Understanding Targeting Selectivity in CRISPR CAS9 (Summarized)

Han Yao Choong

September 8, 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 Key Findings

-Model 1 may provide thermodynamic justifications for a minimum viable targeting sequence length at around $l = 19$

-An analysis of differences between S_{ideal} and S_{real}

-Model 3 may provide thermodynamic justifications for a minimum viable targeting sequence length at around $l = 16 - 18$

-Positive correlation between GC composition of a targeting sequence and targeting selectivity observed with Santa Lucia rules implemented

-Selectivities (S_{ch20}) derived from 'chromosome-walk' with Santa Lucia rules have a bias to be larger than those derived from model 3 partition function factorization.

-In general, thermodynamic models suggest that targeting selectivity (S_{real}) increases with targeting sequence length

-Solved for the time evolution of the population vector of the zipper model

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 -)
Z	Partition Function
S	Targeting selectivity
S_{ch20}	Targeting selectivity but targeted genome only limited to chromosome 20

Table 1: Definition of variables used

3 Model 1

3.1 Model Description & Definitions

Model 1:

-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.

-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)

Explicitly outlining the model, in the general definition of S, q_{comp} is given by

$$\begin{aligned} q_{comp} &= p_A^{N_S^T} p_T^{N_S^A} p_C^{N_S^G} p_G^{N_S^C} e^{-\beta l \epsilon} \\ &(= p_c e^{-\beta \Delta G_c}) \end{aligned} \tag{1}$$

and Z for model 1 is given by

$$\begin{aligned} Z &= \sum_{N_{cp}, N_{G_S}^A, N_{G_S}^T, N_{G_S}^C, N_{G_S}^G}^{4^l seqs} q_{(N_{cp}, N_{G_S}^A, N_{G_S}^T, N_{G_S}^C, N_{G_S}^G)} \\ &= \sum_{N_{cp}, N_{G_S}^A, N_{G_S}^T, N_{G_S}^C, N_{G_S}^G}^{4^l seqs} p_A^{N_{G_S}^A} p_T^{N_{G_S}^T} p_C^{N_{G_S}^C} p_G^{N_{G_S}^G} e^{-\beta N_{cp} \epsilon} \\ &(= Z_{ideal}) \end{aligned} \tag{2}$$

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

$$\begin{aligned} Z &= \prod_{i=1, X_i \in \{ts\}}^{l} [1 + p_{X_i^*}(e^{-\beta\epsilon} - 1)] \\ &(= Z_{ideal}) \end{aligned} \tag{3}$$

With Z and q_{comp} defined as above, targeting selectivities such as S_{ideal} can be defined

$$S_{ideal} = \frac{1}{\frac{Z}{q_{comp}} - 1} \tag{4}$$

3.2 Results & Discussion

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 one type of nucleotide (it does not matter at this stage which one). The genome nucleotide frequencies are equal, i.e. p_A=p_T=p_C=p_G=0.25. 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 uniform access to the entire human genome in a cell nucleus.

The plots below show the result for these targeting sequences from length of 1 to 40 nucleotides with the aforementioned conditions satisfied.

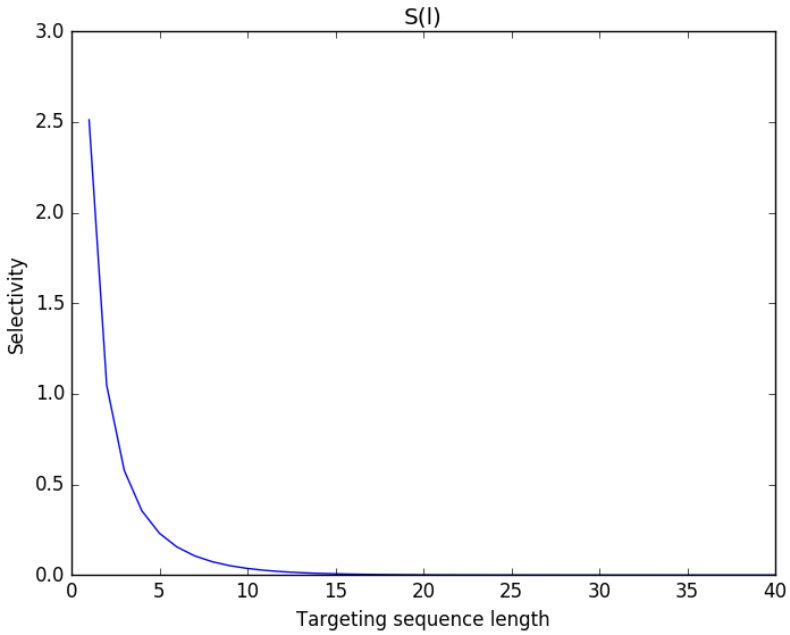


Figure 1: A plot of S_{ideal} against the targeting sequence length l.

The figure above (fig. 1) shows a monotonic decrease of S against l. This is because for the simplistic S_{ideal} the contribution to Z associated with the complementary sequence is scaled by p^l. Hence as l increases, the p^l factor suppresses the probability of complementary binding, causing the non-complementary contribution to Z to increase as a proportion of Z.

With $S_{ideal} = \frac{(pe^{-\beta\epsilon})^l}{[1+p(e^{-\beta\epsilon}-1)]^l-(pe^{-\beta\epsilon})^l}$, directly substituting in values used in the simulation, one arrives at the expressions $S_{ideal} \approx \frac{1.88^l}{2.63^l-1.88^l} = \frac{1}{1.40^l-1}$.

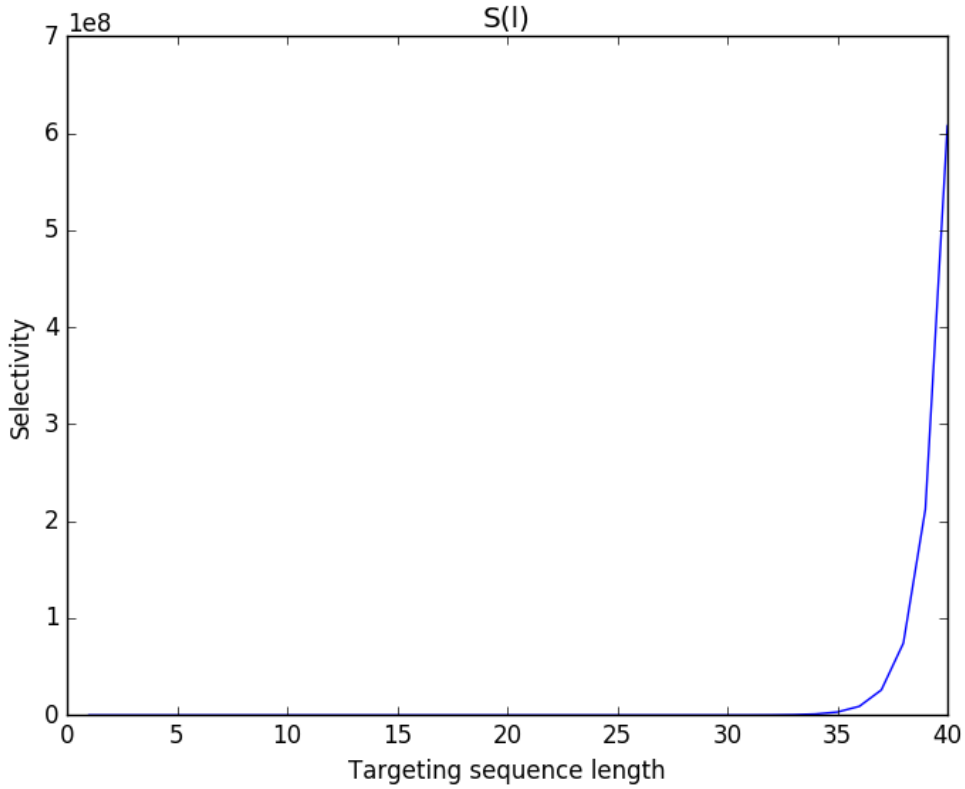


Figure 2: A plot of S_{real} against the targeting sequence length l.

