correlation and different defect energies.

The combined model is best realised using partition function factorisation, as it was found that combinatorial algorithms written, though operational, had excessive run times.

4.4 Simulations and Results

(+Carry out more simulations on this, see if it provides any insights)

5 Model 3

5.1 Nearest Neighbour Model

A more realistic description of DNA binding considers nearest neighbour interaction of different nucleotide pairs.

In such a model, the total binding energy is given by

$$E_{tot} = E_{init} + E_{sym} + \sum_{i=1}^{l-1} E_{dpx_i} \tag{12}$$

Here, the total free energy is given by a sum of the initiation energies, which correspond to the binding energies associated with the base pairs at the terminals of the oligonucleotide, the free energies of dimer duplexes in the bounded strands and a symmetry correction from self-complementary sequences. In the computational model developed, the symmetry correction was ignored for now.

This nearest neighbour interaction model is described in Santa Lucia et al (1998) and pioneered in papers by Zimm et al and Tinoco et al. The model is well established for complementary sequences, and cases of single and bubble defects. However the case of adjacent double defects has not been investigated. Therefore due to this lack of experimental data for dimer duplex energies consisting of adjacent defects (double defects), in the models and simulations constructed a first approximation was utilised to calculate the double defect dimer duplex energies. The energies were derived by subtracting average complementary pair energies (derived in turn from halving the average of the appropriate complementary dimer duplex energies) from single defect energies and subsequently averaging them, yielding a neighbour independent estimate for the binding energies for non-Watson Crick pairs. Subsequently, for double defects, the appropriate neighbour-independent estimates are simply summed.

(+Comment more on method of approximating the double defect duplex energies)

The following table outlines the energies of the 10 possible duplexes, and the initiation energies, taken from Santa Lucia et al (1998). The values quoted correspond to 'unified' energies which

Item	Quantity	eV
AA/TT	-1.00	-0.0434
AT/TA	-0.88	-0.0382
TA/AT	-0.58	-0.0252
CA/GT	-1.45	-0.0629
GT/CA	-1.44	-0.0624
CT/GA	-1.28	-0.0555
GA/CT	-1.30	-0.0564
CG/GC	-2.17	-0.0941
GC/CG	-2.24	-0.0971
GG/CC	-1.84	-0.0798
Init, GC	+0.98	+0.0425
Init, AT	+1.03	+0.0447

Table 2: These unified energy values are valid under a sodium concentration of 1.0M and with the rank of 12 for the stacking matrix.

are derived from data collated from 7 separate experiments at varying salt concentrations and conditions.

Algorithmically, this nearest neighbour model is easily realised by modifying the previous 'combined' model 2. The only change here then lies in the energetics, and the algorithmic approach towards computing nucleotide correlation remains the same.

Finally as a general prediction, it may be useful to note that by applying the Santa Lucia rules as shown in equation 4, the nearest neighbour model should theoretically give rise to increased favourability towards longer targeting sequences (as compared to the simple mean field models of model 1 and 2) due to the positive free energies associated with the initiation terms.

5.2 Chromosome 20 Simulation

"Translation binding model"- a **"chromosome walk"** model

5.2.1 Initial Test

To test model 3, an optimal value may be found for the optimal length l of the targeting sequence using an algorithm which simulates binding of a targeting sequence to the various positions in the genome by shifting along it as if it were a perfectly straight, non-conformed strand of DNA.

The partition functions and selectivities associated with particular targeting sequences can be derived by considering multiple allowed sites in the genome (e.g. satisfying the requirement of existence of the PAM sequence), from which an average can be derived. (+However this has not been done yet)

Again, the human chromosome 20 is used for this simulation due to a high percentage of sequenced base pairs.

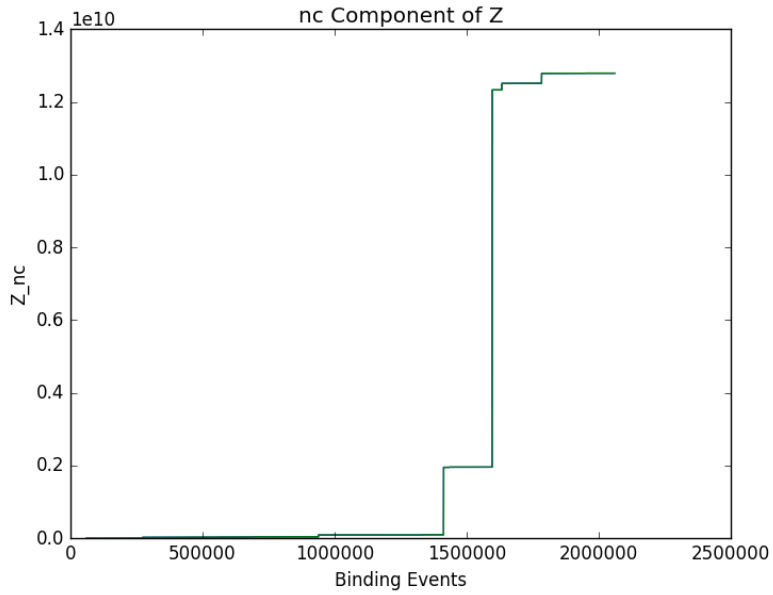


Figure 9: The Z_{nc} in the figure above refers to the non-complementary component of the partition function. The number of binding events refers to the position along the chromosome strand the targeting sequence has moved.

The figure above shows an initial simulation for the translational binding model conducted with chromosome 20 done to test the algorithm. It provides a general sense of the result to be expected from the simulation. As seen, as the targeting sequence shifts along the 'chromosome strand', a very dramatic increase in Z_{nc} occurs when a near complementary sequence is encountered.

The Z_{nc} in the figure above refers to the non-complementary component of the partition function associated with the binding between the targeting sequence and the genome. The definition and nature of Z_{nc} can be unambiguously understood using the expression $S = \frac{1}{\frac{Z}{q_{comp}}-1} = \frac{q_{comp}}{Z-q_{comp}} = \frac{q_{comp}}{Z_{nc}}$. Hence from this, Z_{nc} is defined as $Z - q_{comp}$.

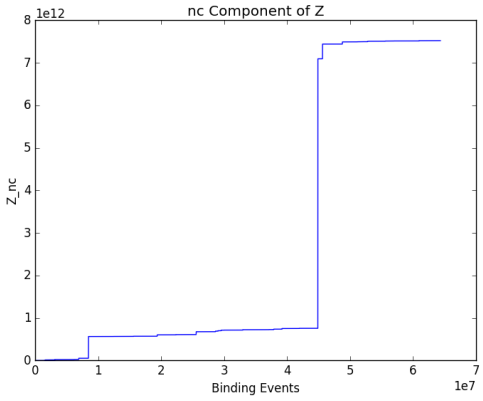
The complementary sequence to the targeting sequence is at the very beginning of the sequenced region of the 64Mbp chromosome. For this simulation, the first allowed 20nt sequence satisfying the condition of presence of a NCT PAM was chosen. This sequence happens to end at position 60045. The 'Binding Events' variable on the x axis simply refers to the number of nucleotides the targeting sequence has 'translated' over.

```
5'  ...TTATCTGTTCTGGTGTTCGT(CGA)...  3'
3'  ...AATAGACAAGACCACAAGCA(GCT)...  5'    Targeted Genome

    5' TTATCTGTTCTGGTGTTCGT 3'            Targeting Sequence
```

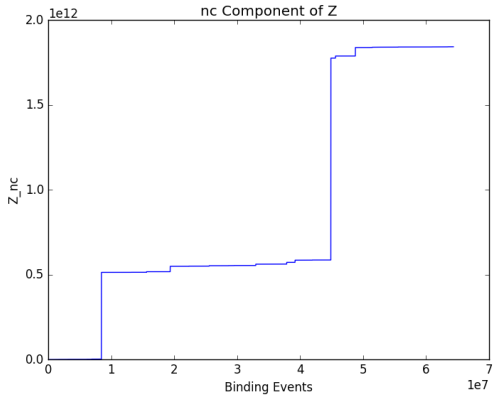
Figure 10: A visualisation of the binding of the user defined targeting sequence to the region 60023-60045 of chromosome 20. Notice the presence of the 5'-NGA-3' PAM sequence in the region 60043-60045.

```
5'  ...TTATCTGTTCTGGTGTTCGT(CGA)...  3'
3'  ...AATAGACAAGACCACAAGCA(GCT)...  5'    Targeted Genome

    5' UUAUCUGUUCUGGUGUUCGU 3'            Targeting Sequence
```

Figure 11: Evidently, a more realistic model would replace thymine by uracil in the RNA targeting sequence.



(a)



(b) The evolution of Z_{nc} over the positions on the genome walked by the targeting sequence. This time the initiation energies were not set to zero

Figure 12

Having successfully conducted the initial simulation, the binding simulation for the entire 64Mbp chromosome 20 was carried out. The figure above is the 'translation binding' simulation conducted over the entire chromosome 20. It can be seen that at around position 45Mbp a non-complementary sequence with a high number of Watson-Crick pairs is present.

