Cellular trade-offs in the non-equilibrium synthesis of complex molecular information

by

Alkesh Yadav

2022

A thesis submitted to
the Jawaharlal Nehru University for the degree of
Doctor of Philosophy

Certificate:

This is to certify that the thesis entitled **Cellular trade-offs in the non-equilibrium synthesis of complex molecular information** submitted by Alkesh Yadav for the award of the degree of Doctor of Philosophy of Jawaharlal Nehru University is his original work. This has not been published or submitted to any other University for any other Degree or Diploma.

Prof. Tarun Souradeep
(Director)
Raman Research Institute
Bangalore 560 080
India

Prof. Pramod Pullarkat

Prof. Madan Rao
(Thesis Supervisors)

Declaration:

I hereby declare that the work reported in this thesis is entirely original. This thesis is composed independently by me at Raman Research Institute under the supervision of Prof. Madan Rao. I further declare that the subject matter presented in this thesis has not previously formed the basis for the award of any degree, diploma, membership, associateship, fellowship or any other similar title of any university or institution.

Prof. Madan Rao
Simons Centre for the Study of Living Machines
National Centre for Biological Sciences (TIFR)
Bengaluru 560 065
India

Alkesh Yadav

Prof. Pramod Pullarkat
Raman Research Institute
Bangalore 560 080
India

To the three amazing women in my life.

Acknowledgements

I will start this section with the disclaimer that although this thesis has my name on it, it is the result of a collective effort of a lot of people, and I can only try to acknowledge all them.

First, I would like to express my deep gratitude to my advisor Madan Rao. His imagination, creativity and deep love for science have been a constant source of inspiration for me. I will be forever indebted to him for teaching me how to do science, how to develop a problem and ask interesting questions.

Garud Iyengar has been like a second advisor to me. I have learned many technical things that I use in this thesis, namely optimization and information theory, from him. His deep knowledge of these fields and his precise communication have helped me a lot. I thank him for his patience with me.

Quentin Vagne worked closely with me during the initial part of this work. This was a very fruitful collaboration in which I learned how to not get tied down to a model and how to communicate your ideas.

I want to thank Pramod for being helpful with the thesis submission process.

I would like to thank people at the Simon center and RRI for their friendship, support and significant help. They include : Amit, Krishnan, Rahul, Saptarshi, Abhishek, Raj, Debshanker, Kabir, Shubham, Amit Singh, Amit Das, Nida, Ayan, Archisman, Nandita, Simanraj, Deepak, Sanjay, and Sreeja.

I would also like to thank the RRI and NCBS administration, especially Radha and Harini, for being very helpful.

Last,and most important, my deepest gratitude to my family - my mother, sister, wife and my grandparents for being a constant source of support for me. I am grateful to my late father for inculcating the love of science in me.

Synopsis

Cells have evolved to sense and make an internal representation of the outside environment as well as display a representation of itself to the environment. This molecular information flow from the cell to the environment and vice versa, helps the cell to infer, sense and encode molecular information leading to control and coordinated decisions. One example is the process of glycosylation, the sequential covalent attachment of sugar moieties to proteins catalyzed by a set of enzymatic reactions within the Endoplasmic reticulum (ER) and the Golgi complex. Glycans, the final products of this glycosylation assembly line, are delivered to the plasma membrane (PM) conjugated with proteins, whereupon they engage in multiple cellular functions, including immune recognition, cell identity markers, cell-cell adhesion and cell signaling. We focus on the role of glycans as markers of cell identity and tissue niche. For the glycans to play this role, they must inevitably represent a molecular code. In this thesis, we study one aspect of molecular coding, namely the *fidelity* of this molecular code generation. The Golgi complex, where glycosylation primarily takes place, consists of a stacks of flattened, membrane-enclosed compartments called cisternae. Each Golgi stack typically consists of four to six cisternae, although some unicellular flagellates can have more than 20. Maximizing the *fidelity* of this displayed information results in a trade-off among the limited resources, e.g. number, type and specificity of enzymes, accessible to the cell. These cellular trade-offs coupled with the physical forces dictate the intracellular patterning of the organelles inside cell, e.g. the size, shape and number of Golgi cisternae.

This thesis is primarily a theoretical study of the cellular trade-offs involved in high *fidelity* synthesis of *complex* glycan molecular code, which surprisingly constrains the architecture of Golgi complex, specifically the number of cisternae, and the number and specificity of glycosylation enzymes. Subsequently we explore the evolutionary consequence of enzyme specificity in a more general context.

Synopsis

The information theory problem that we address in this thesis is to find the optimal synthesis machinery, subject to the constraints of physics and biology, that produces a given *complex* target signal with high *fidelity*. Since the idea of *complexity* of a molecular code in a general sense is lacking in the conventional information theory, we explore this in the specific context of glycans. Extant glycan distributions have high *complexity*, owing to evolutionary pressures arising from (a) reliable cell type identification amongst a large set of different cell types in a complex organism, and (b) pathogen-mediated selection pressures. We estimate this glycan *complexity* from the mass spectrometry(MS) glycan data by defining complexity of a MS profile as the minimum number of components of a Gaussian Mixture Model (GMM) required to fit the given MS profile. Using this quantitative definition of *complexity* on real MS glycan data of *human* cells, *planaria* and *hydra* we demonstrate that complex organisms have complex glycan profiles. The target signal for the synthesis machinery is given by these de-noised MS glycan profiles of real cells.

The glycan synthesis machinery involves sequential chemical processing via cisternal resident enzymes and cisternal transport of the glycans. Each cisterna has a distinct chemical environment, e.g. pH, which affects the state of enzyme residing in the cisterna. The enzymes are assumed to act via a induced fit mechanism where they deform to match the substrate shape, providing enzymes a specificity towards substrate binding. The cisternal transport of glycans is unidirectional from *cis* to *trans* Golgi. The steady state *synthesized glycan distribution* is therefore a function of cellular parameters such as the number and specificity of enzymes, distribution of the enzymes across the cisternae, inter-cisternal transfer rates, and number of cisternae.

We minimize the Kullback-Leibler (KL) divergence between the synthesized distribution of glycans and the target glycan distribution over the enzyme rates, enzyme distribution and the transport rates for a fixed number of enzymes, enzyme specificity and the number of cisternae. The minimum KL divergence is used as a quantitative measure of *fidelity* of the synthesis machinery. This *fidelity* is a function of the parameters characterizing the glycan synthesis machinery, such as the number of cisternae, and number and specificity of enzymes. We analyze the trade-offs between these parameters, in order to achieve a prescribed target glycan distribution with high fidelity.

Our analysis leads to a number of interesting results, of which we list a few:

- In order to construct an accurate representation of a complex target distribution, such as those observed in real cells, one needs to have multiple cisternae or multiple enzymes. Since having more enzymes invokes a more elaborate genetic cost to the cell, the analysis provides a quantitative argument for the evolutionary requirement of multiple-compartments.

- For fixed number of enzymes and cisternae, there is an optimal level of specificity of enzymes that achieves the complex target distribution with high fidelity.

Our results imply that the pressure to achieve the target glycan profile for a given cell type, places strong constraints on the cisternal number and enzyme specificity. This would suggest that a description of the non-equilibrium assembly of a fixed number of Golgi cisternae must combine the dynamics of chemical processing and membrane dynamics involving fission, fusion and transport.

Inspired by the strong dependence of *fidelity* of synthesis on enzyme specificity, we further study the evolutionary implications of non-equilibrium driving to modulate enzyme specificity in a more general context. Proper functioning of the cell requires enzymes to discriminate its specific substrates from a multi-component mixture of thousands of different substrates that are present in the cell or the cellular compartment. Failure to do so can lead to both accumulation of wrong products and unavailability of the enzyme for the correct substrate, impairing normal cellular function. Enzymes that are too specific can also invoke costs to the cell in terms of the requirement of a large number of enzymes, less robustness to mutations and changing environments. These arguments suggest an optimal intermediate enzyme specificity and a mechanism to modulate it might have evolutionary advantages.

The enzyme substrate specificity is a result of either kinetic or energetic discrimination of substrates by the enzyme. Kinetic discrimination is based on the number and height of barriers encountered along various paths connecting one state to another in the free energy landscape. Here the enzyme substrate complexes which, on average, are separated from the enzyme by multiple high energy barriers are unfavorable as compared to complexes which are separated from the enzyme by fewer

and lower energy barriers. The energetic discrimination, on the other hand is path independent, and is based on the differences in free energies of the enzyme substrate complexes. The steady state relative concentration of products depends on the free energy of the enzyme substrate complexes, the kinetics of formation of the complexes and the nature of the non equilibrium drive.

We explore two general ways of driving the system out of equilibrium without selectively biasing the system towards a particular enzyme substrate complex : (a) a periodic drive to oscillate the states and the barriers between them (b) biasing trajectories which absorb and dissipate work at a fixed rate by a method called dynamical biasing. Within bounds, the non-equilibrium driving performs a time dependent sculpting of the free energy landscape and hence provides a control on enzyme specificity. We will subsequently do a population dynamics calculation, to study the effect of enzyme-substrate specificity on robustness of the cell to genetic mutations and adaptability of the cell to changing environments. The genotype here is defined by the the free energy landscape of the enzyme-substrate interaction which depends on the equilibrium conformation of the enzymes and the substrates, and the interactions between them. The phenotype is determined by the relative concentrations of products, and the environment is characterized by the optimal phenotype for that environment. Inheritance is subject to mutations, and the genotype to phenotype map is stochastic and depends on the the non-equilibrium drive. A specific instance of this general calculation is the improvement of the *fidelity* of glycan synthesis if a cisternal control of enzyme specificity is allowed.

To summarize, in this thesis we provide an information theoretic language to analyze the cellular trade-offs involved in the synthesis of glycan information to show that the functional requirement of generating a high fidelity glycan code puts constraints of the number of Golgi cisternae and number and specificity of enzymes. In future we plan to invoke error correcting mechanisms that might be involved in the generation of the glycan code.