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 because the probability of the complementary sequence occurring is not suppressed by p^l as is the case before. As S_{real} ensures the existence of a complementary sequence, the Z_{real} correction does indeed result to more accurate selectivity compared to the uncorrected mean field model.

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, what can we say about the model's apparent invalidity for l»20, as S increases infinitely?

The increase in selectivity with l is expected as it becomes less and less likely for complementary or near-complementary sequences to occur randomly in the genome as l increases. (further comparison to experimental observations could be made)

As for the second point, it is important to consider that the model assumes the binding energy can increase infinitely with l (which is unrealistic by itself), which leads to the dominance of the exponential 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 S_{real} crosses unity at around 20nt for l may seem a promising indication. Therefore the model may provide some simple equilibrium energetic justifications of why $l << 20$ lead to low selectivity, hence the favourability of $l = 20$ in nature.

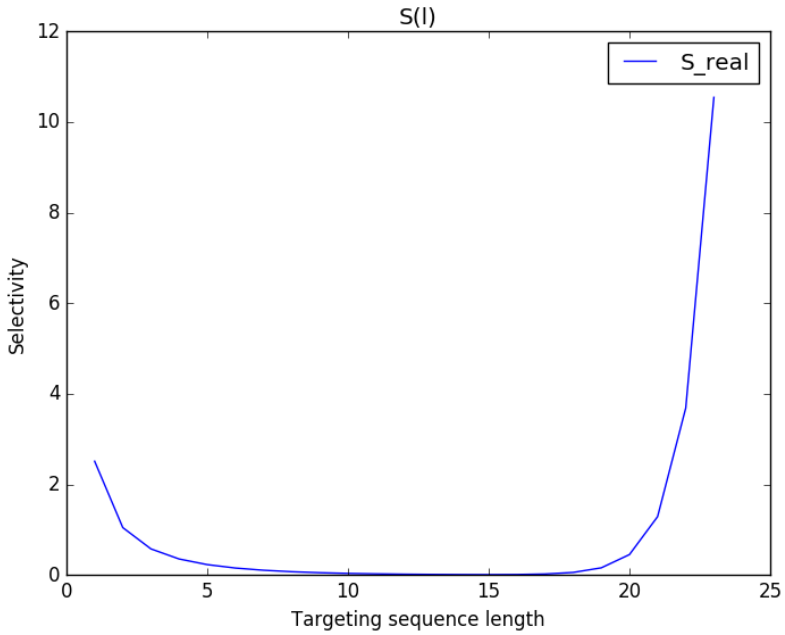


Figure 3: A plot of the modified S_{real} against l. Here the modified S_{real} is that which allows 'correct binding' to be at more than one site in the genome, should the complementary sequence occur 'randomly' via the mean field.

The figure above (fig. 3) 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 (these complementary sequences occur with mean field probabilities). 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.

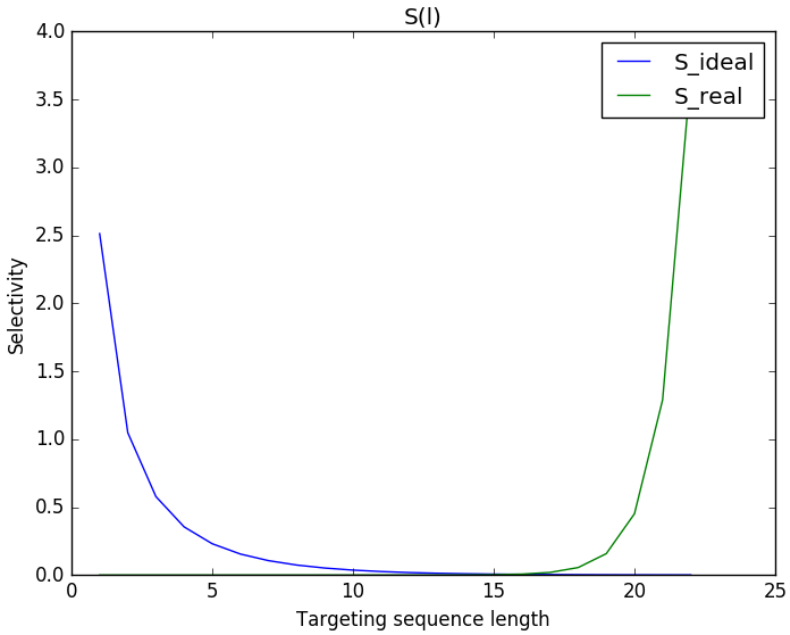


Figure 4: The blue line indicates S_{ideal} while the green line indicates S_{real} against the targeting sequence length l.

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 may depend on the length of the genome accessible to the targeting sequence, associated with the organism under question.

The following section discusses the effect of nucleotide frequency p_x and pairwise binding energy ϵ on targeting selectivity.

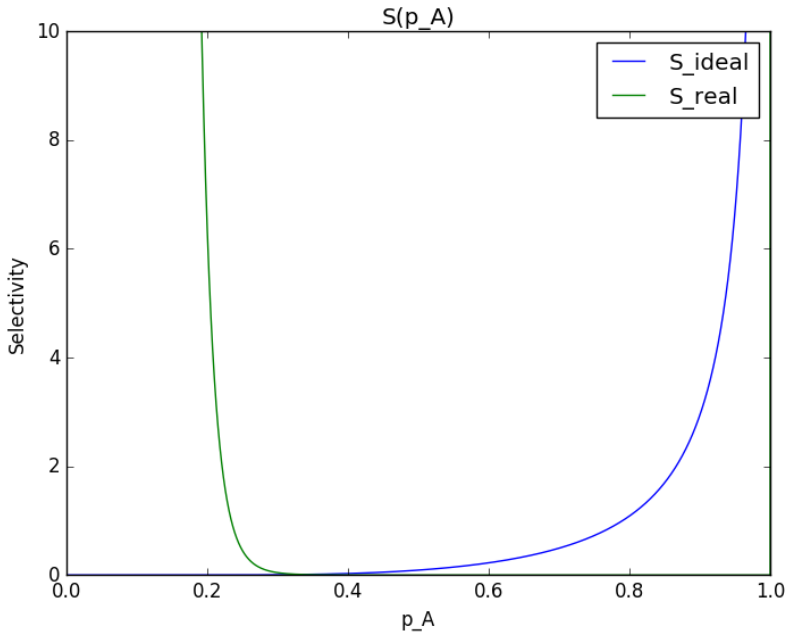


Figure 5: The blue line indicates S_{ideal} while the green line indicates S_{real} against the 'A' nucleotide frequency in the genome. The targeting sequence in this case is a 20nt sequence consisting only of T.