(Note that since at this point insufficient information had been found about non WC-pair initiation energies, the figure above corresponds to a simulation where all initiation energies are set to zero. Therefore, the only contribution to the total binding energy comes from the duplex binding energies.)

Now, as a way of potentially extrapolating this result find a value for the binding selectivity for the entire genome, the possibility of extrapolating linearly was considered. However in this case, treating the Z_{nc} as a linear function over the number of binding events is not valid, as the latter simulation carried out over the entire chromosome has a 'linear gradient' 20 times that of the former. Clearly the effect of unpredictable occurrences of near-complementary sequences can cause significant sudden changes in the value Z_{nc}.

For absolute values of $e^{-\beta\Delta G_{comp}}$, Z and S , 5.0567163×10^{17}, 5.0567914×10^{17} and 67300 (3s.f.) were obtained respectively.

Following the translation binding model, the mean field model 3 was compared against it. Using the average probabilities and nucleotide correlation parameters given in the appendix, values for S_{real} was derived as 5020 (3s.f.). This shows that the mean field model 3 results to a much better match with the result from the translation binding model, differing by one order of magnitude. (Note that again, this mean field result was also achieved by setting the initiation energies to zero for the time being)

5.2.2 Results

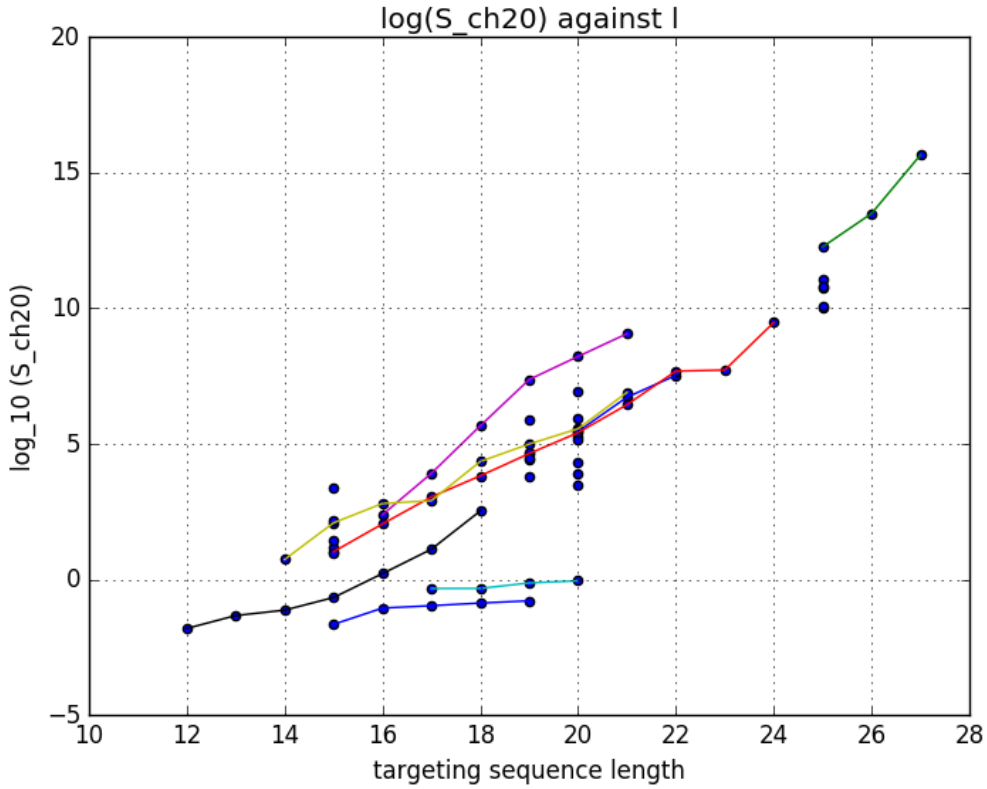


Figure 13: This figure shows the S_{ch20} selectivity of 69 targeting sequences tested on chromosome 20. The data points connected with lines correspond to sequences of different l that end on exactly the same nucleotide in the genome. *See appendix for the data represented in this graph and for more details.*

It is important to note here that discrepancies between the consecutive events is due to the differences in nucleotide composition of the targeting sequence. Notably it can be seen that such differences may result to thermodynamic selectivities up to 8 orders of magnitude apart while the sequences have the same l.

6 Zipper Model

To represent DNA binding/unbinding, a simple statistical physics model can be constructed, emulating the problem of molecular binding using a zipper model.

The basic governing equation of the zipper model is

$\frac{dP_n}{dt} = -P_n(K_b + K_f) + P_{n-1}K_f + P_{n+1}K_b$

$=K_f(P_{n-1} - P_n) + K_b(P_{n+1} - P_n)$

where K_f and K_b are the rate at which a bond forms and breaks respectively, and P_n is a 'population' which describes how many pairs in the duplex are binded. The rates K_f and K_b are related to, explicitly, the free energy associated with bond formation between nucleotides of the two strands.

$\frac{K_f}{K_b} = e^{-\beta \Delta G_f}$

where ΔG_f is the change in free energy during a bond formation process (as consistent with the definition of ΔG used in the report).

Therefore,

$\frac{dP_n}{dt} = K_b[e^{-\beta \Delta G_f}(P_{n-1} - P_n) + (P_{n+1} - P_n)]$.

6.1 Numerical Solution

Using this equation as the model, the system was first solved numerically. The figures below show the population time-evolution at 3 points (positions 1,5,10) of the system. The initial population vector is $\underline{P}$=(1,0,0,...,0), while the product of K_b and the number of x-points simulated is constant at 10.

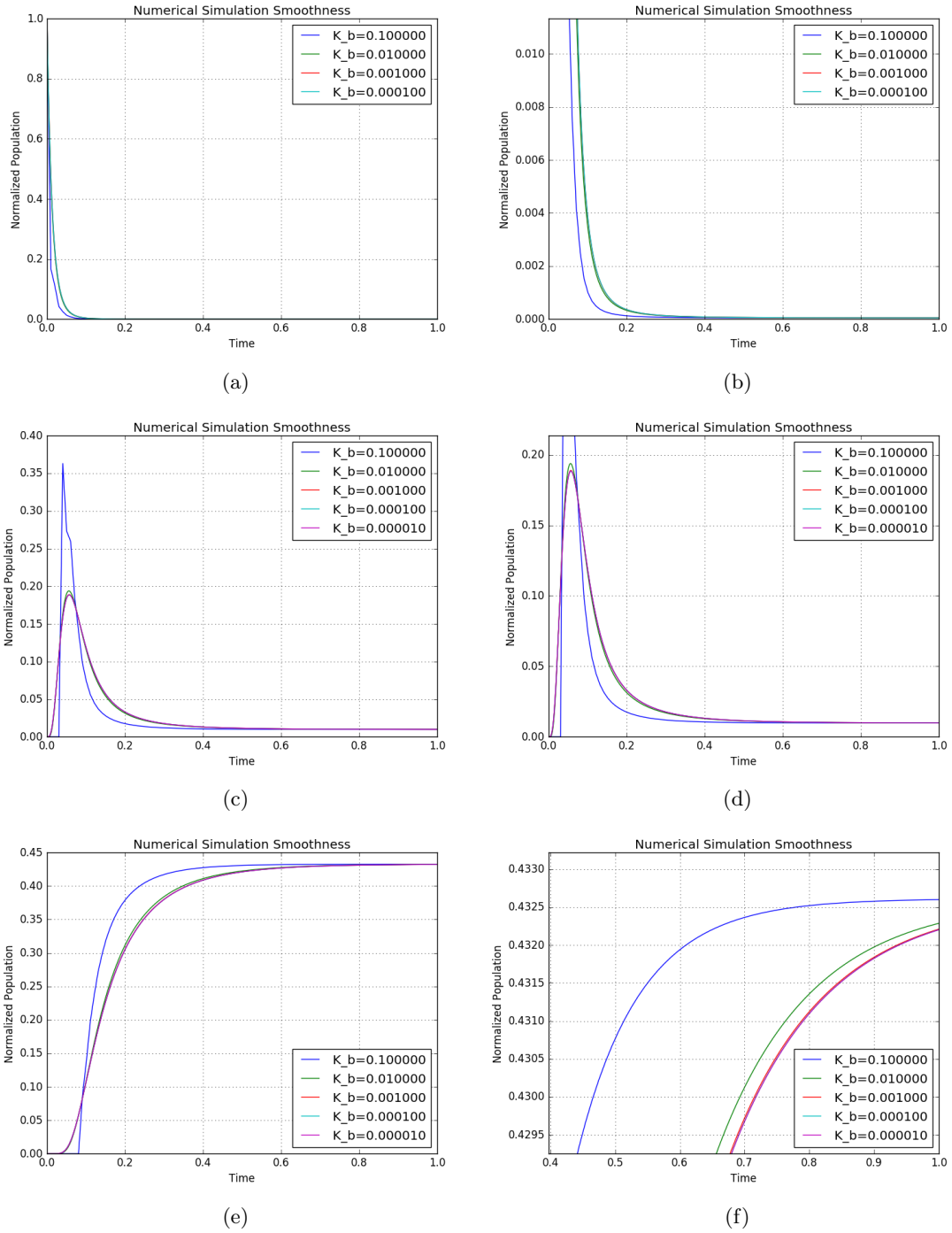
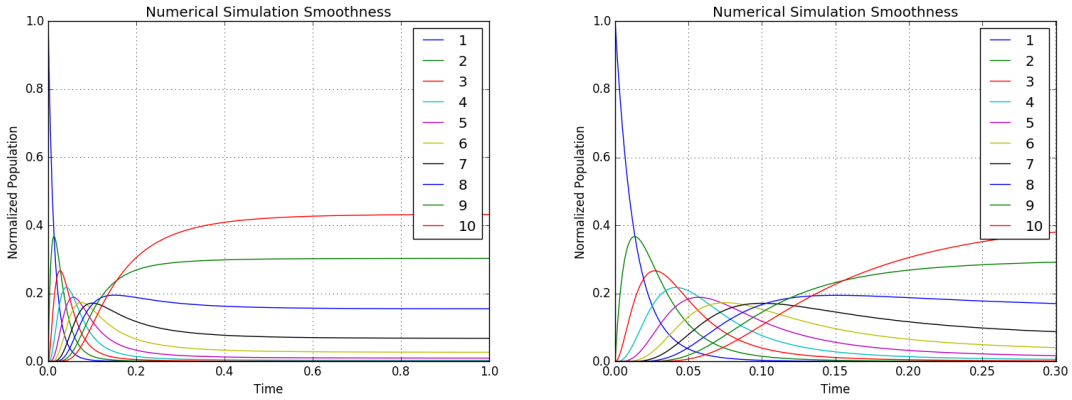