Synopsis

Prof. Madan Rao
Simons Centre for the Study of Living Machines
National Centre for Biological Sciences (TIFR)
Bengaluru 560 065
India

Alkesh Yadav

Prof. Pramod Pullarkat
Raman Research Institute
Bangalore 560 080
India

Publications

- Alkesh Yadav, Quentin Vagne, Garud Iyengar, Pierre Sens, and Madan Rao, *Glycan processing in the Golgi: optimal information coding and constraints on cisternal number and enzyme specificity*, eLife 2022;11:e76757

Contents

Chapter 1

Introduction

1.1 Cell as an information processing entity

Living systems constantly interact with their environment to perform necessary biological functions, e.g. taking in nutrients and external chemical signals [1]. The survival of the organism critically depends on the ability to communicate reliably with the its environment. Complex organisms have a dedicated neuronal system that carries out this functionality at the scale of the organism. However, at the scale of the cell, both within complex organisms as well as in simpler unicellular organism, communication with the environment is carried out by sensing chemical signals from the environment, e.g. in chemotaxis [2], and displaying chemical signals to the external environment, e.g. in immune response [3]. The information flow [4] from the cell to the environment and vice versa, carried out by molecules, is constrained by the laws of thermodynamics and other resource constraints. Cells which maximize this information flow given the physical constraints can get selective advantages at performing a critical function, like finding food in chemotaxis and increase the fitness of the organism.

One way of maximizing the information flow between the cell and the environment is by creating compressed representations. The internal state of the cell at anytime is the detailed chemical composition in the cell, the concentrations of all the molecules and there spatial distribution. This description is way too detailed for any external agent to sense or detect and therefore does not have any functional relevance. Therefore, the functionally relevant part of the internal state should be represented, in a reliable and compressed way, with a smaller subset of molecules which is accessible

to the external environment. Similarly the cell must create a functionally relevant internal representation of the environment. Maximizing the information flow by creating reliable and compressed representations results in a trade-offs among the limited resources, e.g. number of enzymes, accessible to the cell. These trade-offs on resources coupled with the physical and chemical forces dictate the organization inside the cell - intracellular patterning [5] of the organelles.

The classical information theory [6], pioneered by Shannon, deals with the notion of compression for a sequence of immutable symbols. However, the symbols in biology are molecules which carry out certain functions. Therefore, quantification of biological information requires a context, which is set by the functions of the molecule in the cell [7]. This extended information theoretic framework leads to optimization principles which sheds light on the resource and physical trade-offs that might be operational in the cell.

In this thesis, we focus on the information processing in the Golgi complex, a cellular organelle, which consists of a collection of flattened, membrane-enclosed compartments called cisternae. Each Golgi stack typically consists of four to six cisternae, although some unicellular flagellates can have more than 20 [1]. The Golgi complex maintains a stable structure despite being subjected to a constant non-equilibrium flux of vesicles carrying cargo from the Endoplasmic Rerticulum (ER). It is the site of glycosylation [8], a process of sequential attachment of sugar moieties to proteins synthesized in the ER by a set of enzymatic reactions. Glycans, the final products of this glycosylation assembly line, are delivered to the plasma membrane(PM) conjugated with proteins, whereupon they engage in multiple cellular functions, including immune recognition, cell identity markers, cell-cell adhesion and cell signaling [9]. We study the resulting cellular trade-offs from glycans as the carrier of information.

We start with a brief introduction to the classical information theory and define the basic information theoretic quantities that we will use in the thesis later. It is heavily borrowed from the classic textbooks [6, 10]. We follow this up with a section on the use of information theory in biology which sets up the broader context of this work. We provide introduce a essential introduction to thermodynamic and Markov processes in the following section. At last, we give a brief phenomenology of the biological system of interest of this thesis - the Golgi complex and glycosylation.

1.2 Foundations of information theory

The information content in a sequence of random variables in the context of communication is based on asking two basic questions :

1. Data compression: How much can we "compress" a sequence of random variables?

2. Data transmission: What rate can we transmit random variables through a noisy channel and expect to recover them back perfectly?

The answer to both these questions is in the notion of uncertainty in a random variable. We provide the following example to give an intuitive feel for what is meant by 'information', 'uncertainty' and 'compression' before formally defining them.

Suppose we have a set of N hypothesis $\mathcal{Y} := \{y_1, \ldots, y_n\}$ and their associated probabilities $\mathcal{P} := \{p_1, \ldots, p_n\}$ of correctly explaining a particular phenomenon. Our task is to identify the correct hypothesis from the set by doing the least number of experiments. Initially, lets assume that each experiment can only tell whether a single hypothesis is correct or not. If all the hypothesis are equally probable then we can randomly pick a hypothesis to experiment for and rule it out until we find the correct hypothesis. On the other hand, if all the hypothesis are not equiprobable it makes sense to first test for the most probable one, and go to the next most probable one and so on until we hit the correct hypothesis. The average of experiments needed in the first case will be higher than in the second case. What happens if we are allowed to pool hypothesis in experiments, what should be the best pooling strategy? Again, lets first start with the case when the hypothesis are equiprobable. The best pooling strategy is going to when we divide the two halves $N/2$, then $N/4$ and so on. The optimal pooling strategy divides the hypothesis in equiprobable groups at each stage.

Here we are trying to elucidate the fact that a given probability distribution is associated with a certain amount of uncertainty (measured in the previous example by the number of experiments required). With this in mind we now formally define some important information theoretic quantities and state their mathematical properties.

Definition 1 (Shannon information content) *Ensemble X is a triple $(x, \mathcal{A}_X, P_A)$, where x is the outcome which takes values in the discrete set of symbols, called alphabet, $\mathcal{A}_X = \{a_1, a_2, \ldots, a_I\}$ with probabilities $\mathcal{P}_X = \{p_1, p_2, \ldots p_I\}$. The Shannon information content of the the outcome x is defined as*

$$h(x = a_i) = -\log_2 p(x = a_i) = -\log_2 p_i \tag{1.1}$$

This function has the following nice properties which align with our intuitive notion of uncertainty and quantifies our surprise given an outcome.

1. $h(x) \geq 0$ with $h(x) = 0$ only if the outcome is a certainty ($p(x) = 1$).
2. $h(x)$ of a less probable outcomes is higher than more probable outcomes capturing the intuitive notion that a unlikely outcome carries more information than a likely outcome.
3. The information content of a composite event $x' = x_1$ OR x_2 is given by $h(x') = h(x_1) + h(x_2)$.

Definition 2 (Shannon entropy) *The average information content of an ensemble, called Shannon entropy, is given by*

$$H(X) \coloneqq -\sum_{x \in \mathcal{A}_\mathcal{X}} p(x) \log_2 p(x) \tag{1.2}$$

The following properties of Shannon entropy follow from the definition

1. $H(X) \geq 0$
2. $H(X)$ is maximum for ensemble with equiprobable events. $H(X) \leq \log_2 |\mathcal{A}_\mathcal{X}|$
3. The entropy of two independent ensembles, X and Y, $H(X, Y) = H(X) + H(Y)$, for any other pair of ensembles X and Y, $H(X, Y) \leq H(X) + H(Y)$
4. Additivity of composite events

These properties can conversely be used to uniquely define the Shannon entropy function [11]. We will show that Shannon entropy is approximately equal to the logarithm of the number of typical values that the variable can take.

We now describe a key property of a large sequence of independent identically distributed (i.i.d.) random variables which allow us to answer the two basic questions asked in the beginning of the section. Informally, almost all large sequences of i.i.d. random variables belong to a subset, called the typical set, of all possible sequences. The probability of finding a sequence not in the typical set is almost zero as the length of the sequence tends to infinity. We will show that the size of this typical set is related to the Shannon entropy. We can therefore store just the typical sequences rather than all sequences with asymptotically zero probability of making an error.

Let $\mathbf{x_N} = (x_1, x_2, \ldots x_N)$ be a sequence of N identical independently distributed (i.i.d.) random variables drawn from the ensemble $X = (x, \mathcal{A}_\mathcal{X}, P_X)$. We now define a set called the $\epsilon-$ typical set which contains almost all sequence of i.i.d. random variables coming from the ensemble X:

Definition 3 ($\epsilon-$typical set) *For some $\epsilon > 0$, a sequence, $\mathbf{x_N}$, of iid random variables from an ensemble X with entropy $H(X)$ is $\epsilon-$typical if*

$$\left| -\frac{1}{N} \log P(\mathbf{x_N}) - H(X) \right| \leq \epsilon \tag{1.3}$$

The set of all such sequences, $\mathcal{A}_\epsilon^n$ is called the $\epsilon-$ typical set

This set contains all the 'typical' sequences of the i.i.d. random variables We now state the asymptotic equipartition (AEP) theorem which forms the basis of the Shannon source coding theorem.

Theorem 1 (AEP) *Let $\mathcal{A}_\epsilon^N$ be an $\epsilon-$ typical set, then*

1. $\lim_{n\to\infty} P(\mathbf{x_N} \in \mathcal{A}_\epsilon^N) = 1$
2. *For large enough N,* $\exp(N|H(X) - \epsilon|) \leq |\mathcal{A}_\epsilon^N| \leq \exp(N|H(X) + \epsilon|)$
3. *For any $\mathbf{x_N} \in \mathcal{A}_\epsilon^n$,* $\exp(-N|H(X) + \epsilon|) \leq P(\mathbf{x_N}) \leq \exp(-N|H(X) - \epsilon|)$

Loosely, AEP states that the total number of sequences of i.i.d. random variables asymptotically approaches $\exp(NH(X))$ as the length of the sequence goes to infinity. Asymptotically all the sequences are typical and the size of the typical set is given by the Shannon information $\exp(NH(X))$.

1.2.1 Data Compression

Suppose we have several N base pair long DNA sequences and the probability of occurrence of nucleotides A,T,C and G in these sequences is p_A, p_T, p_C and p_G respectively. How can we store these sequences in a database using the least amount of memory? Clearly, we need to define a one to one mapping, called coding scheme, between A,T,C,G to some other symbols, e.g. A is coded as 0, T is coded as 1, C is coded 01 and G is coded as 010. The length of the coded sequence is different from the length of the original sequence. If we design a coding scheme which codes the more probable nucleotides with a shorter character and the less probable nucleotides with a longer character, then we can make the expected length of a long sequence smaller than the original length. The Shannon's source coding theorem shows that there exists a coding scheme in which the average length is less than the original length and is bounded by the Shannon entropy.

Now we formally define codes and then state the Shannon source coding theorem.