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)

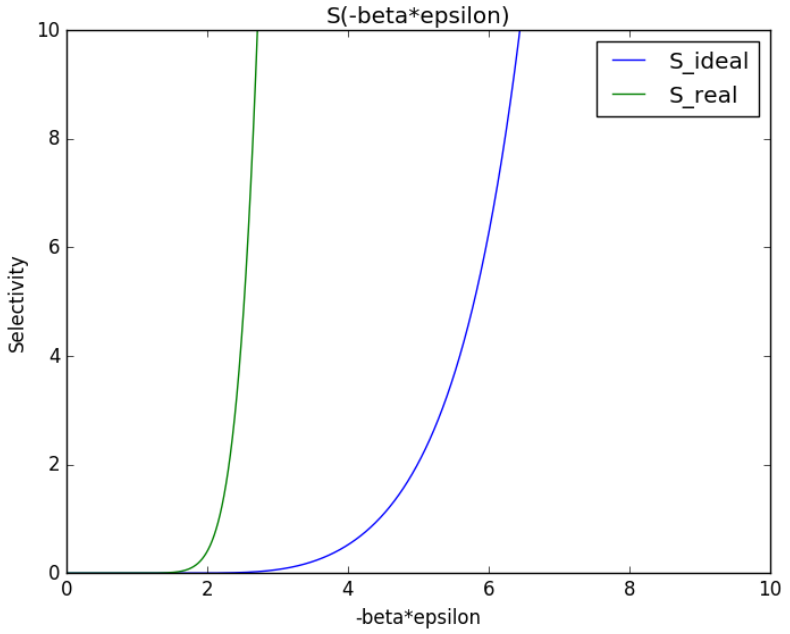


Figure 6: Ideal and real selectivities as a function of the exponential argument $(-\beta\epsilon)$ of the energy term. In this scenario again the targeting sequence is 20nt, consisting only of T. p_A is set as 0.25.

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.

4 Model 3

4.1 Nearest Neighbour Model

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

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

$$E_{tot} = E_{init} + E_{sym} + \sum_{i=1}^{l-1} E_{dpx_i} \tag{5}$$

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. The symmetry correction was ignored for now.

With the binding energy calculated, S_{real} is then given by $\frac{e^{-\beta\Delta G}}{(N_G - l)Z_{ideal}}$, assuming that there is only one 'correct' complementary site in the genome. Z_{ideal} can once again be calculated by partition function factorization, with nucleotide correlation and the differing stacking parameters taken into account.

The following table outlines the energies of the 10 possible no-mismatch propagation sequences, and the initiation energies, taken from Santa Lucia et al (1998). The values quoted correspond to 'unified' energies which are derived from data collated from 7 separate experiments at varying salt concentrations and conditions.

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.

Propagation sequence	X	Y			
		A	C	G	T
GX/CY	A	0.17	0.81	−0.25	WC
	C	0.47	0.79	WC	0.62
	G	−0.52	WC	−1.11	0.08
	T	WC	0.98	−0.59	0.45
CX/GY	A	0.43	0.75	0.03	WC
	C	0.79	0.70	WC	0.62
	G	0.11	WC	−0.11	−0.47
	T	WC	0.40	−0.32	−0.12
AX/TY	A	0.61	0.88	0.14	WC
	C	0.77	1.33	WC	0.64
	G	0.02	WC	−0.13	0.71
	T	WC	0.73	0.07	0.69
TX/AY	A	0.69	0.92	0.42	WC
	C	1.33	1.05	WC	0.97
	G	0.74	WC	0.44	0.43
	T	WC	0.75	0.34	0.68

[a]WC indicates a Watson-Crick pair, which is given in Table 1. Error bars and ΔH° and ΔS° parameters are provided in the original references.

Figure 7: Stacking parameters for propagation sequences with a single defect

Although the model is well established for complementary sequences, and cases of single and bubble defects. However the case of adjacent double defects appears to have not been investigated experimentally. Therefore due to this lack of experimental data for dimer duplex energies consisting of adjacent defects (double defects), in the models and simulations constructed very rough approximations were utilised to calculate the double defect dimer duplex energies.

4.2 Chromosome 20 Simulation

a **"chromosome walk"** model

To test model 3, an algorithm is used, 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. Once again, the human chromosome 20 is used for this simulation due to a high percentage of sequenced base pairs.

4.2.1 Initial Test

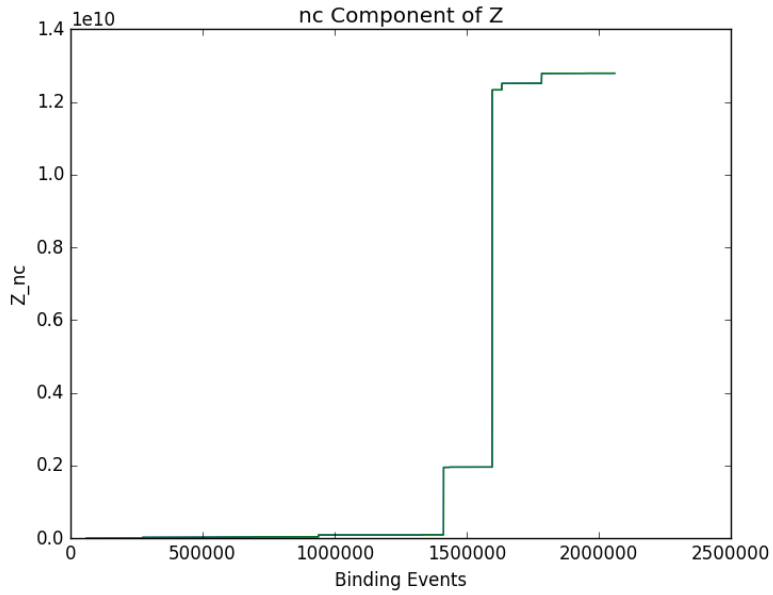


Figure 8: The Z_{nc} in the figure above refers to the non-complementary component of the partition function. The number of binding events refers to the position along the chromosome strand the targeting sequence has moved.

The figure above shows an chromosome-walk simulation for chromosome 20 done to test the program initially. 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 of Z_{nc} is $Z - q_{comp}$. Hence ($S = \frac{1}{\frac{Z}{q_{comp}}-1} = \frac{q_{comp}}{Z-q_{comp}} = \frac{q_{comp}}{Z_{nc}}$).

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.

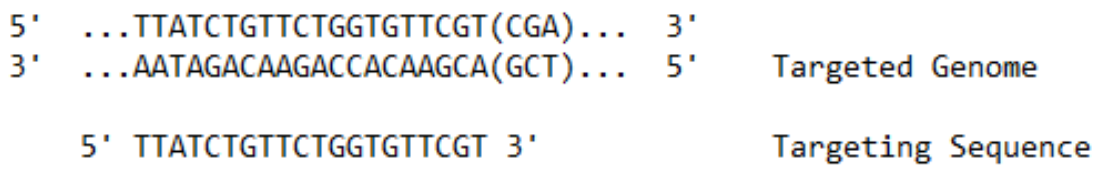
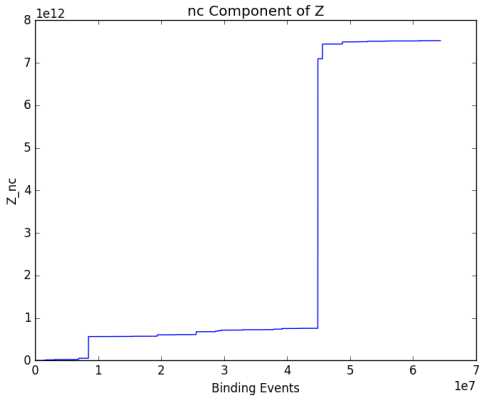


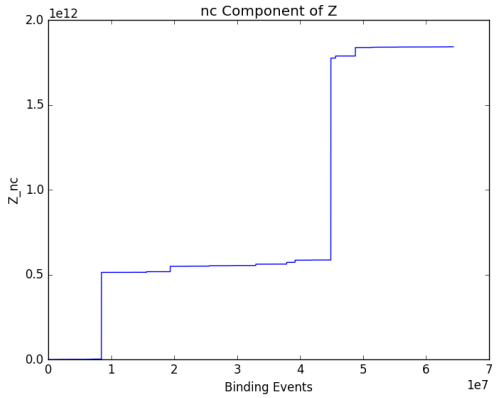
Figure 9: A visualisation of the binding of the user defined targeting sequence to the region 60023-60045 of chromosome 20. Notice the presence of the 5'-NGA-3' PAM sequence in the region 60043-60045.