Figure 14

As seen from these simulations, a setting of $K_b = 10^{-3}$ and $N_x = 10^4$ can be seen as being sufficiently accurate for approximating the time-evolution in the analytic, continuous limit. These values can be referred to when constructing new numerical simulations.

Using the values $K_b = 10^{-3}$ and $N_x = 10^4$, the following figure then shows the time-evolution of $\underline{P}$.



(a) (b)

Figure 15: Numerical simulations of the time-evolution of the population vector with initial state $\underline{P}$=(1,0,0,...,0).

6.2 Analytic Solution

For a system of the form $\frac{\mathrm{d}\underline{x}}{\mathrm{d}t} = \underline{\underline{A}}\underline{x}$, the solution to $\underline{x}$ is given by $\underline{x} = e^{t\underline{\underline{A}}}\underline{x}_0$ [], where $\underline{x}_0$ is the initial state of the system. This form of solution can be motivated by applying Picard iteration $x_{n+1}(t) = x_0 + \int_0^t \underline{\underline{A}}x_n(s)ds$ (subscript indicates iteration number) on the initial condition by setting $x_0(t) = x_0$. It then follows that for the n^{th} iteration, the solution is $x_n(t) = x_0 + t\underline{\underline{A}}x_0 + \frac{t^2}{2}\underline{\underline{A}}^2x_0 + ... + \frac{t^n}{n!}\underline{\underline{A}}^n x_0$. It is hence this infinite sum that defines the matrix exponential $e^{t\underline{\underline{A}}}$. Subsequently through the Picard theorem, it can then be proved that the infinite sum $\sum_{n=0}^{\infty} \frac{t^n\underline{\underline{A}}^n}{n!}$ exists for all t.[]

In the zipper model under consideration, the matrix $\underline{\underline{A}}$ is given by

$$\underline{\underline{A}} = \begin{pmatrix} -(K_f+K_b) & K_b & 0 & \dots & 0 \\ K_f & -(K_f+K_b) & K_b & \dots & 0 \\ 0 & K_f & -(K_f+K_b) & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & -(K_f+K_b) \end{pmatrix}$$

equivalently,

$$\underline{\underline{A}} = K_b \begin{pmatrix} -(e^{-\beta\Delta G}+1) & 1 & 0 & \dots & 0 \\ e^{-\beta\Delta G} & -(e^{-\beta\Delta G}+1) & 1 & \dots & 0 \\ 0 & e^{-\beta\Delta G} & -(e^{-\beta\Delta G}+1) & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & -(e^{-\beta\Delta G}+1) \end{pmatrix}$$

Writing $\underline{\underline{A}}$ in the form $\underline{\underline{S}}\underline{\underline{\Lambda}}\underline{\underline{S}}^{-1}$, it is then the case that $\underline{\underline{A}}^k = \underline{\underline{S}}\underline{\underline{\Lambda}}^k\underline{\underline{S}}^{-1}$. Here $\underline{\underline{\Lambda}}$ and $\underline{\underline{S}}$ have their usual definitions as the diagonal matrix of eigenvalues and the matrix of eigenvectors respectively. (Due to the non-trivial nature of the eigenvalues and eigenvectors, $\underline{\underline{\Lambda}}$ and $\underline{\underline{S}}$ are obtained computationally.)

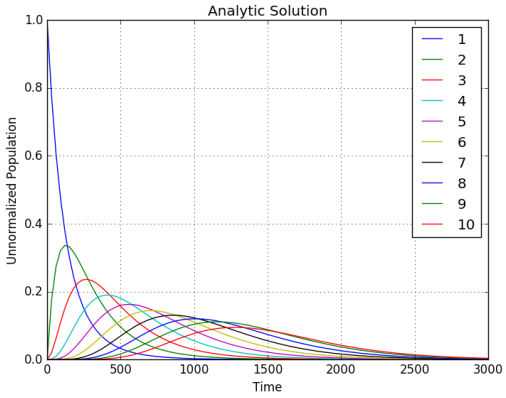
With $\sum_{k>0}(\frac{1}{k!}\underline{\underline{A}}^k) = \sum_{k>0}(\frac{1}{k!}\underline{\underline{S}}\underline{\underline{\Lambda}}^k\underline{\underline{S}}^{-1}) = \underline{\underline{S}}(\sum_{k>0}\frac{1}{k!}\underline{\underline{\Lambda}}^k)\underline{\underline{S}}^{-1}$, it then follows that $e^{\underline{\underline{A}}} = \underline{\underline{S}}e^{\underline{\underline{\Lambda}}}\underline{\underline{S}}^{-1}$ and $e^{t\underline{\underline{A}}} = \underline{\underline{S}}e^{t\underline{\underline{\Lambda}}}\underline{\underline{S}}^{-1}$, yielding the solution $\underline{P} = \underline{\underline{S}}e^{t\underline{\underline{\Lambda}}}\underline{\underline{S}}^{-1}\underline{P}_0$.

Thus explicitly, after solving for the eigenvalues and eigenvectors which are expressed in $\underline{\underline{\Lambda}}$ and $\underline{\underline{S}}$ respectively,

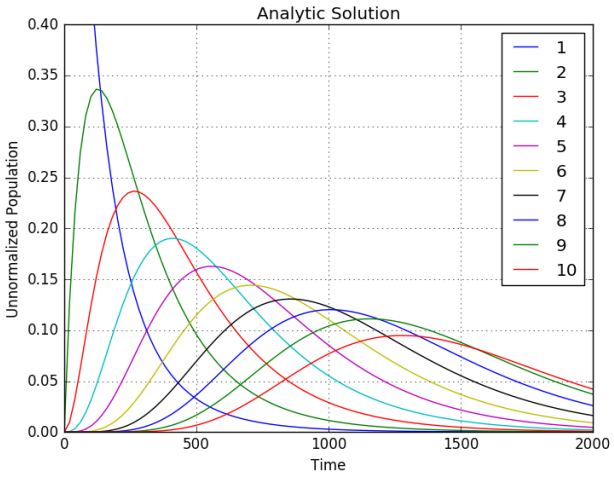
$$\begin{pmatrix} P_1 \\ P_2 \\ P_3 \\ \vdots \\ P_l \end{pmatrix} = \underline{\underline{S}} \begin{pmatrix} e^{\lambda_1 t} & 0 & 0 & \dots & 0 \\ 0 & e^{\lambda_2 t} & 0 & \dots & 0 \\ 0 & 0 & e^{\lambda_3 t} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & e^{\lambda_l t} \end{pmatrix} \underline{\underline{S}}^{-1} \begin{pmatrix} P_1(0) \\ P_2(0) \\ P_3(0) \\ \vdots \\ P_l(0) \end{pmatrix}$$