Definition 4 (Code) *A code $C(X)$ for an ensemble X is a mapping from the set of alphabets, $\mathcal{A}_X$ to $\{0,1\}^+$, $c(x)$ denotes the codeword corresponding to the outcome x and $l(x)$ denotes its length.*

Average length of the code is $L(X) = \sum_x p(x)l(x)$. A code is called uniquely decodable code if and only if for $x \neq y$, $c(x) \neq c(y)$ $\forall x, y \in \mathcal{A}_{\mathcal{X}}$. A code in which no codeword is a prefix of any other codeword is called a prefix code. Prefix codes are self punctuating or instantaneously decodable without looking ahead at the subsequent codeword. Note that the codeword lengths of a uniquely decodable code $C(X)$ over the binary alphabet must satisfy the Kraft's inequality - $\sum_{i=1}^{|\mathcal{A}_{\mathcal{X}}|} 2^{-l_i} \leq 1$.

Theorem 2 (Shannon's source coding theorem) *For an ensemble X, there exists a prefix code $C(X)$ with average length satisfying*

$$H(X) \leq L(C, X) \leq H(X) + 1 \tag{1.4}$$

Shannon's source coding theorem is based on the idea that we can define a coding scheme which shortens the length of typical sequences at the expense of making the atypical sequences longer and since by the AEP theorem almost all long sequences

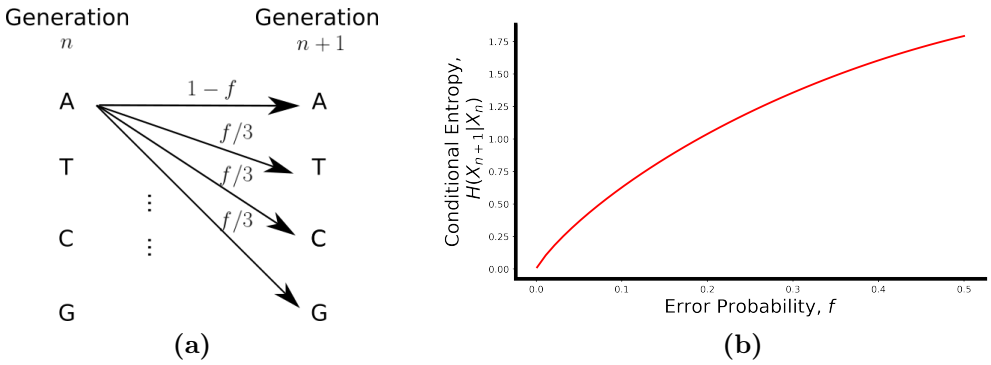


Figure 1.1. (a) DNA replication as an information channel. f is the probability of error, (b) The conditional entropy between the next and the present generation. The uncertainty increases as the probability of error increase.

are typical we end up reducing the expected length. Note that Shannon's theorem only gives the theoretical bound on the minimum length (memory) required to store a sequence of random variables coming from a probability distribution but it does not say anything about how to achieve that limit. Finding out the coding scheme that reaches close to the Shannon limit is part of coding theory and there is no formal way of finding the optimal coding scheme. Another point to note is that Shannon's theorem is about lossless compression, which requires that there is a codeword for every symbol in the alphabet.

1.2.2 Data transmission

We now analyze the transmission of information over space or time through a noisy channel. As a concrete biological example, we look at information transfer from one generation to the next during the process of DNA replication. The replication process is noisy due to the possibility of chemical errors. Let the probability of this error be f, therefore all nucleotides A,T,C and G are correctly replicated in the next generation with a probability $1 - f$ (See Figure 1.1). We want to quantify the information transferred during such process and can we have perfect information transfer in presence of noise?

The conditional probabilities for this channel can be written as:

$$p(X_{n+1}|X_n) = \begin{cases} 1-f & X_{n+1} = X_n \\ f/3 & X_{n+1} \neq X_n \end{cases} \tag{1.5}$$

Here $X_i \in \{A, T, C, G\}$. We can quantify the reduction in uncertainty of the random variable X_{n+1} on the revelation of another random variable X_n by calculating the Shannon entropy of the conditional probability

$$H(X_{n+1}|X_n) = \sum_{i=1}^{4} p(X_n) \left[p(X_{n+1}|X_n) \log_2 p(X_{n+1}|X_n) \right]$$

Assuming all base pairs are equiprobable, the conditional entropy of this system comes out to be $H(X_{n+1}|X_n) = -(1-f)\log_2(1-f) - f\log_2(f/3)$. The graph in Figure 1.1 shows that the conditional uncertainty of the next generation is an increasing function of f implying that the present generation gets less and less informative about the next generation on increasing the noise. We now define a related quantity, the mutual information, which measures the dependence of two random variables. It is the reduction in uncertainty of one random variable on the revelation of the other random variable.

Definition 5 (Mutual information) *Mutual information, I, between two ensembles X and Y is given by*

$$\begin{aligned} I(X,Y) &:= -\sum_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} \\ &= H(X) - H(X|Y) = H(Y) - H(Y|X) \end{aligned} \tag{1.6}$$

Mutual information is a symmetric quantity and captures the dependence of the random variable X on Y and vice versa. If X and Y are independent then $H(X|Y) = H(X)$, $H(Y|X) = H(y)$ and $I(X,Y) = 0$, on the other hand if X, Y are completely dependent, X is a function of Y, then $H(X|Y) = H(Y|X) = 0$ and $I(X,Y) = H(X) = H(Y)$. Note that mutual information is a nonlinear measure of relatedness and can capture the relation between X and Y which are highly informative but not detected by linear correlations [12]. Figure 1.2 illustrates the relationship between

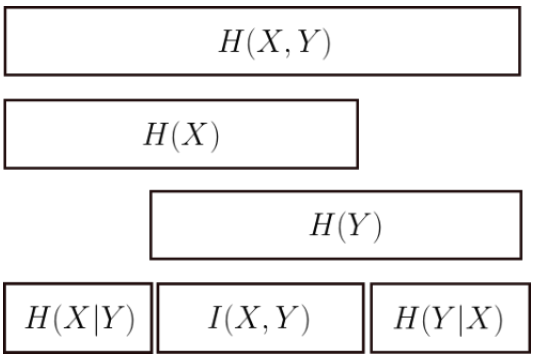


Figure 1.2. Relationship between entropies, $H(X)$ $H(Y)$, conditional entropies, $H(X|Y)$ and $H(Y|X)$, joint entropy, $H(X,Y)$, and the mutual information, $I(X,Y)$, between the set of random variables X and Y.

entropies, $H(X,Y)$, $H(X)$, $H(Y)$, $H(X|Y)$, and the mutual information, $I(X,Y)$.

The mutual information between the next generation and the present generation in the last example decreases on increasing the noise indicating a loss of information during the transmission process due to noise. There are two ways of reducing the loss of information in the channel - (i) we either reduce the noise in the replication process using biophysical means like kinetic proofreading or (ii) we can introduce an encoding-decoding scheme which adds redundancies to help identify and fix errors that might arise during the transmission.

Here we focus on the later, in the previous example, see Figure 1.1, instead of just transmitting the nucleotides (A,T,C,G) directly we first encode them by a rule that introduces redundancies, then transmit the encoded message through the channel and later decode the message. The encoded message can still be corrupted by the noisy channel but if the corruption is below a threshold will still be able to recover the message correctly. A simple example of a encoding-decoding scheme can be to encode A as AAA, T as TTT, C as CCC and G as GGG. Here for every A we will transmit A three times, similarly for the rest (T,C,G). While decoding the transmitted message we wait for three transmissions and then take the majority as the message. This simple scheme can reduce the probability of making errors by $1/f$ but the cost we have to pay is that we use the channel three times to transmit a single message (nucleotide). Shannon worked out that minimum amount of redundancies we need to introduce to ensure lossless transmission of information through a noisy channel and surprisingly he found that the amount of redundancies that we need to

add are finite to ensure lossless transmission. We can therefore transmit losslessly over a noisy channel at a non-zero rate.

We now formally define what we mean by information channel, an encoder - decoder, rate of transmission and then state the Shannon channel theorem.

Definition 6 (Channel) *A discrete memoryless channel Q is characterized by an input alphabet $\mathcal{A}_\mathcal{X}$, an output alphabet $\mathcal{A}_\mathcal{Y}$ and a transition probability matrix $Q(y|x)$ for $x \in \mathcal{A}_\mathcal{X}$ and $y \in \mathcal{A}_\mathcal{Y}$*

Definition 7 (NK Block code) *A (N, K) block code for channel Q is a list of $S = 2^K$ codewords*

$$\{\mathbf{x^{(1)}}, \ldots, \mathbf{x^{(2^K)}}\} \quad \mathbf{x^{(s)}} \in \mathcal{A}_\mathcal{X}^\mathcal{N}$$

each of length N. The encoder encode a signal $s \in \{1, \ldots 2^K\}$ as $\mathbf{x^{(s)}}$. The rate of code $R = K/N$. The decoder for an (N, K) block code is a mapping from the set of N length string of channel output $\mathcal{A}_\mathcal{Y}^\mathcal{N}$ to a codeword label $\hat{s} \in \{0, 1, \ldots, 2^K\}$.

The probability of block error $p_B = \sum_{s_{in}} P(s_{in})P(s_{out} \neq s_{in}|s_{in})$. The maximum probability of error $p_{BM} = \max_{s_{in}} P(s_{out} \neq s_{in}|s_{in})$.