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

    5' UUAUCUGUUCUGGUGUUCGU 3'            Targeting Sequence
```

Figure 10: As a brief comment and reminder, the actual scenario would involve uracil rather than thymine, 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 11

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 stacking parameters.)

4.2.2 Results

The figures below show plots of S_{ch20} against l for sequences randomly selected in chromosome 20.

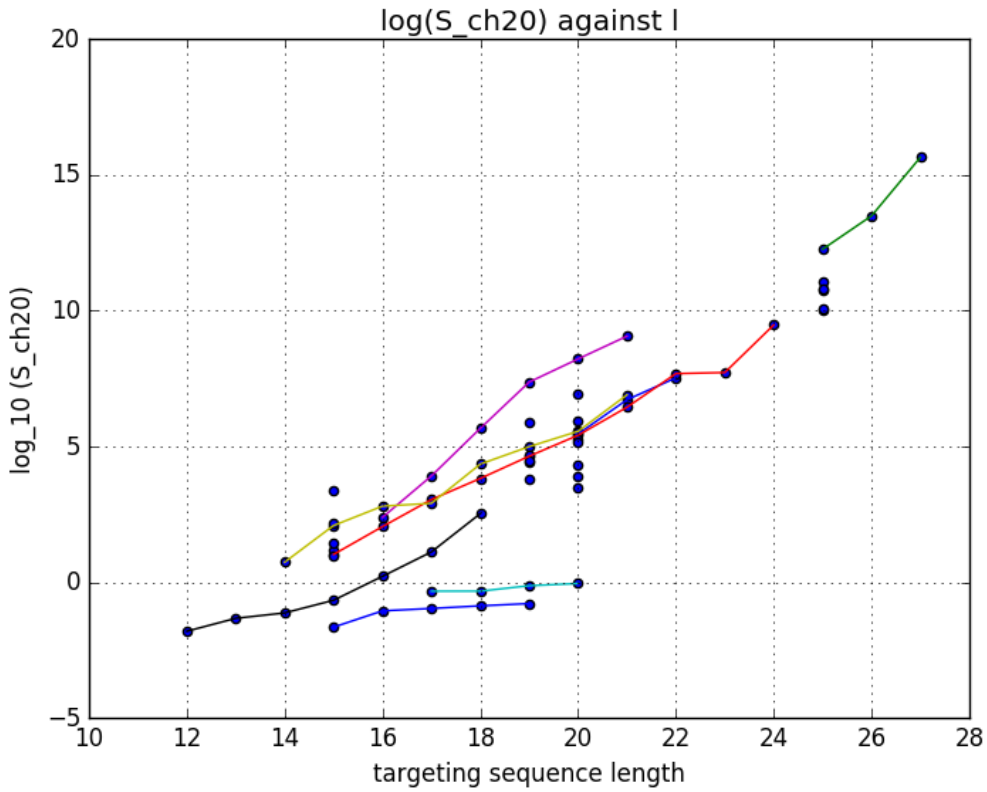


Figure 12: This figure shows the $log_{10}(S_{ch20})$ of 69 targeting sequences tested on chromosome 20. The data points connected with lines correspond to sequences of different l that end on exactly the same nucleotide in the genome. *See appendix for the data represented in this graph and for more details.*

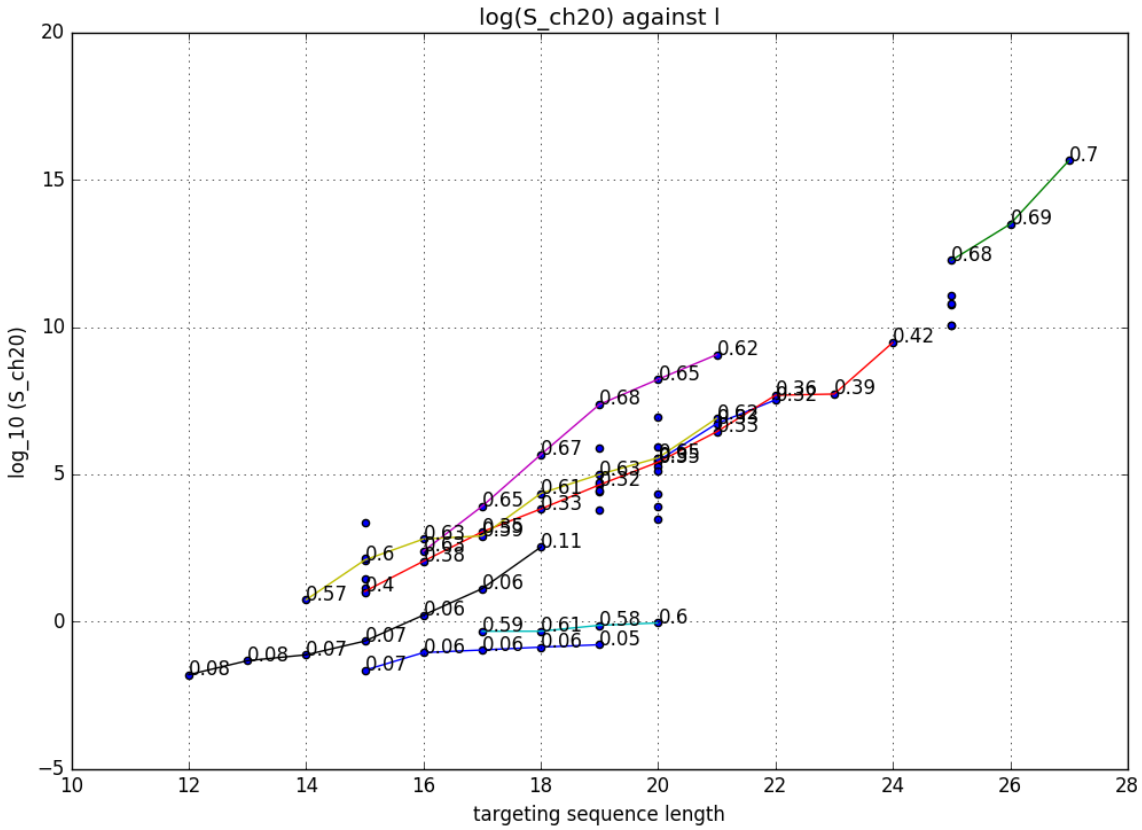
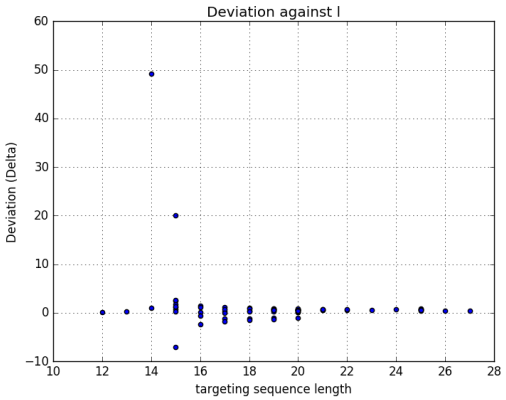


Figure 13: The consecutive data points here are labelled with the fraction of G or C nucleotides in the targeting sequence. There is some positive correlation between GC composition and selectivity.

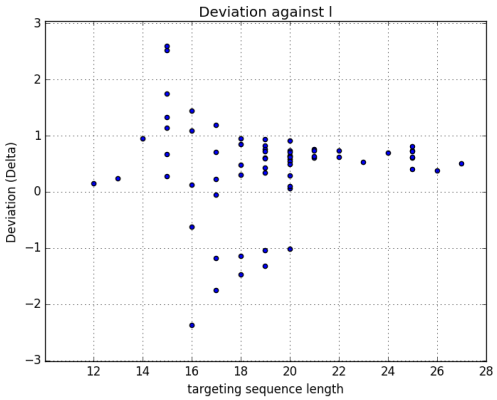
It is important to note here that discrepancies between the consecutive events is due to the differences in nucleotide composition of the targeting sequence. Notably it can be seen that such differences may result to thermodynamic selectivities up to 8 orders of magnitude apart while the sequences have the same l. However these discrepancies in selectivity still cannot completely or reliably be explained purely by differences in GC composition. Therefore a more sophisticated measure of fractional abundance of nucleotides with high average binding energies may prove to correlate better with selectivity.

Therefore the fact that GC composition cannot fully account for this means that another factor that is likely to contribute to the explanation is that chromosome 20, with N_G=64Mbp is not long enough to converge to a mean field. However this does not explain the obvious positive bias of computed selectivities.

It was also seen that the selectivities obtained are generally greater than those predicted by the mean field model. This is shown in fig. 14.



(a)



(b) Enlarged view

Figure 14: Figures show the deviation of the computed selectivities from the those predicted using the NN mean field model (model 3). The deviation is defined by $\Delta = \frac{log(S)-log(S_{pred})}{log(S_{pred})}$. In the second subfigure there is a clear positive bias, which remains to be accounted for.

5 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,

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

5.1 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.

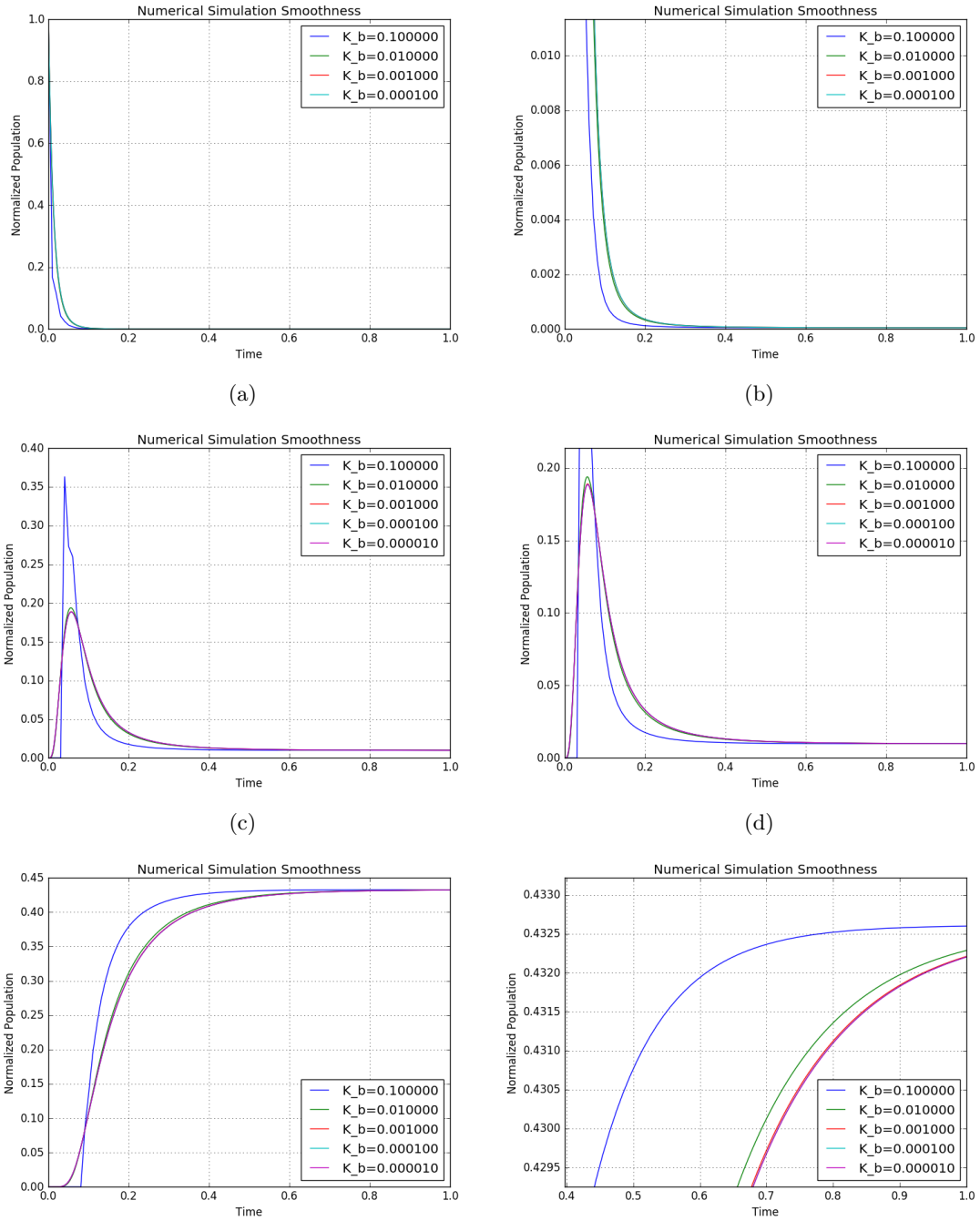
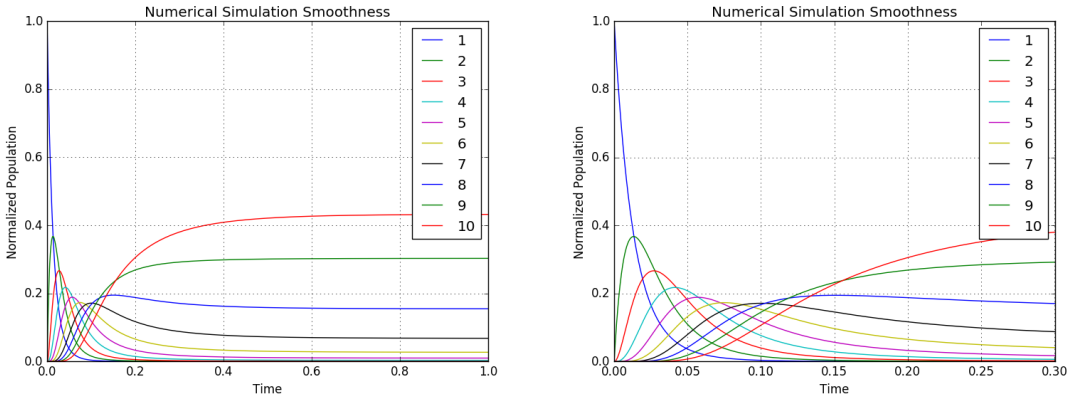


Figure 15

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

5.2 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 $x_0(t) = x_0$. With $x_1(t) = x_0 + \int_0^t \underline{\underline{A}}x_0 ds = x_0 + t\underline{\underline{A}}x_0$, $x_2(t) = x_0 + \int_0^t \underline{\underline{A}}x_1 ds = x_0 + \int_0^t (1 + s\underline{\underline{A}})\underline{\underline{A}}x_0 ds = x_0 + t\underline{\underline{A}}x_0 + \frac{t^2}{2!}\underline{\underline{A}}^2 x_0$,... it =then follows that for the n^{th} iteration, the solution is $x_n(t) = x_0 + t\underline{\underline{A}}x_0 + \frac{t^2}{2}\underline{\underline{A}}^2 x_0 + ... + \frac{t^n}{n!}\underline{\underline{A}}^n x_0$. 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.)