$$\begin{pmatrix} P_1 \\ P_2 \\ P_3 \\ \vdots \\ P_l \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^{l} S_{1i}S_{i1}^{-1}e^{\lambda_i t} & \sum_{i=0}^{l} S_{1i}S_{i2}^{-1}e^{\lambda_i t} & \sum_{i=0}^{l} S_{1i}S_{i3}^{-1}e^{\lambda_i t} & \dots & \sum_{i=0}^{l} S_{1i}S_{il}^{-1}e^{\lambda_i t} \\ \sum_{i=0}^{l} S_{2i}S_{i1}^{-1}e^{\lambda_i t} & \sum_{i=0}^{l} S_{2i}S_{i2}^{-1}e^{\lambda_i t} & \sum_{i=0}^{l} S_{2i}S_{i3}^{-1}e^{\lambda_i t} & \dots & \sum_{i=0}^{l} S_{2i}S_{il}^{-1}e^{\lambda_i t} \\ \sum_{i=0}^{l} S_{3i}S_{i1}^{-1}e^{\lambda_i t} & \sum_{i=0}^{l} S_{3i}S_{i2}^{-1}e^{\lambda_i t} & \sum_{i=0}^{l} S_{3i}S_{i3}^{-1}e^{\lambda_i t} & \dots & \sum_{i=0}^{l} S_{3i}S_{il}^{-1}e^{\lambda_i t} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=0}^{l} S_{li}S_{i1}^{-1}e^{\lambda_i t} & \sum_{i=0}^{l} S_{li}S_{i2}^{-1}e^{\lambda_i t} & \sum_{i=0}^{l} S_{li}S_{i3}^{-1}e^{\lambda_i t} & \dots & \sum_{i=0}^{l} S_{li}S_{il}^{-1}e^{\lambda_i t} \end{pmatrix} \begin{pmatrix} P_1(0) \\ P_2(0) \\ P_3(0) \\ \vdots \\ P_l(0) \end{pmatrix}$$

With K_b=1 and K_f=7.5 set for this model, the following results were obtained. In this simulation the hybridizing strands are fully complementary, with ΔG^{37}=-0.054eV. The initial condition corresponds to the case where only the first base pair is hybridized.



(a)



(b)

Figure 16

It is important to note in the figures above the population is unnormalised, and because $\sum_{i=1}^{l} P_i$ decays, the rate at which the populations change in the model correspondingly decrease as time progresses, resulting to the large amount of over which the time-evolution happens.

```
In [170]: %run "C:
\Users\Choon\Desktop\CRISPR\Code\Algorithms\zipper\P_
evo2.py"
[  5.04469213e-05   2.64932529e-04   1.01311516e-03
3.33476168e-03
   9.92172085e-03   2.71261015e-02   6.81592961e-02
1.54864439e-01
   3.03063052e-01   4.32202134e-01]
```

(a)

```
In [169]: P_tx[-1]/sum(P_tx[-1])
Out[169]:
array([  4.99255617e-05,   2.62376773e-04,   1.00444254e-03,
         3.31088080e-03,   9.86654969e-03,   2.70206591e-02,
         6.80041538e-02,   1.54731329e-01,   3.03137600e-01,
         4.32612083e-01])

In [170]:
```

(b)

Figure 17

The values above correspond to the steady state vector yielded from the numerical and analytical solution respectively. As seen, the steady state $\underline{P}$, $\underline{P}(t \to \infty)$ derived from the two methods are for all essential purposes equivalent.

From these results, a possible definition of the zipping or unzipping time can be the time at which P_1 or P_l becomes the largest component of in $\underline{P}$.

7 Remarks

7.1 Other Factors Affecting Binding Selectivity

For the CRISPR CAS9 system in the bacteria *streptococcus pyogenes* (Is the PAM sequence in humans NGA?), the presence of a protospacer adjacent motif (PAM) consisting of a NGG sequence near the 3' end of the targeted sequence is required. This condition necessarily increases binding selectivity for the targeting RNA.

The effect of the requirement of having the PAM sequence for binding to occur limits the number of possible sites that the targeting sequence can bind to the targeted genome. Ultimately this leads to the selectivity to differ by a factor of F, which is the ratio of the total number of allowed binding sites with the PAM sequence present and the number of all binding sites in the genome.

Thermodynamic differences between N-U and N-T binding have also not been incorporated into the models yet as well. .

An important point to note about the simulations and carried out above is that that those utilising the Santa Lucia rules (and correspondingly the associated thermodynamic database) are valid under the condition that the sodium concentration of the environment is 1.0M. As seen in an array of experimental data quoted Santa Lucia et al 1998, varying salt concentrations can lead to significant differences in the binding energy between sequences. Therefore differing selectivities will be valid for targeting under physiological concentrations.

So far, in the models developed it should be noted that the 'genome' is considered as a strand of DNA, the regions of which are all equally accessible to the targeting sequence. In reality due to bending and conformations of chromosomes around histones, some regions will be more accessible to others.

In 2016, Kleinstiver et al published results detailing that high fidelity CAS9 (spCas9-HFI*) provides no detectable genome wide off target effects [**?**]. As such it provides a more precise alternative in comparison to wild type Cas9.

**Streptococcus pyogenes Cas9*

There should not be a first order effect relating the entropy of a sequence to its selectivity. There may be some subtle effects occurring in the chromosome-walk algorithm, but there should be no correlation between sequence entropy and selectivity in model 1.

Again, note that the symmetry/self-complementarity contribution associated with the Santalucia rules has been ignored.

Very crucially as well, the models and simulations constructed so far involve DNA-DNA binding. Therefore it is still an approximation to the actual CRISPR CAS9 system where DNA-RNA binding is involved. A few differences among others between DNA-DNA binding and DNA-RNA binding include different initiation energies and dimer hybridisation energies. []

7.2 Provisional Conclusion

From the investigations conducted so far, there has been no direct indication of any non-monotonic behaviour of selectivity as a function of l (though investigations for model 3 are still yet to be concluded). It suggests that the optimality of a 20nt targeting sequence in nature may be due to other energetic penalties (e.g. DNA conformation, twisting and looping, etc) associated with binding between very long sequences, penalties which are not considered in any of the models developed.

Therefore it is a provisional conclusion from these investigations that the insight offered by the models 1-3 developed lies in the fact that some models appear to suggest that the selectivity starts to rise significantly around the region of $l \approx 20$, when expressed as a function of N_G. It should be said that the models have the underlying assumption that the binding energy can rise infinitely with l, and once combined with additional energetic penalties, optimality may emerge. In addition it seems that particularly when using the real partition function modification in model 1 it seems possible that an expression for the optimal length as a function of the length of the overall genome accessible to the targeting sequence in a cell.

Another comment worth making on the minimum l value necessary for a 'functional' selectivity- is that a rapid rise in selectivity is also encouraged by the fact that at around l=15, it was seen that there were multiple re-occurrences of the complementary sequences within a single chromosome, leading to low selectivity. But when l is increased to slightly larger values, one can expect a rapid increase in selectivity due to the disappearance of these complementary sequences. Therefore as another rule of thumb it can be considered that a l value associated with a rapid rise in selectivity can be approximated by the equation $l = \log_4 N_G$. Indeed as clearly it becomes increasingly likely for complementary sequences to re-occur in the genome below this l, the value can also be considered to be a fundamental threshold as it becomes impossible for the targeting sequence to target one specific site in the genome.

Commenting on the thermodynamic models developed, the results give a good justification for

some of the observations of targeting selectivity involving truncated sgRNAs. In that it can clearly be seen from the 'walking' of targeting sequences over chromosome 20 that at around 16-18, there is a rapid exponential rise in selectivity at equilibrium. As such the models developed can be considered as a useful reference to consider the lower bound of l to produce a sufficient targeting selectivity to achieve practical gene editing purposes.

Generally more can still be done to investigate the formulation of an effective ΔG which combines binding energetics AND statistics.