Theorem 3 (Shannon's channel coding theorem)

1. *For every discrete memoryless channel, the channel capacity $C = \max_{p(x)} I(X;Y)$ has the following property. For any $\epsilon > 0$ and $R < C$, for large enough N, there exists a code of length N and rate $\geq R$ and a decoding algorithm, such that the maximal probability of error $p_{BM} < \epsilon$.*

2. *If probability of bit error p_b is acceptable, rates up to $R(p_b)$ are achievable, where $R(p_b) = \frac{C}{1-H_2(p_b)}$*

3. *For any p_b, rates greater than $R(p_b)$ are not achievable.*

It is a remarkable fact that one can transmit information without error along a channel that has a nonzero noise level provided that the rate of transmission does not exceed this capacity. Note that, the channel capacity is obtained by maximizing over all possible source (input) probability distributions and is therefore just a function of the channel, i.e the transition probabilities. Shannon's channel coding theorem again

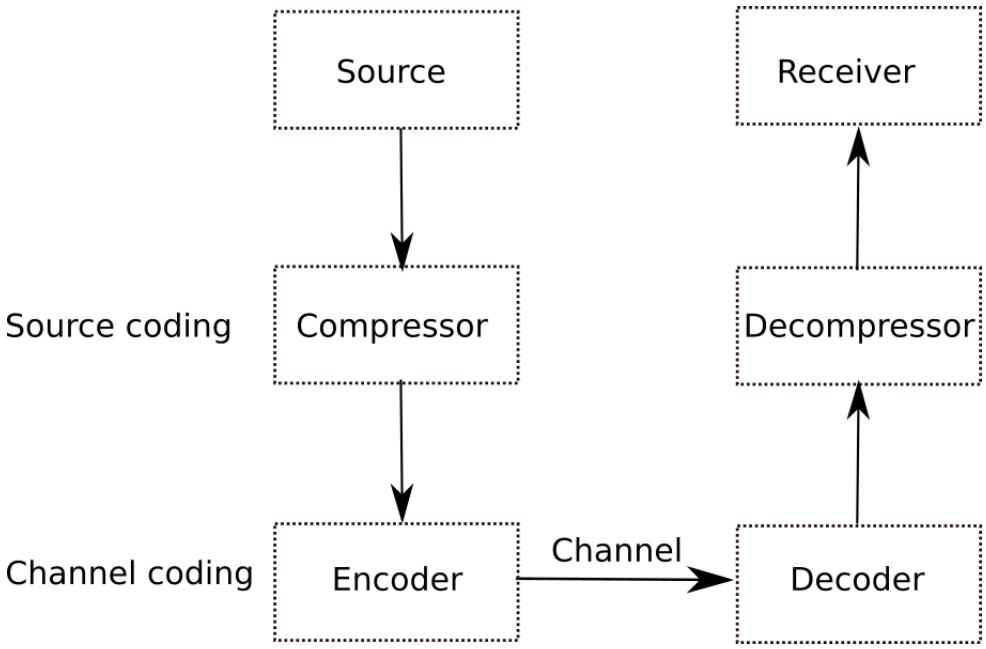


Figure 1.3. Schematic of a communication system. A source message is first compressed using some coding scheme which reduces the size of the message. It is then endoded into the transmitted signal using an encoding scheme which adds redundancies so that the message can be transmitted through a noisey channel without errors. The reciever then performs the inverse operations of decoding and decompressing to receive the message.

only gives the theoretical bounds on the capacity of lossless transmission by a channel but does not say anything about how to design the encoder-decoder scheme that achieves this capacity. We note that actually finding the capacity of a channel, given a detailed model of the input-output relation and noise as represented by $P(y|x)$, can be challenging and, only simple models, such as y being a linearly filtered version of x with added Gaussian noise, are tractable. The complete communication system can be schematically represented as shown in the Figure 1.3. Shannon's theorems give the fundamental theoretical bounds for lossless compression and transmission of data.

There is a formalism in information theory which allows deals with lossy compression or transmission of information - the rate distortion theory. Let $\mathcal{X}$ be an ensemble defined by random variable X, alphabet $\mathcal{A}_\mathcal{X}$ and probabilities $p_X(x)$. As the size of the alphabet, $|\mathcal{A}_\mathcal{X}|$, increases and the input probability is well distributed over the alphabet $\mathcal{A}_\mathcal{X}$, both lossless compression and transmission becomes more demanding. This is especially relevant in biology where lossless compression and transmission of information maybe too costly and unnecessary for the cell. From this detailed ensemble we want to go to a smaller ensemble, $\mathcal{Y}$ defined by random variable Y, alphabet $\mathcal{A}_\mathcal{Y}$, probabilities $p_Y(y)$ and $|\mathcal{A}_\mathcal{Y}| < |\mathcal{A}_\mathcal{X}|$, while maintaining a certain level of functionality. The functionality is measured by a distortion function, denoted by $d(x, y)$ and defined on $|\mathcal{A}_\mathcal{X}| \times |\mathcal{A}_\mathcal{Y}|$ space, which measures the distance between X and Y in some relevant sense. Since both $x \in \mathcal{A}_\mathcal{X}$ and $y \in \mathcal{A}_\mathcal{Y}$ are random variables we are interested in the average distortion, $\langle d(x, y)\rangle = \sum_{x,y} p(x, y)d(x, y)$. The average distortion is a measure of how badly we are doing on a particular function by moving from the ensemble $\mathcal{X}$ to $\mathcal{Y}$ and we want this average distortion to be less than certain threshold, D. These considerations can be formally represented by the following optimization problem

$$\begin{aligned} R(D) = \quad & \min_{p(y|x)} I(X, Y) \\ & \text{s.t. } \langle d(x, y)\rangle \leq D \end{aligned}$$

which can be equivalently written as

$$R(\lambda) = \min_{p(y|x)} I(X, Y) + \lambda \langle d(x, y)\rangle \tag{1.7}$$

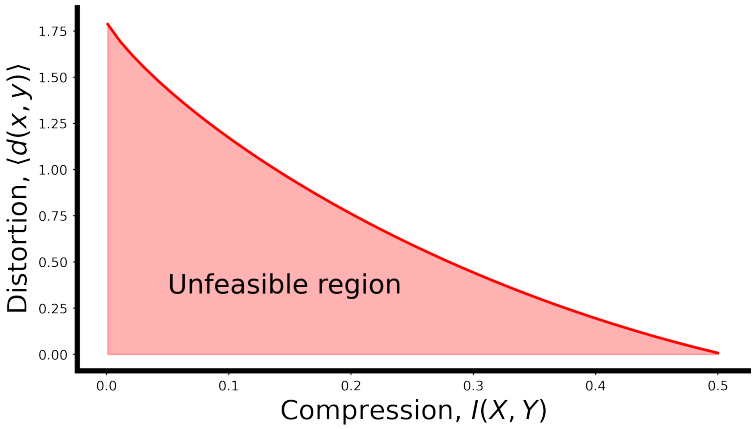


Figure 1.4. An illustration of rate distortion function, $R(\lambda)$. The each point on the red curve is the optimal solution for a given value of λ. The curve represents the minimum distortion acheivable for a particular compression.

Minimizing $I(X, Y)$, the mutual information between X and Y, compresses X to Y and minimizing the average distortion, $\langle d(x, y)\rangle$, increases the functionality of this change of ensemble. The optimization, therefore, compresses the information while maintaining the functional relevance, λ decides the weight of these two competing forces. The balance between these two factors, compression and the functional relevance, is given by the rate distortion curve, $R(\lambda)$ (See Figure 1.4 for a typical curve). The region in the figure above the rate distortion curve is unachievable. and this can potentially lead to physical or biological constraints in biological systems [**?**]. One practical limitation for the use of rate distortion framework in biology is the difficult of coming up with a with relevant distortion function. In [13], proposed a Information Bottleneck(IB) principle to circumvent the requirement of a distortion function by having a target ensemble which represents the function. We discuss more about IB and use it in our calculation in Chapter 2 of the thesis.

We now define one more information theoretic quantity that we use in this thesis to measure the similarity of two probability distributions.

Definition 8 (KL divergence) *The KL Divergence between two probability dis-*

tribution is given by

$$D_{KL}(p||q) \coloneqq \sum_x p(x) \log_2 \frac{p(x)}{q(x)} \tag{1.8}$$

The KL divergence has the following properties:

1. $D_{KL}(p||q) \geq 0$ This is sometimes called the Gibbs inequality, $D_{KL}(p||q) = 0$ iff $p = q \; \forall x$
2. in general $D_{KL}(p||q) \neq D_{KL}(q||p)$
3. The entropy is a convex function, convex functions satisfy the Jenson inequality which is very useful in proving properties of information theoretic quantities. For a convex function $f(\mathbf{x})$, $f(\mathrm{E}(x)) \leq \mathrm{E}(f(x))$, here E denotes the expectation value. This inequality directly follows from the definition of convex functions.

1.3 Information theory in biology

Living systems process information and perform computations at various spatial and temporal scales, ranging from microns in bacterial chemotaxis to meters in the neural system of large organisms. The mechanisms that carry information are also widely different ranging from electrical signals in neuronal systems to molecules and mechanical signals in cellular information processing. It is attractive to discuss information transmission in these wide variety of biological cases in the same units (bits) using the unifying language of information theory. The framework described in the previous section quantified the information carried by a probability distribution over immutable symbols and described the theoretical limits of compression and transmission of that information. While this framework has been greatly successful in communication applications, describing biological system using this framework requires more care and meaningful extensions.

In engineering applications, the main concern is devising coding schemes, i.e., algorithms that transform inputs x into messages to be sent through the channel (and likewise recover x from the channel output), so that information can be transmitted

over noisy channels as close to capacity as possible using bounded processing resources. Biological information processing systems, on the other hand, are driven out of equilibrium by a continuous consumption of energy and they are shaped by evolution to operate robustly in the presence of thermal, active or chemical noise [2, 14], be energy and resource efficient [15], exhibit parametric robustness [16], i.e. do not require precise tuning of parameters, and be evolvable [17], i.e. can adapt to changing environments. Capacity attaining encoding-decoding schemes, therefore, may not be relevant in biological information processing systems. However, optimization principles arising from information theoretic considerations (like maximizing mutual information between two random variables) in biological systems have resulted in physical constraints on the system. One example of physical resources limiting the information transmission is the synthesis of transcription factor proteins (TFs) that bind to DNA and influence the rate at which encoded information is read out to make other proteins; here the transcription factor concentration is the input, and the resulting protein concentration is the output. The number of molecules of TF can limit the maximum amount of information that can be transmitted through this channel [18].

Quantifying information in biological systems in a meaningful way is critical in arriving at relevant optimization principles which put physical constraints on the system. Inside the cell, information is carried by molecules which are not like the immutable symbols of the classical information theory but are involved in carrying out certain functions, have different life spans, utility and production cost to the cell. One possible direction in quantifying biological information is the idea of functional complexity given by Jack Szostak and coworkers [7]. The idea is based of different mRNA sequences resulting in proteins that after folding perform the same function and therefore are functionally equivalent. This additional redundancy can be used to further compress the sequence beyond the Shannon entropy giving rise to a new limiting complexity, called functional complexity, which is dependent on the function that the sequence is performing.

Another way of formally using function to come up with relevant information theoretic optimization principles is the framework rate-distortion theory described in the previous section. If we can come up with a relevant distortion measure for a biological information processing system, the framework allows for lossy compression

of data while keeping functionality above a threshold. We now give a few instances of where a biologically important function can be replaced with an abstract information theoretic optimization principle. that can put constraints the system.