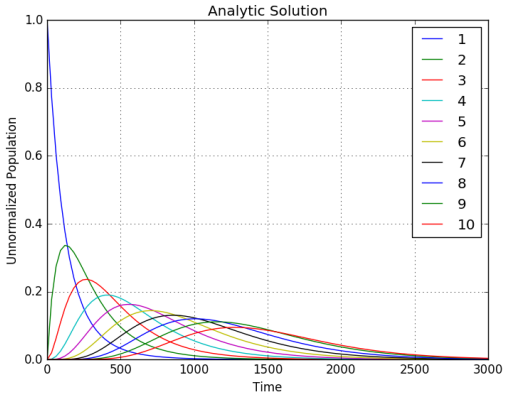
With $\sum_{k>0}(\frac{1}{k!}\underline{\underline{A}}^k) = \sum_{k>0}(\frac{1}{k!}\underline{\underline{S}}\underline{\underline{\Lambda}}^k\underline{\underline{S}}^{-1}) = \underline{\underline{S}}(\sum_{k>0}\frac{1}{k!}\underline{\underline{\Lambda}}^k)\underline{\underline{S}}^{-1}$, it then follows 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,

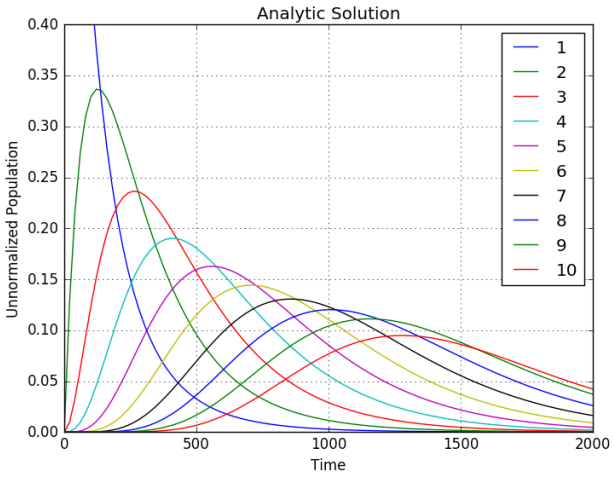
$$\begin{pmatrix} P_1 \\ P_2 \\ P_3 \\ \vdots \\ P_l \end{pmatrix} = \underline{\underline{S}} \begin{pmatrix} e^{\lambda_1 t} & 0 & 0 & \dots & 0 \\ 0 & e^{\lambda_2 t} & 0 & \dots & 0 \\ 0 & 0 & e^{\lambda_3 t} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & e^{\lambda_l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

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

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 17

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 18

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 $\underline{P}$.

6 Remarks

6.1 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.

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

The thermodynamic models constructed- especially model 3 (NN model) are partly constrained by a lack of specific data. Importantly, data for double-defect (adjacent defect) stacking parameters are missing from SantaLucia et al.'s DNA thermodynamic database. Because of this, the adjacent defect energies are derived from the known single defect stacking parameters by effectively decoupling the effect of the identity of the nearest neighbour nucleotide. In addition, initiation energies involving defects are also missing. As such, in practice when simulating the NN model, the contribution of the initiation energy to the total ΔG has been neglected.

6.2 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. 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 selectivity sufficient for practical purposes- 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.

7 Appendix

7.1 Real Selectivity Corrections

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.

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

7.2 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.

7.3 Miscellaneous Results

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

7.3.1 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, 0.380427517737918 83],
[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 19: 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.

7.3.2 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 20: 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 21: Chromosome 20 statistical parameters for the regions: 0-12.9Mbp, 12.9-25.8Mbp and 25.8-38.6Mbp

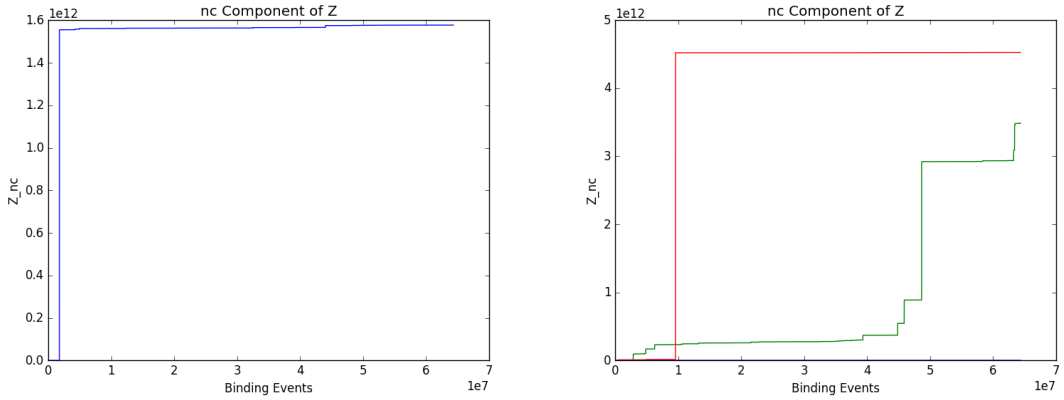
```
In [86]: runfile('C:/Users/Choon/Desktop/CRISPR/Code/Algorithms/
nt_data_param_extraction.py', wdir='C:/Users/Choon/Desktop/CRISPR/Code/Algorithms')
[0.2737806653077641, 0.2741630894094981, 0.2265305343416688, 0.225525710941069]
[[0.30584096808099726, 0.23698923658569002, 0.1859799876045459,
0.27118980772876683], [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 22: 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.