8 Appendix

8.1 Real Partition Function Corrections

For a mean field model, a naive 'ideal' selectivity S_{ideal} can be written as $S_{ideal} = \frac{q_{comp}}{Z_{ideal} - q_{comp}}$, which essentially is a ratio of $\frac{complementary\_events}{noncomplementary\_events}$. Thus using this notion, several S_{real} can be defined, manifesting as corrections to S_{ideal}, with the more realistic consideration that a sequence complementary to the targeting sequence **must** exist in the genome. Because differences in what constitutes a 'correct' binding and conditions imposed on the genome are allowed, there can be several definitions of S_{real}

1. **At least** one complementary sequence exists in the genome, there is a **unique** 'correct' binding site:
   $$S_{real} = \frac{1}{\frac{\frac{1}{N_G-l+1}e^{-\beta\Delta G_c} + \frac{N_G-l}{N_G-l+1-l}Z_{ideal}}{\frac{1}{N_G-l+1}e^{-\beta\Delta G_c}} - 1}$$
   $$S_{real} = \frac{e^{-\beta\Delta G_c}}{(N_G-l)Z_{ideal}}$$
   $$S_{real} = \frac{Z_{ideal} - p_c e^{-\beta\Delta G_c}}{p_c(N_G-l)Z_{ideal}} S_{ideal}$$
2. **At least** one complementary sequence exists in the genome, **any** site containing a complementary sequence is 'correct':
   $$S_{real} = \frac{1}{\frac{\frac{1}{N_G-l+1}e^{-\beta\Delta G_c} + \frac{N_G-l}{N_G-l+1-l}Z_{ideal}}{\frac{1}{N_G-l+1}e^{-\beta\Delta G_c} + \frac{N_G-l}{N_G-l+1}p_c e^{-\beta\Delta G_c}} - 1}$$
   $$S_{real} = \frac{\frac{1}{N_G-l}e^{-\beta\Delta G_c} + p_c e^{-\beta\Delta G_c}}{Z_{ideal} - p_c e^{-\beta\Delta G_c}}$$
   $$S_{real} = (\frac{1}{p_c(N_G-l)} + 1) S_{ideal}$$
3. **Exactly** one complementary sequence exists in the genome, (it must follow that) there is a **unique** 'correct' binding site:
   $$S_{real} = \frac{1}{\frac{\frac{1}{N_G-l+1}e^{-\beta\Delta G_c} + \frac{N_G-l}{N_G-l+1-l}\frac{1}{1-p_c}(Z_{ideal} - p_c e^{-\beta\Delta G_c}}{\frac{1}{N_G-l+1}e^{-\beta\Delta G_c}} - 1}$$
   $$S_{real} = \frac{e^{-\beta\Delta G_c}}{\frac{N_G-l}{1-p_c}(Z_{ideal} - p_c e^{-\beta\Delta G_c})}$$
   $$S_{real} = \frac{1-p_c}{p_c(N_G-l)} S_{ideal}$$

8.2 ϵ for Model 1

For model 1, ϵ was ostensibly set at -0.054eV. This sub-section discusses how this value was derived.

From Santalucia et al 1998, an average value for the unified duplex energies found in oligonucleotides was given as -1.42kcal/mol (an average over all of the 10 possible Watson-Crick duplexes). The initiation energies of an AT and CG pairs were given as +1.03kcal/mol and +0.98kcal/mol respectively. Ignoring the energetic contribution from symmetric/self-complementary sequences, an average ϵ for 20nt sequences can be found as $\frac{1}{20}[(19 \times -1.42) + 1.03 + 0.98]$ kcal/mol, which corresponds to 0.054eV.

8.3 Miscellaneous Results

This sub-section displays the results of some miscellaneous simulations carried out on the side of primary simulations discussed in previous sections.

8.3.1 Chromosome 1 Statistical Parameters

To test an algorithm to extract mean field statistical parameters for complete chromosomes, parts of the human chromosome 1 was analysed.

For the figures to follow, the format of the matrices containing the values of the statistical parameters is in the convention shown below.

$M_p = \begin{pmatrix} P_G^A & P_G^T & P_G^C & P_G^G \end{pmatrix}$

This is a matrix containing the values of the genome nucleotide frequencies.

The correlation matrix is in the form of:

$$M_c = \begin{pmatrix} P(A|A) & P(T|A) & P(C|A) & P(G|A) \\ P(A|T) & P(T|T) & P(C|T) & P(G|T) \\ P(A|C) & P(T|C) & P(C|C) & P(G|C) \\ P(A|G) & P(T|G) & P(C|G) & P(G|G) \end{pmatrix}$$

where $P(X_{n+1}|X_n)$ is the probability that probability for a particular nucleotide to be at position $n+1$ of the genome given that identity of the previous nucleotide at position n.

The results shown in the figure below shows parameters corresponding to the first 6Mbp of the human chromosome 1, divided into 3 equal and chronological intervals (i.e. 0-2Mbp, 2-4Mbp, 4-6Mbp). Unsequenced 'N' nucleotides are neglected from the computation. Due to minor initial difficulties in handling large data files, not all of the 249Mbp of the chromosome.

```
In [72]: runfile('C:/Users/Choon/Desktop/CRISPR/Code/Algorithms/
nt_data_param_extraction.py', wdir='C:/Users/Choon/Desktop/CRISPR/Code/Algorithms')
[0.24857730900940708, 0.23166914764627589, 0.26138383770860746, 0.2583697056357096]
[[0.28664223777936165, 0.1977821602074415, 0.21684062774661006, 0.2987349742665868],
[0.16919128531514138, 0.26106284998439955, 0.2520239567033962, 0.3177219079970629],
[0.3004682438361063, 0.27796675718271907, 0.31623716591883494, 0.10532783306233964],
[0.2306420501515569, 0.19107738972946936, 0.25714189554458466, 0.3211386645743891]]

In [73]: runfile('C:/Users/Choon/Desktop/CRISPR/Code/Algorithms/
nt_data_param_extraction.py', wdir='C:/Users/Choon/Desktop/CRISPR/Code/Algorithms')
[0.21111224693200578, 0.21488839707844148, 0.28732168321841667, 0.28667767277113604]
[[0.21972017170291425, 0.1759954871089655, 0.24062860892070725, 0.363655732267413],
[0.11182202855668166, 0.23073922464542002, 0.2770112290599795, 0.38042751773791883],
[0.2832029686502141, 0.2669276925184251, 0.34829339969865214, 0.1015759391327087],
[0.20694727233179172, 0.17949051672187732, 0.26832738436961784, 0.3452348265767131]]

In [74]: runfile('C:/Users/Choon/Desktop/CRISPR/Code/Algorithms/
nt_data_param_extraction.py', wdir='C:/Users/Choon/Desktop/CRISPR/Code/Algorithms')
[0.26576763288381644, 0.25856212928106465, 0.2396126198063099, 0.236057618028809]
[[0.28805628980217673, 0.22926430056346242, 0.19737928828769508,
0.28530012134666577], [0.17101314191567207, 0.2837481919230204, 0.2448348945320658,
0.30040377162924176], [0.3387072878084407, 0.3117032709061506, 0.2930272836350357,
0.056562157650373], [0.2704235196932951, 0.2100187454327865, 0.22722218103640002,
0.2923355538375184]]
```

Figure 18: Chromosome 1 statistical parameters for different regions

As can be seen, there is notable variation of the nucleotide frequencies and correlations over the three intervals. With the length of chromosome 1 at about 250Mbp, the results suggest that with the chromosome split into about 100 equal segments of about 2Mbp in length, the chromosome can by no means be approximated as having homogeneous parameter values. This result notably suggests that over relatively large chromosome segments of the order of Mbp, the statistical parameters of the segments do not converge to averages, but are instead notably varying over the different segments, as affected by the different genetic functions and gene compositions of the chromosome segments.

8.3.2 Chromosome 20 Statistical Parameters

```
In [82]: runfile('C:/Users/Choon/Desktop/CRISPR/Code/Algorithms/
nt_data_param_extraction.py', wdir='C:/Users/Choon/Desktop/CRISPR/Code/Algorithms')
[0.27941908841008195, 0.2825336761673531, 0.21762912969651052, 0.2204181057260545]
[[0.31483548892924956, 0.24448373801023404, 0.1804939179951752,
0.26018685506534117], [0.19529943620157347, 0.32078468820493483, 0.2151456301221506,
0.2687702454713411], [0.3425408413430566, 0.3332099271728796, 0.2686678106168629,
0.055581420867200866], [0.2800238221292369, 0.23170341136920625, 0.2174956284235974,
0.27077713807795945]]
```

Figure 19: Matrices containing statistical parameter values for the entire chromosome 20, 64.4Mbp in total length.

```
In [83]: runfile('C:/Users/Choon/Desktop/CRISPR/Code/Algorithms/
nt_data_param_extraction.py', wdir='C:/Users/Choon/Desktop/CRISPR/Code/Algorithms')
[0.2897117244188874, 0.29333556385209447, 0.20799070353872165, 0.2089620081902965]
[[0.3253080289189719, 0.2575684395847331, 0.17183292779432432, 0.24529060370197067],
[0.21327047468203053, 0.32904643709439957, 0.2054882618378183, 0.2521948263857516],
[0.35037938627313187, 0.34351051768564067, 0.25984727102365895,
0.04626282501756852], [0.287280733660202, 0.2428524050888854, 0.2100181788541092,
0.25984868239680337]]

In [84]: runfile('C:/Users/Choon/Desktop/CRISPR/Code/Algorithms/
nt_data_param_extraction.py', wdir='C:/Users/Choon/Desktop/CRISPR/Code/Algorithms')
[0.2886027996951159, 0.2915946146245059, 0.20973079639799788, 0.2100717892823803]
[[0.3226431970079734, 0.2554231423430534, 0.17510544382920049, 0.2468282168197727],
[0.2085618710943258, 0.32710066269077037, 0.20706838225864044, 0.25726908395626336],
[0.3556770994216082, 0.3399242005264161, 0.2595171991558106, 0.04488150089616508],
[0.2859738216414658, 0.2437523545010674, 0.21129015578487062, 0.2589836680725962]]

In [85]: runfile('C:/Users/Choon/Desktop/CRISPR/Code/Algorithms/
nt_data_param_extraction.py', wdir='C:/Users/Choon/Desktop/CRISPR/Code/Algorithms')
[0.27914956527042517, 0.2830192600158269, 0.2128182424235331, 0.22501293229021485]
[[0.31823356078210124, 0.23319084390119654, 0.1812423363322952,
0.26733325898440696], [0.17640427817235235, 0.33469119749552834,
0.21516951827518244, 0.2737350060569369], [0.3447515225977487, 0.3389492083314783,
0.25273981545879587, 0.06355945361197711], [0.297843046295048, 0.2269456265929813,
0.2112782808912309, 0.2639330462207398]]
```