In [19], the authors analyzed the problem of sensing the time-dependent ligand profile outside the cell by a collection of distributed and mobile sensors on the surface of the cell to find the optimal placement of the sensors. There are two competing objectives : to faithfully read the ligand concentration at a given position requires the sensor to stay there and take multiple measurements of the ligand in time. But this sampling requires one to cluster the receptors at some position leaving the ligand at other locations undetected by the sensors. They found that depending on the sensor concentrations and the ease of clustering there are three phases of optimal sensor placement - (a) at low sensor concentrations the receptors diffuse freely to sense the ligand signal (b) at high sensor concentrations but low ease of clustering, the optimal distribution is sensors fixed on a lattice (c) at high sensor concentrations but high ease of clustering, the optimal distributions is some sensors randomly forming a cluster and some sensors diffusing freely. All these phases are found in real biological systems with consistent sensor concentration and ease of clustering.

In [20], the authors show that sensing the ligand(morphogen) outside to infer cells position inside the development tissue accurately requires two kind of receptors (specific and non specific to the ligand) and a negative correlation, due to a feedback mechanism, between the bound specific and bound non-specific receptor. This calculation is an instance of optimizing over the channel properties to achieve better functionality.

One roadblock in applying information theory to biology is the difficulty of quantifying real information flows in biological systems. Estimating information theoretic quantities, like mutual information, is extremely difficult because they require the knowledge of the whole probability distribution, whereas most of the time we have access to only a small number of samples coming from that distribution. The requirement of high quality data, therefore, has pushed for both better experimental and statistical techniques (See [4] for a review).

Information theory is also used as a tool for the analysis of biological data, e.g. the use of maximum entropy models to infer the underlying Hamiltonian that describes

correlations observed in experiments [21], and using mutual information to detect relatedness of two experimentally observed variables that is not captured by linear correlations [12].

In the next section we will discuss the physical constrained put on by thermodynamics on the molecules that carry biological information.

1.4 Physics of molecular information processing systems

The link between information and thermodynamics goes back to the idea of Maxwell's demon which revealed the relationship between entropy and information. It demonstrated that, by using information, one can relax the restrictions imposed by the second law on the energy exchanged between a system and its surroundings [22]. Later, Rolf Landauer pointed out that erasure of information is necessarily a dissipative process. His insight is that erasure always involves the compression of phase space, and so is irreversible [22]. Information manipulations such as measurement, erasure, copying and feedback can be thought of as physical operations with a thermodynamic costs. Therefore, thermodynamics provides constraints on the information processing capabilities of a physical system. Information in biology is carried by molecules, which are again subjected to the laws of thermodynamics. We are primarily interested in the thermodynamics of these molecules and how it constraints the information processing capabilities.

Thermodynamics describes the properties of macroscopic equilibrium systems in form of thermodynamic laws. The first law of thermodynamics is a statement about the conservation of energy and the second law of thermodynamics places constraints on what thermodynamic processes are physically realizable: only those that increase entropy. Emergence of structure and in particular of life, the most complex structure we know of, seems contradictory in the face of the second law, but these systems are not equilibrium systems and are kept out of equilibrium by a continuous influx of energy.

Here we provide a brief description of non-equilibrium thermodynamics [23] and the Markov formalism [24].

1.4.1 Non-equilibrium Thermodynamics

We express the change in entropy $dS = d_eS + d_iS$, as change due to equilibrium (reversible) exchange of entropy to the environment and irreversible entropy production. Irreversible process are caused by dissipative thermodynamic forces driving thermodynamic flows. For example, concentration gradient causing flow of matter or temperature gradient causing flow of heat. We can write the irreversible entropy production as follows

$$d_iS = \sum_k F_k dX_k \tag{1.9}$$

Here F_k is the generalized thermodynamic force and dX_k is the thermodynamic flow. The equilibrium flow of entropy can be written as

$$Td_eS = dU - dW - \sum_k \mu_k d_eN_k \tag{1.10}$$

Here T is the temperature, dU is the change in internal energy, dW is the work done, μ is the chemical potential and d_eN_k is the reversible exchange of particle. Note that, $d_eS = 0$ for a closed system which does not exchange matter or energy from the surrounding. A stronger form of the second law of thermodynamics for a system with many subsystems can be stated as $d_iS = d_iS^{(1)} + d_iS^{(2)} + \ldots + d_iS^{(r)} \geq 0$ and each $d_iS^{(k)} \geq 0 \; \forall \; k$.

When a system is isolated, $d_eS = 0$, the entropy of the system will continue to increase due to irreversible processes and reach the maximum possible value, the state of thermodynamic equilibrium. In the state of equilibrium, all irreversible processes cease. When a system begins to exchange entropy with the exterior, then, in general, it is driven away from equilibrium and the entropy-producing irreversible processes begin to operate. The exchange of entropy is due to the exchange of heat and matter. The entropy flowing out of the system is always larger than the entropy flowing into the system, the difference arising due to entropy produced by irreversible processes within the system. Systems that exchange entropy with their exterior do not simply increase the entropy of the exterior, but may undergo dramatic spontaneous 'self-organization'. The irreversible processes that produce entropy create

these organized states. Such self-organized states range from convection patterns in fluids to life. Irreversible processes are the driving force that creates this order.

We now describe the thermodynamics of chemical systems. For a general chemical reaction

$$a_1A_1 + a_2A_2 + \ldots + a_nA_n \underset{K_r}{\overset{K_f}{\rightleftharpoons}} b_1B_1 + b_2B_2 + \ldots b_nB_n \tag{1.11}$$

The entropy production for this chemical system can be written as

$$dS = dS_e + dS_i = \frac{dU - dW - \sum_k \mu_k d_e N_k}{T} - \sum_k \frac{\mu_k}{T} d_i N_k \tag{1.12}$$

Here, $dN_k = d_eN_k + d_iN_k$, the change in number of molecules can be divided into the reversible exchange of matter with the environment and d_iN_k is the irreversible change. The entropy production for a closed system is given by $d_eN_k = 0$ and $d_iN_K =$

$$\frac{d_iS}{dt} = \sum_k \frac{\mu_k}{T} \frac{d_iN_K}{dt} \tag{1.13}$$

Almost no chemical system is in equilibrium due to constant influx of matter and energy from an external source but almost all systems are in “local" equilibrium [23]. Local equilibrium is based on the idea that at sufficiently small length-scales, the timescale of equilibrium relaxation is much faster than the timescale of non-equilibrium drive. This allows us to meaningfully assign a temperature and other thermodynamic variables to every elemental volume. Thermodynamic relations are valid for the thermodynamic variables assigned to the elemental volume. For systems in local equilibrium, the intensive thermodynamic variables like temperature, chemical potential and pressure become a function of space and time - $T(x,t)$, $\mu(x,t)$ and $p(x,t)$ respectively, and the extensive thermodynamic variables like entropy, internal energy and particle number are replaced by their densities - $s(x,t), u(x,t), n(x,t)$ respectively. The extensive quantities obey conservation equations like the one described below

$$\frac{\partial s}{\partial t} + \vec{\nabla} \cdot \boldsymbol{J_s} = \sigma \tag{1.14}$$

Here $\boldsymbol{J}_s$ is the entropy current and σ is the local entropy produced at that position

due to irreversible processes, the equation represents the idea that local entropy can change by the flow of entropy from the surrounding into the infinitesimal volume and the local irreversible entropy production. Similar equation can be written for other extensive thermodynamic quantities like energy and number density.

Chemical systems can be formally described by Markov systems and the thermodynamics of chemical reactions can be extended for a general Markov system. We give a brief overview of Markov systems and their thermodynamics in the next section. The presentation is based on [24, 25]

1.4.2 Markov systems

Markov systems do not have memory and the stochastic evolution in future only depends only on the current state and not on the past. They are described by the general Chapman -Kolmogrov equation

$$p(x_n;t_n) = \int p(x_n|x_{n-1})p(x_{n-1}|x_{n-2})\dots p(x_2|x_1)p(x_0;t_0)dx_0\dots dx_{n-1} \tag{1.15}$$

For discrete random variables, the differential form of the Chapman-Kolmogrov equation gives rise to the Master equation which describes the dynamics of jumps between the states of the random variable.

$$\frac{dp_i}{dt} = \sum_j W_{i,j}p_j - W_{j,i}p_i \tag{1.16}$$

Here p_i is the probability of occupancy of the $i-$th state and $W_{i,j}$ is the transition rate from the j-th to i-th state. Note that the matrix W is a stochastic matrix with column sum($\sum_j W_{ij} = 0$) equal to zero, which preserves the normalization. Such systems can be represented on a graph with nodes being the states and the edges being the transition rates. We can solve the master equation using the method of generating functions which turns this system of ordinary differential equations to a partial differential equation. The steady state can be obtained by the null space of W and using $\sum_i p_i = 1$ or by a graphical method described in [25]. The net current flowing through two nodes i and j is given by $J_{ij} = W_{ij}p_j - W_{ji}p_i$. In the special condition of detailed balance, which correspond to equilibrium systems, the

net current is zero($W_{ji}p_i = W_{ij}p_j$). The entropy production for a general discrete Markov system [25] is given by

$$\frac{dS}{dt} = \sum_{\text{all edges}} J_{ij} \log \frac{W_{ij}p_j}{W_{ji}p_i} = \sum_{\text{all edges}} J_{ij} \left(\log \frac{W_{ij}}{W_{ji}} + \log \frac{p_j}{p_i} \right) \tag{1.17}$$

Clearly, entropy production is always positive, $\dot{S} \geq 0$, entropy production is zero if and only if the system is in detail balance. Entropy production is the energy requirement to make the system stay out of equilibrium. An application of this formalism in biology is kinetic proofreading [14] where the non-equilibrium driving leads to better discrimination by the enzyme between a right and a wrong product. The manipulation of the flux in such driven biological Markov system can lead to interesting speed, accuracy and energy trade-offs [15].