7.3.3 Model 3: More results from chromo-walk simulations



(a) Z_{nc} is the non-complementary component of the partition function. The number of binding 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.

Figure 23

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.

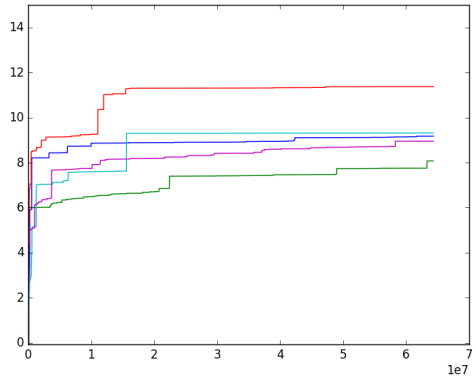
Targeting seq. (l)	Z_{ch20}	S_{ch20} (pred)
TTTATAATGTATAAAAACTA (20)	6.21333×10^{11}	2898 (1735)
GGTGGGTGGACAACCTGAGA (20)	$2.96949442 \times 10^{18}$	851471 (2877)
TCTGGGTCTTGACATCTAGG (20)	3.82331×10^{16}	8445 (3832)
CAGAAAATTAGTATGGGAAA (20)	1.4557813×10^{14}	341724 (2617)
CACTAACTCAGGGGAATGAT (20)	$7.53451162 \times 10^{15}$	8443608 (4210)
TCTATGACGTACGAATCTG (19)	$1.12671881 \times 10^{15}$	750146 (1096)
TTTAGAATAAGTAATAACA (19)	7.548886×10^{11}	6363 (455)
GTTAGGGGACAAAGGAGAC (19)	$1.22648875 \times 10^{16}$	51148 (530)
GTATCACTGACTATATGGT (19)	5.5842427×10^{13}	26930 (2058)
TGTTTTCTGTTATTATCGA (19)	2.6029006×10^{13}	29155 (625)
TGGGTTACGACACCT (15)	2.077×10^{13}	14.50 (3.49)
TCAACACGCCAACGA (15)	7.411×10^{13}	144.4 (19.7)
CCACACTTCACGACT (15)	9.663×10^{12}	9.78 (5.97)
GAACGTACTTTGGTC (15)	1.1923×10^{12}	28.11 (2.52)
CCATTCTGCGAACTT (15)	1.619981×10^{12}	2302 (27.5)
TCACGATTTCATTTCAACAGTATTG (25)	$8.2611793685997 \times 10^{18}$	1.201×10^{11} (2528881)
TTGAAACGGTAGTAAGTCGTGAAAG (25)	$1.3112403059344 \times 10^{21}$	5.828×10^{10} (911827)
CAAGAAGAAAGGATTTAATCACTAG (25)	$1.45323317407 \times 10^{18}$	1.104×10^{10} (1810805)
GAGAGGACTCGATGTCTTCCTTACC (25)	$3.309194806425 \times 10^{21}$	6.161×10^{10} (1867507)
GACGAGGGGAACGGGTGACCCCTGA (25)	$1.996684335525 \times 10^{25}$	1.139×10^{10} (1642360)
GGTCGTCGTGTAGTTGTAGA (20)	4.9752800×10^{17}	191105 (2562)
ATAGAGGTTTTATTGACTGA (20)	1827521×10^{14}	20935 (2259)
TGAAAGAATGGGGACAACAG (20)	$2.27344769 \times 10^{16}$	136046 (2679)

[H]

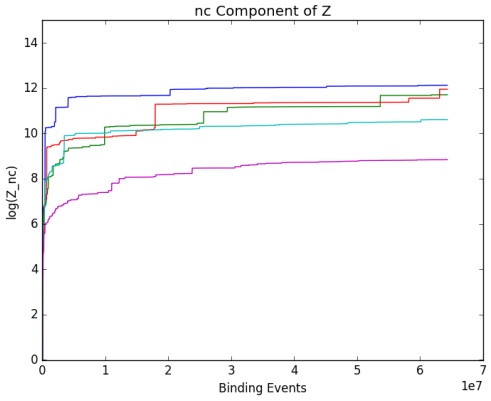
Table 3

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.