Figure 20: Chromosome 20 statistical parameters for the regions: 0-12.9Mbp, 12.9-25.8Mbp and 25.8-38.6Mbp

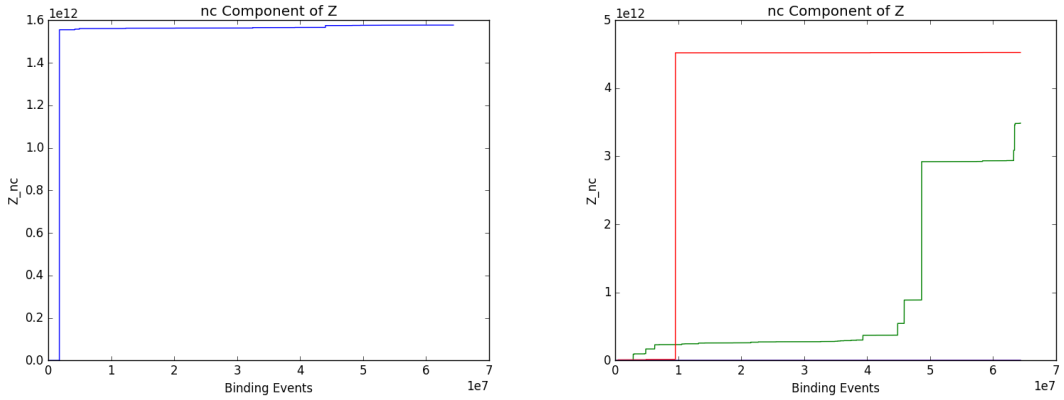
```
In [86]: runfile('C:/Users/Choon/Desktop/CRISPR/Code/Algorithms/
nt_data_param_extraction.py', wdir='C:/Users/Choon/Desktop/CRISPR/Code/Algorithms')
[0.2737806653077641, 0.2741630894094981, 0.2265305343416688, 0.225525710941069]
[[0.30584096808099726, 0.23698923658569002, 0.1859799876045459,
0.27118980772876683], [0.1920712398143332, 0.3062426810638995, 0.2245949080297325,
0.2770911710920348], [0.33614833507665476, 0.32985867793900364, 0.2808922380753731,
0.05310074890896848], [0.27154570221239227, 0.2243492159491503, 0.22350669800049058,
0.2805983838379668]]

In [87]: runfile('C:/Users/Choon/Desktop/CRISPR/Code/Algorithms/
nt_data_param_extraction.py', wdir='C:/Users/Choon/Desktop/CRISPR/Code/Algorithms')
[0.2657759490595191, 0.27051660030041863, 0.23102145070344832, 0.23268599993661393]
[[0.3006618659620582, 0.23762869493246636, 0.18940042817095157,
0.27230901093452387], [0.18404767063545902, 0.305490170780821, 0.22475597264243813,
0.28570618594128183], [0.3277488801511017, 0.31587055549383103, 0.2873496367435092,
0.06903092761155809], [0.2594153066330896, 0.222392394573856l5, 0.22991999881619235,
0.28827229997686193]]
```

Figure 21: Chromosome 20 statistical parameters for the regions: 38.6-51.5Mbp and 51.5-64.4Mbp

Note that for the data used, there are 499910bps which are not sequenced out of a total chromosome length of 64444167bps. I.e. 499910 'N's were counted in the data file.

An important observation to note (which indeed serves to verify the algorithm) is that the phenomenon CG suppression is well represented in the results. As seen in the cases of both chromosomes 1 and 20, the frequency of the CG dinucleotide is about $\frac{1}{5}$ that of the other dinucleotides.

8.3.3 Model 3: More results from chromo-walk simulations



(a) Z_{nc} is the non-complementary component of the partition function. The number of binding eventsrefers to the position along the chromosome strand the targeting sequence has moved.

(b) Z_{nc} is the non-complementary component of the partition function. The number of binding events refers to the position along the chromosome strand the targeting sequence has moved.

Figure 22

Chromo-walk simulation carried out for the entire chromosome 20 with a randomly selected targeting sequence of AAAGGAGAAGGGAAGTCTTT. **Here, the initiation energies for non-WC pairs are set to +1.005kcal/mol**. The PAM chosen is NCT (equivalently NGA). Assuming that this strict PAM requirement is valid, it was found that there are 4636926 available complementary sequences in the sequenced portion of chromosome 20. At the end of chromosome 20, S=4940, $Z = 7.78484 \times 10^{15}$, $q_{comp} = 7.78326 \times 10^{15}$. Once again, it is important to note that the sequencing of chromosome 20 is incomplete, while if the simulation were carried out over the entire genome, the selectivity would notably decrease.

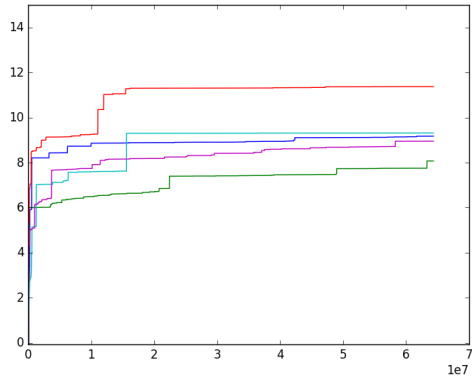
Targeting seq. (l)	Z_{ch20}	S_{ch20}
TTTATAATGTATAAAAACTA (20)	6.21333×10^{11}	2898
GGTGGGTGGACAACCTGAGA (20)	$2.96949442 \times 10^{18}$	851471
TCTGGGTCTTGACATCTAGG (20)	3.82331×10^{16}	8445
CAGAAAATTAGTATGGGAAA (20)	1.4557813×10^{14}	341724
CACTAACTCAGGGGAATGAT (20)	$7.53451162 \times 10^{15}$	8443608
TCTATGACGTACGAATCTG (19)	$1.12671881 \times 10^{15}$	750146
TTTAGAATAAGTAATAACA (19)	7.548886×10^{11}	6363
GTTAGGGGACAAAGGAGAC (19)	$1.22648875 \times 10^{16}$	51148
GTATCACTGACTATATGGT (19)	5.5842427×10^{13}	26930
TGTTTTCTGTTATTATCGA (19)	2.6029006×10^{13}	29155
TGGGTTACGACACCT (15)	2.077×10^{13}	14.50
TCAACACGCCAACGA (15)	7.411×10^{13}	144.4
CCACACTTCACGACT (15)	9.663×10^{12}	9.785
GAACGTACTTTGGTC (15)	1.1923×10^{12}	28.11
CCATTCTGCGAACTT (15)	1.619981×10^{12}	2302
TCACGATTTCATTTCAACAGTATTG (25)	$8.2611793685997 \times 10^{18}$	1.201×10^{11}
TTGAAACGGTAGTAAGTCGTGAAAG (25)	$1.3112403059344 \times 10^{21}$	5.828×10^{10}
CAAGAAGAAAGGATTTAATCACTAG (25)	$1.45323317407 \times 10^{18}$	1.104×10^{10}
GAGAGGACTCGATGTCTTCCTTACC (25)	$3.309194806425 \times 10^{21}$	6.161×10^{10}
GACGAGGGGAACGGGTGACCCCTGA (25)	$1.996684335525 \times 10^{25}$	1.139×10^{10}
GGTCGTCGTGTAGTTGTAGA (20)	4.9752800×10^{17}	191105
ATAGAGGTTTTATTGACTGA (20)	1827521×10^{14}	20935
TGAAAGAATGGGGACAACAG (20)	$2.27344769 \times 10^{16}$	136046

[H]

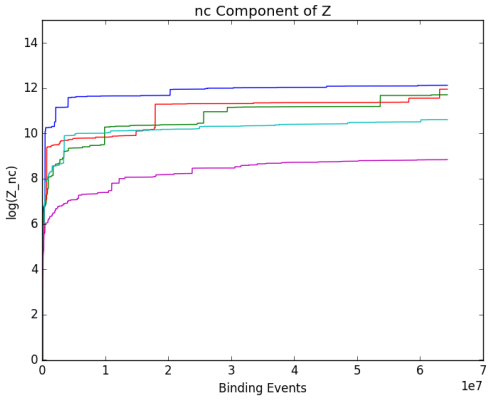
Table 3

Note that the results presented in the table above are computed without neglecting the contribution of binding to sites in the genome that do not satisfy the PAM requirement. That is to say the PAM requirement has not been strictly enforced in the calculation.