For a continuous Markov systems, the expansion of differential form of Chapman-Kolmogrov equation results in the Fokker-Planck equation

$$\frac{\partial p(x)}{\partial t} = \frac{\partial}{\partial x} \left(F(x)p(x) - D(x)\frac{\partial}{\partial x}p(x) \right) \tag{1.18}$$

Here the first term on the right hand side is the drift term and the second term is the diffusion term. The Fokker Planck equation is a local conservation equation, the term inside the bracket on the right hand side represents the probability current. There is an equivalent description of the system in terms of the dynamics of the random variable, x, instead of the probability, $p(x)$, given by the Langevin equation

$$\frac{dx}{dt} = F(x) + D(x)\eta(t) \tag{1.19}$$

In this formulation, $F(x)$ is a deterministic driving force and $\eta(t)$ is the white noise which which characterizes the fluctuations and $D(x)$ is the diffusion constant which decides the strength of fluctuations. Many of the thermodynamic results for macroscopic equilibrium systems can be extended to these non-equilibrium Markov systems by meaningfully assigning thermodynamic properties, like work, entropy etc., to the trajectories of these Markov systems (See [26] for a review).

Recent studies in non equilibrium Markov system have proven a series of bounds on fluctuations in these systems, called the thermodynamic uncertainty relations

(TUR) [27]. An example of TUR in Markov systems is the following bound:

$$\frac{\mathrm{Var}(J_\tau)}{\langle J_\tau \rangle} \geq \frac{2k_B}{\Sigma_\tau} \tag{1.20}$$

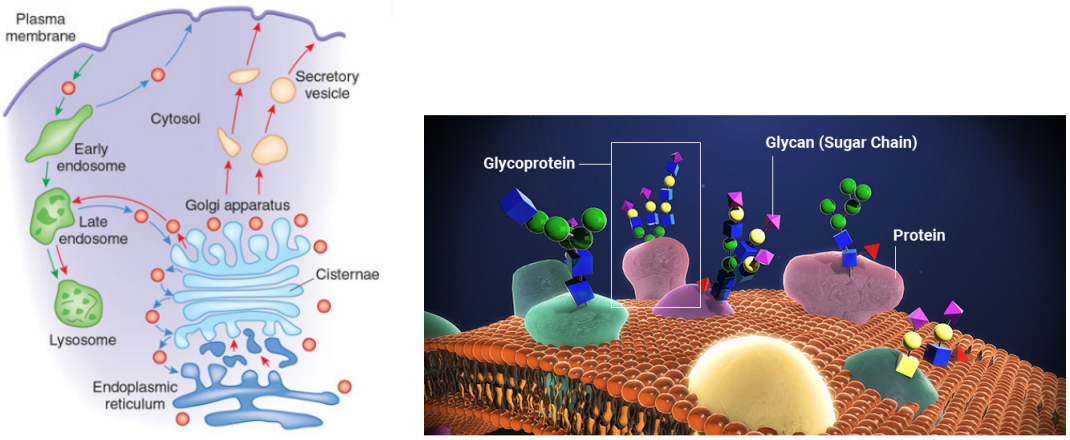
here J_τ and $\Sigma_\tau = \tau \sum_{i,j} J_{ij} \log \frac{W_{ij}}{W_{ji}}$ is the dissipation. Recently, [28] has given a biological relevant TUR, in the context of sensing of a ligand, bounding the fluctuations in the occupancy time of discrete Markov states by the entropy dissipation.

1.5 Gycans, Glycosylation and the Golgi Complex

Eukaryotic cells are internally divided into membrane enclosed functionally distinct compartments called organelles. Each of these organelles have a specialized set of enzymes and other molecules for optimal chemical processing required to perform a function. There is also a complex transport system, consisting of gated/channels in the membrane, protein translocation mechanisms and vesicular transport, which carries cargo between these organelle [1]. These organelle are functionally important for the cell because they increase the membrane area to host biochemical reactions and provide functionally specialized aqueous region optimized for specific biochemical reactions. Here we discuss one of these organelle, the Endoplasmic Reticulum(ER) - Golgi secretion system, which is the focus of this thesis.

1.5.1 The ER- Golgi Complex secretion system

ER is a large labyrinth of membraneous sac surrounding the cell nucleus which is held together by the cytoskeleton (See Figure 1.5). Part of the ER is dotted with ribosomes and is the site for protein and lipid synthesis. The post-translational modifications of the synthesized proteins and lipids also begins in the ER before they are transported to various target locations in the cell, mainly the Golgi complex. The Golgi complex is a collection of flattened membrane bound sac like structures called, cisternae, held together by microtubules (See Figure 1.5). Each cisternae has distinct chemical composition in terms of the pH, the enzymes present inside the cisternae and other molecules. The Golgi complex maintains a stable structure in the face of the constant influx of vesicles from the ER and outflux to plasma



(a) (b)

Figure 1.5. (a)The ER Golgi secretion system: Proteins and lipids from the ER are transported to the Golgi complex by vesicles where they are chemically modified before finally being transported to different parts of the cell. (b) Cartoon of glycans on the surface on the cell attached to protein embedded in the cell membrane.
Picture credit: (a) https://www.nature.com/scitable/topicpage/ how-do-proteins-move-through-the-golgi-14397318, (b) https://www.glytech-inc.com/glycan/glycans-and-cells/

membrane and other parts of the cell.

The ER-Golgi secretion systems (See Figure 1.5) deals with the production, modification and transport of proteins and lipids. Proteins are synthesized by the ribosomes on the surface of ER in the cytoplasm. They are then flipped into the lumen of ER where the post translational modification begin, e.g. addition of sugar tree like structures on the surface top of the protein core by a process called glycosylation [1]. The modified proteins are then transported to the cis-Golgi by vesicles, from where they are further transported to the trans-Golgi network, plasma membrane and other cellular destinations through vesicles. The transport of synthesized protein depends on the sorting signals that direct their delivery to locations outside the cytosol or to organelle surfaces.

1.5.2 Non equilibrium self assembly of the Golgi complex

As previously stated, the Golgi complex is a stable membranous structures, subject to and driven by a continuous flux of membrane-bound material from the ER to the plasma membrane (PM). The morphology and chemical identity of compartments over large spatio-temporal scales should therefore emerge as the self organized steady

states of a driven non-equilibrium system [29].

The flux of vesicles from ER transports proteins and lipids to the Golgi complex for further modification. A complete understanding of protein transport through the Golgi stack is lacking with two competing models representing the extreme cases - the vesicular transport model and the cisternal progression model. In the vesicular exchange model, cisternae are stable "static" structures and transit proteins move along the stack by anterograde vesicular transport. In the cisternal progression model, the entire cisternae progress through the stack and Golgi resident enzymes undergo retrograde (backward) transport to remain in a fixed position. A theoretical model of cargo transport within the Golgi cisternae should therefore accommodate both these extreme cases, one such model is given in [30]. The localization of resident proteins/enzymes in Golgi cisternae against a bulk flow of material suggests the presence of cargo specific retrograde transport to maintain chemical identity of the cisternae [31].

The size, shape, composition, and location are all important and regulated features of these organelles that ultimately contribute to the organelle's function. In this thesis we focus on how the number of Golgi cisternae affect the process of glycosylation and motivate the need for a multi-cisternal system. Implying that there should be a coupling between the enzyme kinetics, which carry out the function, and the dynamics of compartmentalization responsible for the non-equilibrium self assembly of the Golgi cisternae.

1.5.3 Glycosylation

Glycosylation of protein, arguably the most diverse post-translational modification, is a complex, multistep process that employs around many glycosyltransferase enzymes (200 in humans [8]) that determine which proteins are to become glycoproteins, the positions of glycans on those proteins and the glycan structures assembled [8]. There is a tremendous diversity in the kind of glycoconjugates that can be obtained, much more than polynucleotides and polypeptides. There are nine common monosaccharides found in vertebrate glycoconjugates which can be linked through glycosidic bonds, to make oligosaccharides or polysaccharides. The diversity arises not only from the choice of sugars but also from the way they are linked,

allowing not only more linear products, but also branched products. Branching is a prime characteristic of many glycans found on mammalian cell surfaces with glycans having two, three or four branches. The are two major types of eukaryotic protein glycans - N-glycans and O-glycans. An N-glycan makes a glycosidic bond with the side-chain nitrogen of an asparagine residue. An O-glycan makes a glycosidic bond with the terminal oxygen of a serine or threonine residue. Glycosylation also involves addition of other groups, like sulfation etc., to the glycan structure. One particularly interesting one, abundant in mammalian cells, is the process of Sialylation, a terminal modification, which prevents further addition of sugar monomers to the glycan chain [8].

All forms of glycosylation in the secretory pathway are highly ordered and sequential processes, typically involving glycosyltransferase reactions. The general glycosylation reaction, shown below, involves the catalysis of a group transfer reaction in which the monosaccharide moiety of a simple sugar donor substrate, e.g. UDP-Gal, is transferred to the acceptor substrate [8]

$$\begin{aligned} \text{Acceptor} + \text{glycosyl donor} + \text{Enzyme} \quad &\rightleftharpoons \quad [\text{Acceptor} \cdot \text{glycosyl donor} \cdot \text{Enzyme}] \\ &\rightarrow \quad \text{glycosylated acceptor} + \text{nucleotide} + \text{Enzyme} \end{aligned}$$

Most glycosylation reactions use activated forms of monosaccharides (nucleotide sugars) as donors for glycosyltransferases. In eukaryotes, most of these donors are actively transported across a membrane bilayer by specific multipass transporter proteins, becoming available for reactions within the lumen of the ER–Golgi pathway [32, 33].

The majority of Golgi glycosylation enzymes are membrane proteins placing their catalytic sequences in the Golgi lumen, where they participate in the synthesis of the glycan chains on proteins and lipids during their transit through the secretory pathway. These enzymes, their glycan substrates (attached to protein or lipid), and the appropriate nucleotide sugar donor must be located in the same compartment. Biochemical and ultrastructural studies indicate that glycosyltransferases segregate into distinct overlapping compartments within the secretory pathway. Generally speaking, enzymes acting early in the biosynthetic pathway localize to cis- and medial-Golgi compartments, whereas those acting later in the pathway tend to lo-

calize in the trans-Golgi cisternae and the TGN [8].

Glycosyltransferases constitute a very large family of enzymes, most show a high degree of specificity for both their donor and acceptor substrates but some are promiscuous. Generally speaking, the enzymes that elongate glycans act sequentially so that the product of one enzyme yields a preferred acceptor substrate for the subsequent action of another. The end result is a linear and/or branched structure composed of monosaccharides linked to one another. Acceptor recognition by these glycan-elongating glycosyltransferases does not typically care about the polypeptide or lipid moiety at the root of the acceptor substrate [8]. Apart from the the glycosyltransferases there are glycosidases that remove monosaccharides to form intermediates that are then acted on by glycosyltransferases also play a role in the biosynthesis of some glycan types. In addition, glycans can be modified by many other enzyme types, like sulfotransferases, which add groups other than monosaccharides .