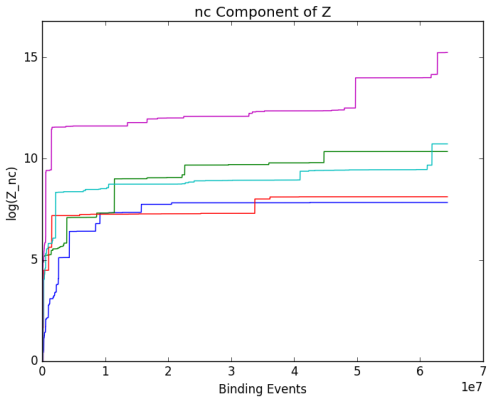
(a) Z_{nc} is the non-complementary component of the partition function. The number of binding events refers to the position along the chromosome strand the targeting sequence has moved.

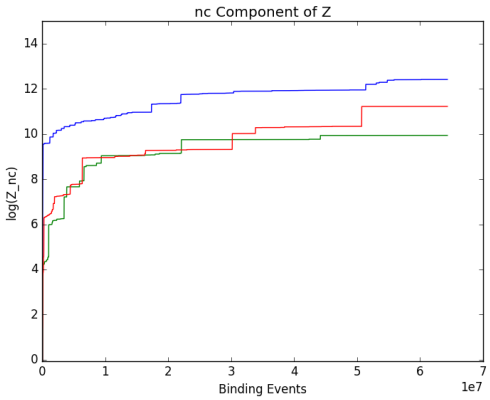


(b) Z_{nc} is the non-complementary component of the partition function. The number of binding events refers to the position along the chromosome strand the targeting sequence has moved. Chromosome walk simulation for 5 random $l = 15$ targeting sequences

Figure 24

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

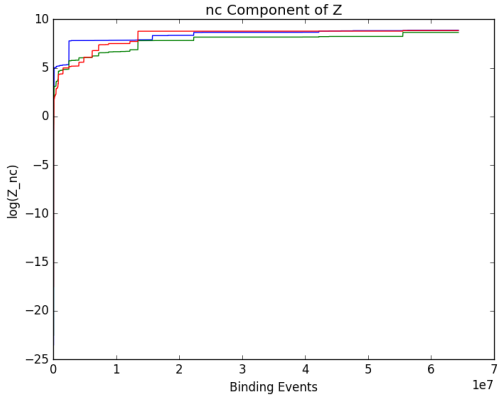
(b) Z_{nc} is the non-complementary component of the partition function. The number of binding events refers to the position along the chromosome strand the targeting sequence has moved. Chromosome walk simulation for 3 random $l = 20$ targeting sequences

Figure 25

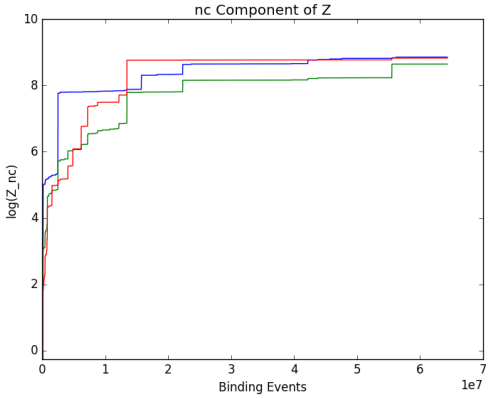
For the next few simulations, the portion of the genome complementary to the targeting sequence was chosen to be at the same location in the genome (i.e. they all border a particular PAM sequence in the genome). (With the exception for the first one)

Targeting seq. (l)	Z_{ch20}	S_{ch20} (Pred.)
AATTACTATAAAACTGGTGG (20)	$2.08098868 \times 10^{14}$	289784(2732)
CAATTACTATAAAACTGGTGG (21)	$2339995381 \times 10^{15}$	5.288344×10^{6} (15706)
ACAATTACTATAAAACTGGTGG (22)	$2.27343098 \times 10^{16}$	3.3764416×10^{7} (44673)
CCGAGGTAGGCGCGGCACCAGTTGT (25)	$3.31540211187376 \times 10^{26}$	$1.84 \times 10^{12}(5.15 \times 10^{8})$
GCCGAGGTAGGCGCGGCACCAGTTGT (26)	$1.124889101466300 \times 10^{28}$	$3.13 \times 10^{13}(5.77 \times 10^{9})$
GGCCGAGGTAGGCGCGGCACCAGTTGT (27)	$2.2331771897961938 \times 10^{29}$	4.50×10^{15} (2.53×10^{10})
AACTCAAACGAGGAA (15)	7.7377×10^{11}	10.640 (1.95)
TAACTCAAACGAGGAA (16)	2.9794×10^{12}	112.71 (6.88)
TTAACTCAAACGAGGAA (17)	1.49974×10^{13}	1115.8 (24.5)
ATTAACTCAAACGAGGAA (18)	3.84402×10^{13}	6681.2 (90.7)
AATTAACTCAAACGAGGAA (19)	1.9501351×10^{14}	44167 (359)
GAATTAACTCAAACGAGGAA (20)	1.6910248×10^{15}	260699 (1310)
AGAATTAACTCAAACGAGGAA (21)	$1.287701600 \times 10^{16}$	2807236 (4620)
CAGAATTAACTCAAACGAGGAA (22)	$1.4479776892 \times 10^{17}$	48533801 (27600)
CCAGAATTAACTCAAACGAGGAA (23)	$2.874586248 \times 10^{18}$	53155464 (108030)
CCCAGAATTAACTCAAACGAGGAA (24)	$5.70674950220 \times 10^{19}$	3004200269 (398792)
GTACAACCCGTCCTACC (17)*	$\times 10^{29}$	0.47120 (60.3)
GGTACAACCCGTCCTACC (18)*	$\times 10^{29}$	0.47315 (179)
TGGTACAACCCGTCCTACC (19)*	$\times 10^{29}$	0.75220 (764)
GTGGTACAACCCGTCCTACC (20)*	$\times 10^{29}$	0.90475 (3360)
CCTCTGGAGCTGAGTC (16)	$\times 10^{29}$	246 (135)
GCCTCTGGAGCTGAGTC (17)	$\times 10^{29}$	8243.6 (1520)
GGCCTCTGGAGCTGAGTC (18)	$\times 10^{29}$	467039 (6645)
CGGCCTCTGGAGCTGAGTC (19)	$\times 10^{29}$	23653413 (38172)
ACGGCCTCTGGAGCTGAGTC (20)	$\times 10^{29}$	168204324 (101473)
TACGGCCTCTGGAGCTGAGTC (21)	$\times 10^{29}$	1160892396 (340733)
TCGGGGAGATAACG (14)	$\times 10^{29}$	5.52 (1.03)
CTCGGGGAGATAACG (15)	$\times 10^{29}$	123.5 (5.77)
CCTCGGGGAGATAACG (16)	$\times 10^{29}$	638.1 (21.8)
TCCTCGGGGAGATAACG (17)	$\times 10^{29}$	812.7 (50.8)
GTCCTCGGGGAGATAACG (18)	$\times 10^{29}$	22600 (227)
GGTCCTCGGGGAGATAACG (19)	$\times 10^{29}$	99300 (675)
GGGTCCTCGGGGAGATAACG (20)	$\times 10^{29}$	364000 (2480)
AGGGTCCTCGGGGAGATAACG (21)	$\times 10^{29}$	8010000 (9480)
AAAAATCTAAAA (12)	$\times 10^{29}$	0.0158 (0.0276)
AAAAAATCTAAAA (13)	$\times 10^{29}$	0.0473 (0.0854)
AAAAAAATCTAAAA (14)	$\times 10^{29}$	0.0748 (0.264)
TAAAAAAATCTAAAA (15)	$\times 10^{29}$	0.219 (0.930)
ATAAAAAAATCTAAAA (16)	$\times 10^{29}$	1.66 (3.81)
AATAAAAAAATCTAAAA (17)	$\times 10^{29}$	13.2 (15.0)
CAATAAAAAAATCTAAAA (18)	$\times 10^{29}$	338.7 (86.1)
GTAAATAAATAAAAA (15)	$\times 10^{29}$	0.022367 (1.87)
AGTAAATAAATAAAAA (16)	$\times 10^{29}$	0.089086 (5.84)
AAGTAAATAAATAAAAA (17)	$\times 10^{29}$	0.10925 (19.0)
TAAGTAAATAAATAAAAA (18)	$\times 10^{29}$	0.13638 (67.1)
ATAAGTAAATAAATAAAAA (19)	$\times 10^{29}$	0.16489 (275)