A chromo-walk simulation involving 5 randomly selected targeting sequences from chromosome 20. As seen, the variation is substantial, necessitating logarithmic representation.



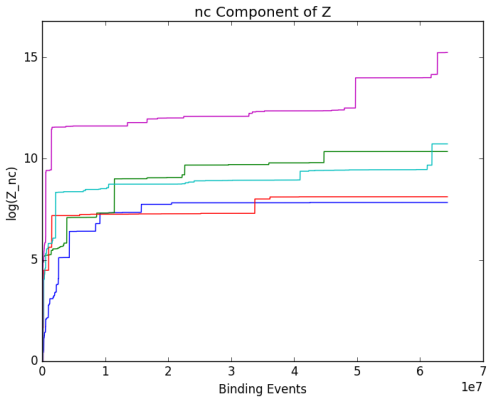
(a) Z_{nc} is the non-complementary component of the partition function. The number of binding events refers to the position along the chromosome strand the targeting sequence has moved.

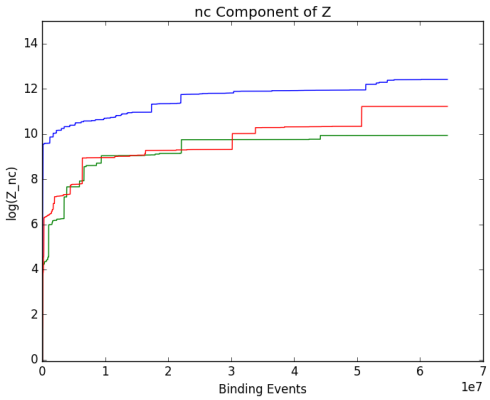


(b) Z_{nc} is the non-complementary component of the partition function. The number of binding events refers to the position along the chromosome strand the targeting sequence has moved. Chromosome walk simulation for 5 random $l = 15$ targeting sequences

Figure 23

From the table in this section it is evident that targeting sequences with high CG composition have high values of q_{comp} due to the high binding energy of CG pairs. It is important to note that some of the $l = 20$ sequences tested have very low selectivity because of an anomalous small CG composition in those sequences.





(a) Z_{nc} is the non-complementary component of the partition function. The number of binding events refers to the position along the chromosome strand the targeting sequence has moved. Chromosome walk simulation for 5 random $l = 25$ targeting sequences

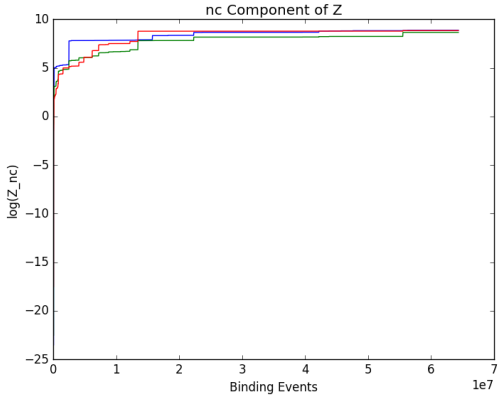
(b) Z_{nc} is the non-complementary component of the partition function. The number of binding events refers to the position along the chromosome strand the targeting sequence has moved. Chromosome walk simulation for 3 random $l = 20$ targeting sequences

Figure 24

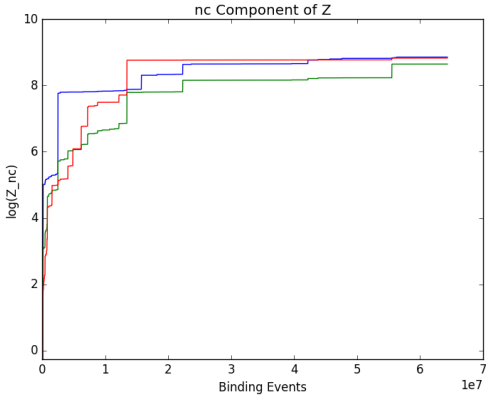
For the next few simulations, the portion of the genome complementary to the targeting sequence was chosen to be at the same location in the genome (i.e. they all border a particular PAM sequence in the genome). (With the exception for the first one)

Targeting seq. (l)	Z_{ch20}	S_{ch20}
AATTACTATAAAACTGGTGG (20)	$2.08098868 \times 10^{14}$	289784
CAATTACTATAAAACTGGTGG (21)	$2339995381 \times 10^{15}$	5288344
ACAATTACTATAAAACTGGTGG (22)	$2.27343098 \times 10^{16}$	33764416
CCGAGGTAGGCGCGGCACCAGTTGT (25)	$3.31540211187376 \times 10^{26}$	1847251692933
GCCGAGGTAGGCGCGGCACCAGTTGT (26)	$1.124889101466300 \times 10^{28}$	31274997412295
GGCCGAGGTAGGCGCGGCACCAGTTGT (27)	$2.2331771897961938 \times 10^{29}$	4503599627370496
AACTCAAACGAGGAA (15)	7.7377×10^{11}	10.640
TAACTCAAACGAGGAA (16)	2.9794×10^{12}	112.71
TTAACTCAAACGAGGAA (17)	1.49974×10^{13}	1115.8
ATTAACTCAAACGAGGAA (18)	3.84402×10^{13}	6681.2
AATTAACTCAAACGAGGAA (19)	1.9501351×10^{14}	44167
GAATTAACTCAAACGAGGAA (20)	1.6910248×10^{15}	260699
AGAATTAACTCAAACGAGGAA (21)	$1.287701600 \times 10^{16}$	2807236
CAGAATTAACTCAAACGAGGAA (22)	$1.4479776892 \times 10^{17}$	48533801
CCAGAATTAACTCAAACGAGGAA (23)	$2.874586248 \times 10^{18}$	53155464
CCCAGAATTAACTCAAACGAGGAA (24)	$5.70674950220 \times 10^{19}$	3004200269
GTACAACCCGTCCTACC (17)*	$\times 10^{29}$	0.47120
GGTACAACCCGTCCTACC (18)*	$\times 10^{29}$	0.47315
TGGTACAACCCGTCCTACC (19)*	$\times 10^{29}$	0.75220
GTGGTACAACCCGTCCTACC (20)*	$\times 10^{29}$	0.90475
CCTCTGGAGCTGAGTC (16)	$\times 10^{29}$	246.23
GCCTCTGGAGCTGAGTC (17)	$\times 10^{29}$	8243.6
GGCCTCTGGAGCTGAGTC (18)	$\times 10^{29}$	467039
CGGCCTCTGGAGCTGAGTC (19)	$\times 10^{29}$	23653413
ACGGCCTCTGGAGCTGAGTC (20)	$\times 10^{29}$	168204324
TACGGCCTCTGGAGCTGAGTC (21)	$\times 10^{29}$	1160892396
TCGGGGAGATAACG (14)	$\times 10^{29}$	5.5198
CTCGGGGAGATAACG (15)	$\times 10^{29}$	123.52
CCTCGGGGAGATAACG (16)	$\times 10^{29}$	638.07
TCCTCGGGGAGATAACG (17)	$\times 10^{29}$	812.65
GTCCTCGGGGAGATAACG (18)	$\times 10^{29}$	22559
GGTCCTCGGGGAGATAACG (19)	$\times 10^{29}$	99287
GGGTCCTCGGGGAGATAACG (20)	$\times 10^{29}$	364060
AGGGTCCTCGGGGAGATAACG (21)	$\times 10^{29}$	8011340
AAAAATCTAAAA (12)	$\times 10^{29}$	0.015800
AAAAAATCTAAAA (13)	$\times 10^{29}$	0.047302
AAAAAAATCTAAAA (14)	$\times 10^{29}$	0.074813
TAAAAAAATCTAAAA (15)	$\times 10^{29}$	0.21869
ATAAAAAAATCTAAAA (16)	$\times 10^{29}$	1.6635
AATAAAAAAATCTAAAA (17)	$\times 10^{29}$	13.151
CAATAAAAAAATCTAAAA (18)	$\times 10^{29}$	338.71
GTAAATAAATAAAAA (15)	$\times 10^{29}$	0.022367
AGTAAATAAATAAAAA (16)	$\times 10^{29}$	0.089086
AAGTAAATAAATAAAAA (17)	$\times 10^{29}$	0.10925
TAAGTAAATAAATAAAAA (18)	$\times 10^{29}$	0.13638
ATAAGTAAATAAATAAAAA (19)	$\times 10^{29}$	0.16489

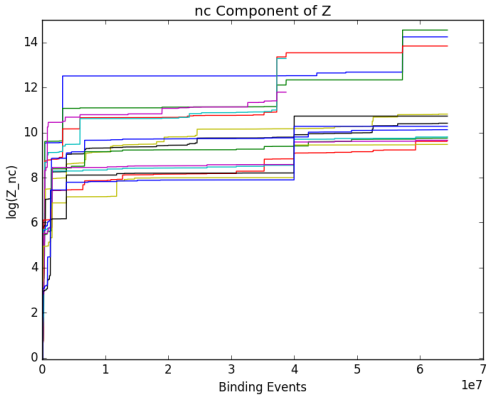
Table 4: *These sequences surprisingly occurred 2-3 times in just chromosome 20 itself. Therefore the Z_{nc} associated with the sequences were very large, rendering a low selectivity.