Functions of glycans

Glycosylation greatly amplifies the proteome by producing diverse proteoforms with different properties, thereby instructing myriad functions. Glycans, being the most diverse and flexible molecules, are ideal to position at the interface between cells and the extracellular milieu (See Figure 1.5) possibly due to their relative hydrophilicity, flexibility, and mobility in aqueous environments and their extreme diversity, allowing short-term and long-term adaptations to changing environments and pathogen regimes [9].

Glycans are involved in structural functions, e.g., extracellular scaffolds: cell walls and extracellular matrices, in energy metabolism and as carriers of molecular information. In this thesis we focus on the role of glycans as information carriers. Glycan binding proteins (GBPs) recognise the molecular patterns of glycans and selectively bind to the glycans. The role of glycans as information carriers is particularly prominent in the assembly of complex multicellular organs and organisms, which requires interactions between cells and the surrounding matrix. Being on the outer surface of cellular and secreted macromolecules, glycans are in a position to modulate or mediate a variety of events in cell–cell, cell–matrix, and cell–molecule interactions

critical to the development and function of a complex multicellular organism [8]. They can also mediate interactions between organisms, e.g., between host and a parasite, pathogen, or a symbiont. This internal and external glycan recognition in a multicellular organism can also act as opposing selective forces, simultaneously constraining and driving evolutionary change respectively, likely accounting for the enormous diversity of glycans in nature. Diversity is further enhanced by microbial pathogens engaged in "molecular mimicry", evading immune reactions by decorating themselves with glycans typical of their hosts [8].

Studies of deficiencies in glycosylation enzymes in animal models and human diseases have advanced understanding of biological functions of protein glycosylation and demonstrated that most glycosyltransferases serve essential roles in mammalian physiology [8]. The glycome is produced and regulated by the glycosylation machinery in a single cell, yet analysis of glycans at the single-cell level is not possible with current glycomics methods, which are limited to probing with glycan-specific antibodies and glycan-binding proteins (such as lectins). It is therefore often perceived as a daunting task to uncover and dissect specific biological functions of glycans and the underlying molecular mechanisms. Advances in next-generation sequencing and proteomics are beginning to provide single-cell transcriptomes and proteomes, which has opened the way for global analysis of the network of enzymes that orchestrate protein glycosylation and the assessment of the glycosylation capacities of any given cell [34]. Accompanying these are the nuclease-based gene editing technologies that — through precise manipulation of glycosylation enzymes — provide virtually unlimited opportunities for engineering, exploration and custom design of cellular glycosylation capacities. We can now probe glycosylation systematically through a genetic entry point, and with additional efforts we will be able to connect information on cellular glycosylation capacities with the actual outcome of the glycome and roles of glycosylation in cells. The biological consequences of experimentally altering glycosylation in various systems seem to be highly variable and unpredictable. Also, a given glycan can have different roles in different tissues, at different times in development (organism-intrinsic functions) or in different environmental contexts (organism-extrinsic functions) [8].

1.6 Scope of the thesis

In this thesis, we look at glycosylation from the lens of information theory in an effort to give meaning to what is colloquially called the “glycan code" [35]. We then show that how this “glycan code” puts constraints on the synthesis machinery made up of the Golgi cisternae and the glycosylation enzymes.

- In the second chapter, we characterize the complexity of the glyan code and show that complex organisms indeed have complex glycan code.
- In the third chapter, we provide a simple model of glycosylation as a sequential chemical modification by enzymes that can catalyze more than one substrate in chemically distinct compartments. We then discuss what kind of concentration profiles are obtained by this synthesis model and how are they controlled by the parameters of the model.
- In the fourth chapter, we study the constraints put on the glycosylation synthesis machinery by the requirement of having a complex glycan code. We bring out the various trade-offs in terms of number of compartments, number and substrate specificity of enzymes and the complexity of the observed glycan distributions.
- In the fifth chapter, we study the modulation of enzyme specificity by non-equilibrium driving and its implication to the fidelity of synthesis of a complex glycan profile.

We conclude in the sixth chapter with a short discussion and outline the directions of future work.

Bibliography

[1] B Alberts et al. *Molecular Biology of the Cell*. Garland Science, 2002.

[2] Uri Alon, Michael G Surette, Naama Barkai, and Stanislas Leibler. Robustness in bacterial chemotaxis. *Nature*, 397(6715):168–171, 1999.

[3] Judith A Owen, Jenni Punt, Sharon A Stranford, et al. *Kuby immunology*. WH Freeman New York, 2013.

[4] William Bialek. *Biophysics: searching for principles*. Princeton University Press, 2012.

[5] Wallace F Marshall. Pattern formation and complexity in single cells. *Current Biology*, 30(10):R544–R552, 2020.

[6] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[7] James M Carothers, Stephanie C Oestreich, Jonathan H Davis, and Jack W Szostak. Informational complexity and functional activity of rna structures. *Journal of the American Chemical Society*, 126(16):5130–5137, 2004.

[8] A Varki et al. *Essentials of Glycobiology*. Cold Spring Harbor Laboratory Press, 2009.

[9] Ajit Varki. Biological roles of glycans. *Glycobiology*, 27(1):3–49, 2017.

[10] David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[11] Imre Csiszár and János Körner. *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press, 2011.

[12] Justin B Kinney and Gurinder S Atwal. Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, 111(9):3354–3359, 2014.

[13] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[14] John J Hopfield. Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity. *Proceedings of the National Academy of Sciences*, 71(10):4135–4139, 1974.

[15] Arvind Murugan, David A Huse, and Stanislas Leibler. Speed, dissipation, and error in kinetic proofreading. *Proceedings of the National Academy of Sciences*, 109(30):12034–12039, 2012.

[16] Naama Barkai and Stan Leibler. Robustness in simple biochemical networks. *Nature*, 387(6636):913–917, 1997.

[17] Marc W Kirschner and John C Gerhart. *The plausibility of life*. Yale University Press, 2008.

[18] Aleksandra M Walczak, Gašper Tkačik, and William Bialek. Optimizing information flow in small genetic networks. ii. feed-forward interactions. *Physical Review E*, 81(4):041905, 2010.

[19] Garud Iyengar and Madan Rao. A cellular solution to an information-processing problem. *Proceedings of the National Academy of Sciences*, 111(34):12402–12407, 2014.

[20] Krishnan Swaminathan Iyer, Chaitra Prabhakara, Satyajit Mayor, and Madan Rao. Cellular compartmentalisation and receptor promiscuity as a strategy for accurate and robust inference of position during morphogenesis. *bioRxiv*, 2022.

[21] Simona Cocco, Christoph Feinauer, Matteo Figliuzzi, Rémi Monasson, and Martin Weigt. Inverse statistical physics of protein sequences: a key issues review. *Reports on Progress in Physics*, 81(3):032601, 2018.

[22] Juan MR Parrondo, Jordan M Horowitz, and Takahiro Sagawa. Thermodynamics of information. *Nature physics*, 11(2):131–139, 2015.

[23] Dilip Kondepudi and Ilya Prigogine. *Modern thermodynamics: from heat engines to dissipative structures*. John Wiley & Sons, 2014.

[24] Crispin Gardiner. *Stochastic methods*, volume 4. Springer Berlin, 2009.

[25] Jürgen Schnakenberg. Network theory of microscopic and macroscopic behavior of master equation systems. *Reviews of Modern physics*, 48(4):571, 1976.

[26] Udo Seifert. Stochastic thermodynamics, fluctuation theorems and molecular machines. *Reports on progress in physics*, 75(12):126001, 2012.

[27] Jordan M Horowitz and Todd R Gingrich. Thermodynamic uncertainty relations constrain non-equilibrium fluctuations. *Nature Physics*, 16(1):15–20, 2020.

[28] Sarah E Harvey, Subhaneil Lahiri, and Surya Ganguli. Universal energy accuracy tradeoffs in nonequilibrium cellular sensing. *arXiv preprint arXiv:2002.10567*, 2020.

[29] Pierre Sens and Madan Rao. (re) modeling the golgi. In *Methods in Cell Biology*, volume 118, pages 299–310. Elsevier, 2013.

[30] Serge Dmitrieff, Madan Rao, and Pierre Sens. Quantitative analysis of intra-golgi transport shows intercisternal exchange for all cargo. *Proceedings of the National Academy of Sciences*, 110(39):15692–15697, 2013.

[31] Benjamin S Glick, Timothy Elston, and George Oster. A cisternal maturation mechanism can explain the asymmetry of the golgi stack. *FEBS letters*, 414(2):177–181, 1997.

[32] Carlos B Hirschberg, Phillips W Robbins, and Claudia Abeijon. Transporters of nucleotide sugars, atp, and nucleotide sulfate in the endoplasmic reticulum and golgi apparatus, 1998.

[33] Carolina E Caffaro and Carlos B Hirschberg. Nucleotide sugar transporters of the golgi apparatus: from basic science to diseases. *Accounts of chemical research*, 39(11):805–812, 2006.

[34] Katrine T Schjoldager, Yoshiki Narimatsu, Hiren J Joshi, and Henrik Clausen. Global view of human protein glycosylation pathways and functions. *Nature reviews Molecular cell biology*, 21(12):729–749, 2020.

[35] Hans-Joachim Gabius. The sugar code: why glycans are so important. *BioSystems*, 164:102–111, 2018.

Chapter 2

Complexity of the glycan code

In the introduction we described the role of glycans as information carriers. Specifically, glycans on the cell surface are recognized by the Glycan Binding Proteins (GBP) of host cells to identify particular cell types and niches within a multi-cellular organism [11]. Simultaneously these glycans are also recognized by pathogens for identification of cells to infect [11]. Due to the role of glycans as the markers of cell type and niche identity, the distribution of glycans on the surface of the cell can be thought of as a code [1] which is shaped by evolution to have certain properties. The competition between the ability to reliably and precisely identify different cell types and niche within a multi-cellular organism on one hand while simultaneously evading recognition from pathogens in a changing environment on the other requires a “complex" and “diverse" distribution, and hence code, of glycans on the cell surface.