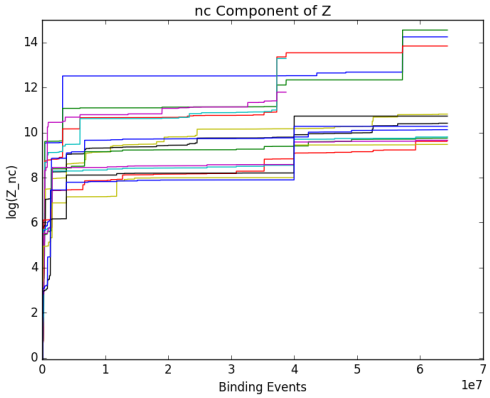
Table 4: *These sequences surprisingly occurred 2-3 times in just chromosome 20 itself. Therefore the Z_{nc} associated with the sequences were very large, rendering a low selectivity.



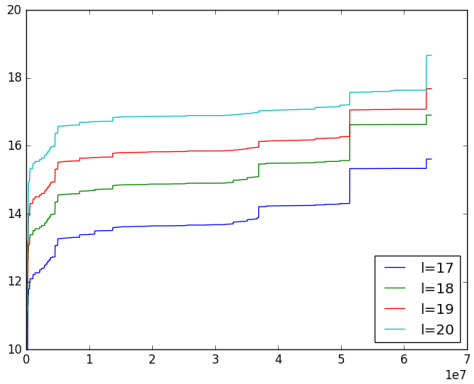
(a) Chromosome walk simulation for 3 sequences at the same position in the genome with $l = 20 - 22$



(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 $l = 15 - 24$



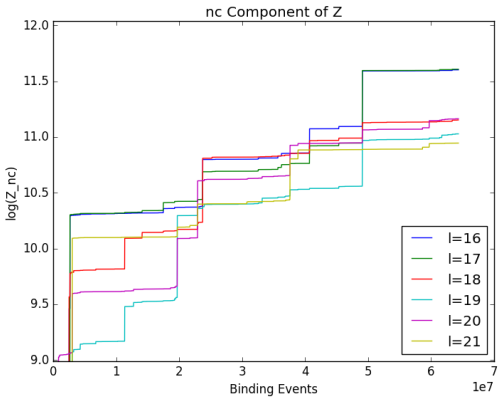
(d) Chromosome walk simulation for 4 sequences at the same position in the genome with $l = 17-20$. For the results associated with this figure, it is very interesting to note that sequences complementary to the targeted sequence appear 3 times in the just the chromosome (20) for $l = 17, 18$ and 2 times 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 26

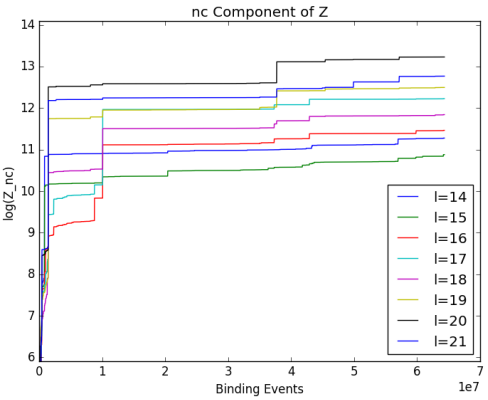
Note that the selectivities displayed in the tables above are only calculated having 'walked' over chromosome 20. It does not represent the selectivity associated with the binding over the entire genome. Therefore a correction factor of 50 can be utilised, which scales the selectivity to the correct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

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

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

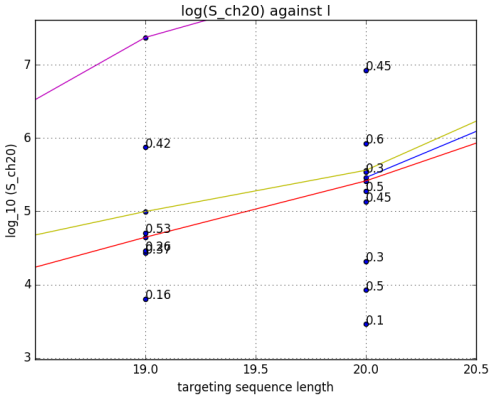


(a) Chromosome walk simulation for 6 sequences at the same position in the genome with $l = 16 - 21$

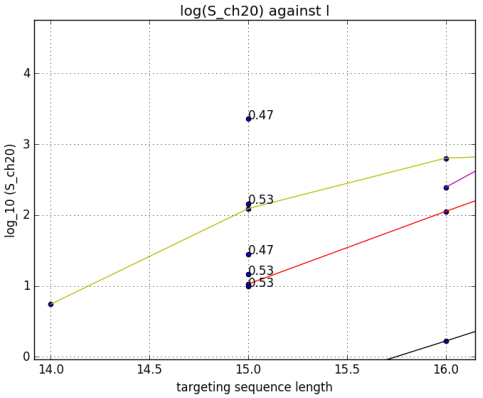


(b) Chromosome walk simulation for 8 sequences at the same position in the genome with $l = 14 - 21$

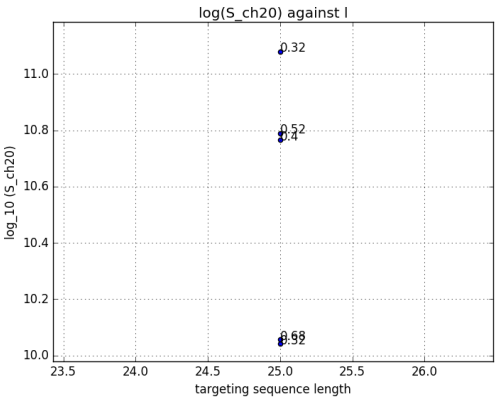
Figure 27



(a)



(b)



(c)

Figure 28

7.3.4 List of Assumptions

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

ϵ==0.054eV for model 1

The DNA thermodynamic database stacking parameters are valid under the condition of sodium concentrations being at 1M, which is not physiological.

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

References