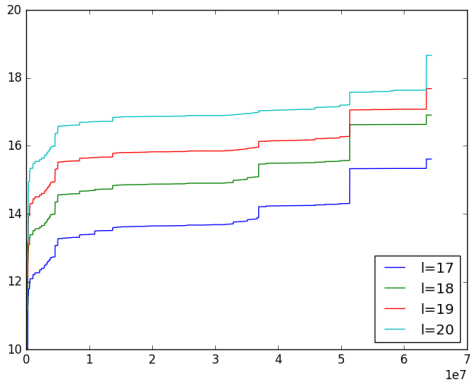
(a) Chromosome walk simulation for 3 sequences at the same position in the genome with $l = 20 - 22$



(b) Chromosome walk simulation for 3 sequences at the same position in the genome with $l = 25 - 27$



(c) Chromosome walk simulation for 10 sequences at the same position in the genome with $l = 15 - 24$



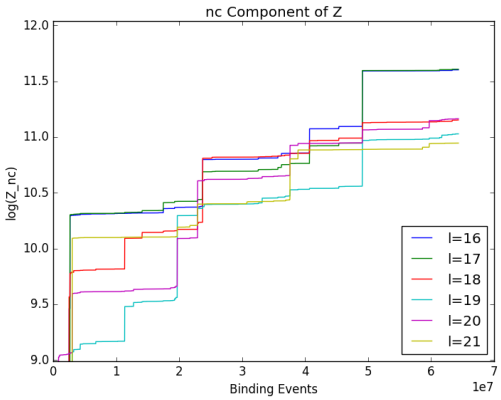
(d) Chromosome walk simulation for 4 sequences at the same position in the genome with $l = 17-20$. For the results associated with this figure, it is very interesting to note that sequences complementary to the targeted sequence appear 3 times in the just the chromosome (20) for $l = 17, 18$ and 2 times for $l = 19, 20$, which is a very rare occurrence. The randomly chosen site initially chosen by the program was 63511633. From the figure itself it can be seen that sharp rises occur at around positions 50Mbp and

Figure 25

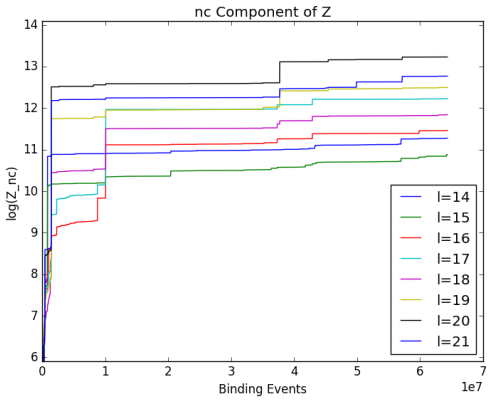
Note that the selectivities displayed in the tables above are only calculated having 'walked' over chromosome 20. It does not represent the selectivity associated with the binding over the entire genome. Therefore a correction factor of 50 can be utilised, which scales the selectivity to the correct value.

$$S_{ch20} = \frac{q_comp}{Z_{ch20} - q_{comp}}$$

$$S_{tot} \approx \frac{N_{ch20}}{N_G} \frac{q_comp}{Z_{ch20} - q_{comp}}$$



(a) Chromosome walk simulation for 6 sequences at the same position in the genome with $l = 16 - 21$



(b) Chromosome walk simulation for 8 sequences at the same position in the genome with $l = 14 - 21$

Figure 26

8.3.4 Monte Carlo Simulation of 'Kinetic Model'

(+Mathematical clarification of the 'kinetic model' to be included here)

The figure below shows the results of an example Monte Carlo simulation for the binding selectivity of a 20nt targeting sequence attempted to be bound to an example 30nt sequence taken from between positions 12888830 and 12888860 in human chromosome 20. The 20nt targeting sequence is TGATTTAGAACCTGAAAGCA and the 30nt targeted sequence is AGGAAACTAAATCTTGGACTTTCGTTCTGA. The simulation involves 10000 binding attempts per iteration, and 10 iterations.

For the random simulation, a kinetic **shotgun** approach is taken, where the probability of binding between two sequences is given by a The particular simulation is based on an expression $\frac{1}{1+Ae^{\beta\Delta G}}$ where ΔG is the total free energy associated with the 20nt binding of the targeting and targeted sequence. In the simulation, the constant A is 'normalised' such that the different between the binding probability of the complementary sequence and that of the single-defect sequence.

The expression

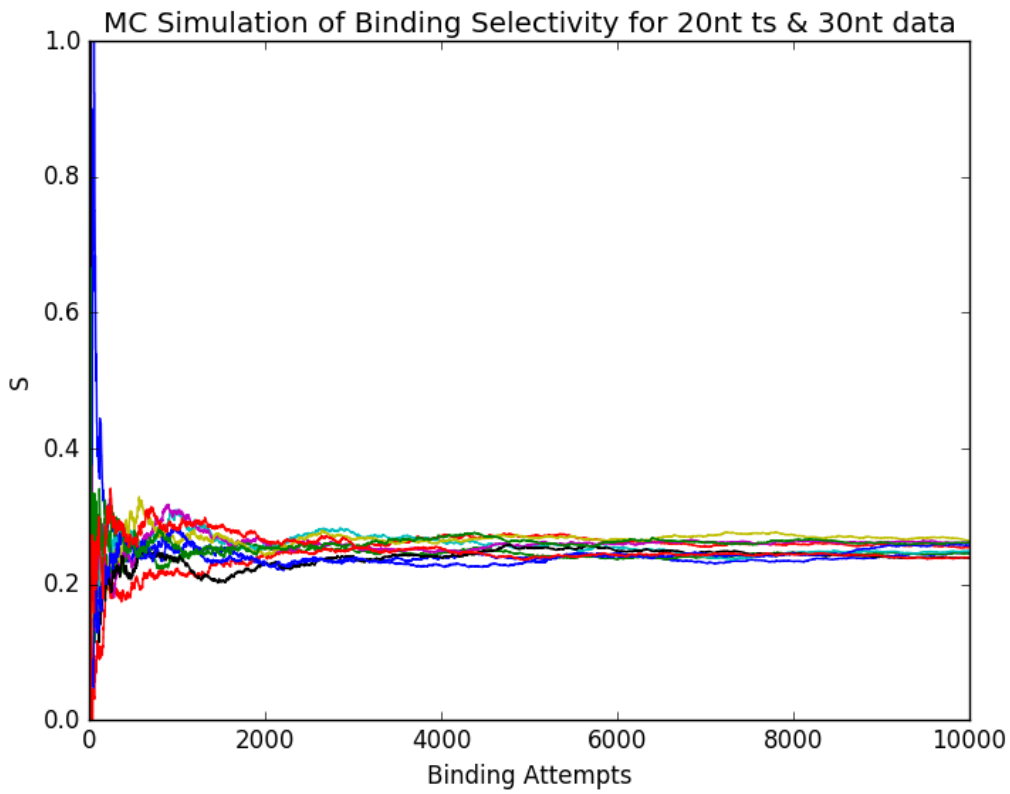


Figure 27: Chromosome 20, 0-12.9Mbp, 12.9-25.8Mbp and 25.8-38.6Mbp

8.3.5 List of Assumptions

+1.005eV for non-WC initiation energies in chromo-walk simulations

ϵ==0.054eV for model 1

Strict binary energetic states in thermodynamic models- resolved at least in part by kinetic models

References

[1] J SantaLucia. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proceedings of the National Academy of Sciences of the United States of America*, 95(4):1460–5, 1998.

[2] John SantaLucia and Donald Hicks. The thermodynamics of DNA structural motifs. *Annual review of biophysics and biomolecular structure*, 33:415–40, 2004.

[3] X Wu, A J Kriz, and P A Sharp. Target specificity of the CRISPR-Cas9 system. *Quant Biol*, 2(2):59–70, 2014.