In this chapter we develop the notion of complexity of the glycan code. We start with a general notion of complexity explored in various contexts and the properties shared by a complex system. We will then explore the notion of complexity for glycan codes before finally coming to quantifying the complexity of real mass spectrometry (MS) glycan profiles. We will find that, indeed, the glycan distribution in complex multi-cellular organisms is more “complex" than simpler organisms.

2.1 Complexity depends on the context

The idea of complexity of a system has been explored in various fields like computer science. We start with a few examples and look the the general characteristics shared by systems which we intuitively understand to be complex.

In computer science, problems are classified in various complexity classes - P, NP and NP complete depending upon how the computational time or memory required by an algorithm to solve the problem of scales with the size of the problem [2]. This scaling is a fundamental property of the problem and is independent of the particulars of how the computations are performed. Problems within a complexity class can be mapped to one-another and therefore the complexity classification has both conceptual and practical importance. Here the most complex problems have exponential scaling of required computation time or memory with problem size.

In dynamical systems, complexity is associated with systems having the large number of interacting parts generating complex behavior. The following general properties are shared by all dynamical systems which we intuitively call complex [3] - (a) They are between total order and disorder, e.g. liquids are much more complex than solids and gases (b) They have hierarchies of timescales and length-scales, and interaction between these hierarchies (c) They have many interacting parts and, strong non-trivial correlation between these parts (d) There are correlations between the system and the environment. One way to quantify complexity in these dynamical systems is by using the Kolmogrov-Sinai complexity [3] defined on the trajectories of this system. It is defined by partitioning the trajectories into discrete regions and generating a sequence of symbols from the trajectories and defining a Shannon entropy on these sequences. This is related to difficulty of prediction the future trajectory given the past.

In communication, complexity is defined in terms of memory required to store and retrieve the shortest code that can generate a given sequence of symbols. This is called the algorithmic or Kolmogrov complexity [4, 5]. A sequence of symbols that is generated by a longer code is more complex than the one generated by a shorter code. The Kolmogrov complexity of a sequence of symbols is related to the Shannon entropy of the sequence.

In all these cases the complexity is related to the difficulty of performing a task, e.g. in computer science it is related to the difficulty of solving a problem, in dynamical systems it is related to the difficulty of predicting the future trajectories, in communication it is the difficulty to generate a sequence of symbols and in statistical modeling it is the difficulty to model the probability distribution of a data. Grassberger fittingly defines complexity as the “difficulty of meaningful task" [3] .

The difficulty in most cases can be quantified by the Shannon information but the "meaningful" part requires a definition of complexity which depends on the context.

In the following section we describe how Shannon information is inadequate in characterizing complexity of a system and how can the notion of "meaning" be added to information theory resulting in a more reasonable information theoretic description of complexity.

2.1.1 Information and complexity

A key obstacle in quantifying complexity using information theoretic approach is to differentiate between a "complex" system and a totally disordered system. The Shannon entropy and other similar quantifications of information are good at quantifying the uncertainty in predicting the next symbol in a sequence of symbols but are maximized by a totally random distribution. We give a brief summary of a couple of approaches that prevent totally random distributions from becoming the most complex ones by associating "meaning" or functional relevance with appropriate information theoretic quantities.

In **??**, the authors quantify complexity of a time series by relating it to the predictability, defined as "predictive information", $I_p(T)$ which is the mutual information between the past and the future of the time series. They argue that this is the part of information that has functional consequences or "meaning" for a living organism. Let $x(t)$ be a stream of data, $x_{\text{past}} = x(t)$ for $-T < t < 0$ and $x_{\text{future}} = x(t)$ for $0 < t < T'$ represent the past and future data respectively. Then the predictive information is defined by

$$\begin{aligned} I_p(T,T') &= \left\langle \log \frac{P(x_{\text{future}}|x_{\text{past}})}{P(x_{\text{future}})} \right\rangle \\ &= -\langle \log P(x_{\text{future}}) \rangle - \langle \log P(x_{\text{past}}) \rangle - (-\langle \log P(x_{\text{future}}, x_{\text{past}}) \rangle \\ &= H(T) + H(T') - H(T+T') \end{aligned} \tag{2.1}$$

The last line is obtained by assuming that the data is coming from stationary, i.e. time translation invariant source. $H(T)$ is extensive in T, i.e $\lim_{T\to\infty} \frac{H(T)}{T} = H_0$ and

$H(T) = H_0 T + H_1(T)$. In the limit of future going to infinity we can write

$$I_p(T) = \lim_{T' \to \infty} I_p(T, T') = H_1(T) \tag{2.2}$$

All the useful information of the past is in the sub-extensive part of the entropy of the time series. Scaling of the predictive information or the sub-extensive entropy ($H_1(T)$) with system size reveals complexity of the system. As an illustration, the authors take an Ising system and study the predictability of a sequence of words of a fixed length formed by combining a fixed number of consecutive Ising spins. They show that the predictive information is constant with respect to the word length for a simple Ising system with a constant nearest neighbor interaction but for a complex Ising system with random long range decaying interactions the predictive information scales logarithmically with the word length.

Another way of incorporating the functionality into information theory framework, called the Information Bottleneck, was developed by Tishby, Pereira and Bialek in [6]. Here instead of complexity of a time series data we can define the complexity of any arbitrary sequence of random variables as long as we have a notion of a target/relevant set of random variables, information about which should be preserved while compressing the original sequence of random variables.

The information between two random variables X and Y is squeezed through a bottleneck representation, T (See Figure 2.1). Here the random variable Y represents some relevant variable, e.g. in physics and biology many times the microscopic details of a system are not relevant to a description in terms of meso or macro variables at a longer length or timescale. The random variables X and Y represent the microscopic and relevant variables(respectively) in the description. The microscopic description of the system might be too elaborate ($|X|$ too large) for any practical computation and thus requires compression while still preserving enough information about Y to be functional. These two opposing forces gives rise to what is called the relevance-compression trade-off.

This is like the rate distortion principle discussed in the previous chapter where the distortion function is given by the mutual information between the target(Y) and compressed(T) set. The information bottleneck principle [6] for data X, and a target or relevant random variable set Y coming from a joint probability $P(X, Y)$,

and the compressed random variable set T, is described by the the optimization of the Information Bottleneck(IB) functional, $I(X,T) - \beta I(Y,T)$, over the transition probabilities, $p(t|x)$).

$$R(\beta) = \min_{p(t|x)} I(X,T) - \beta I(Y,T) \tag{2.3}$$

Here the mutual informations, $I(X,T)$ and $I(T,Y)$ represent the compression and the relevance respectively and β is the trade-off parameter. Notice that the random variables X, Y, T form a Markov chain represented by $T \leftrightarrow X \leftrightarrow Y$, the Data Processing Inequality(DPI) on this Markov process implies $I(T,Y) \leq I(X,Y)$. We obtain the IB function, like the rate distortion function, by doing the optimization in (2.3) for various values of β. This gives a curve in the space of $I(T,X)$ and $I(T,Y)$ which characterizes the compression-relevance of the joint probability $P(X,Y)$. The area above the curve marks the unfeasible region in this space. The complexity of the system can then be defined on the basis of where a particular system is placed in the relevance-compression curve.

In the following section we apply the Information bottleneck framework to get the complexity of the glycan distribution keeping in mind their role in cell type and niche identification.

2.2 What drives the glycan complexity?

One of the important function of glycans is cell-type and niche differentiation in a multicellular organisms; they are markers of cell type identity and niche which are seen by other cells of the organism like the immune cells. Each cell type (in a niche) is identified with a distinct glycan profile [1, 7, 8], and this glycan profile is noisy because of both the cell to cell variations in the synthesis and transport machinery, and the stochastic noise associated with the synthesis and transport [8, 9, 10]. Moreover, the cell type and niche differentiation has to be performed while evading pathogen in a changing environment. This is further complicated by the fact that pathogens themselves can engage in mimicry of glycans of the host cell to evade the immune response [11]. We are interested in defining the complexity of glycan distribution in the context of reliable cell type differentiation in presence of cellular variations and pathogens. Initially, we focus on cell type differentiation in

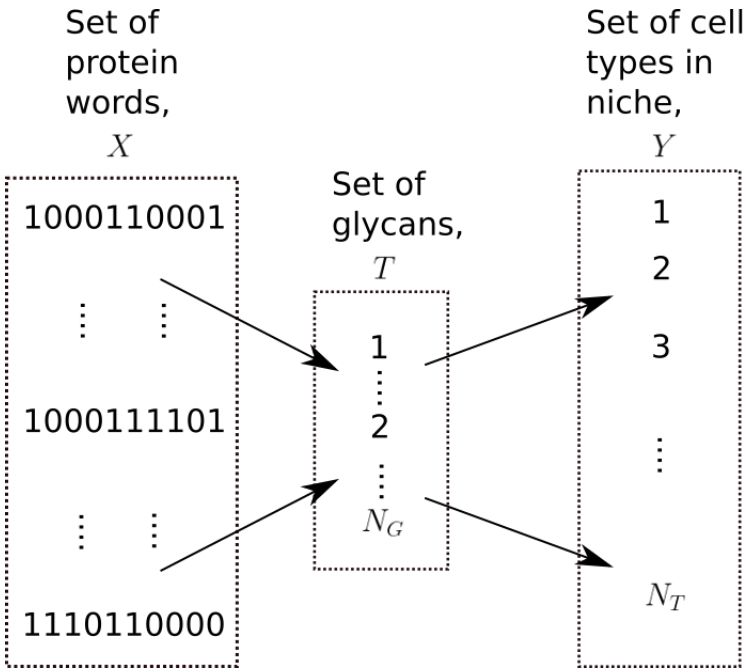


Figure 2.1. Glycan information bottleneck schematic. Each element of the set X is called a protein word and it represents the protein expression of a cell, each element of the set Y is the a cell type and niche, and each element of the set T is a glycan. The information of cell type and niche is characterized by the protein words and contained in the joint $P(X, Y)$. This information is squeezed through the smaller ($|X| \gg |T|$), bottleneck set of glycans (T). Information bottleneck principle gives the the optimal encoding, transition probabilities $p(t|x)$, that preserves information about the set Y.

presence of cellular variations, and identify the characterstics of a more complex code vs a less complex code in this context. We subsequently use these characteristics to characterize the complexity of the mass spectrometry glycan profile of the real cell types.

Cells of different cell types and niches in an organism can be characterized on the basis of proteins expressed in the cell. The signature of cell type and niche is in a subset of all the proteins expressed in the cell as evident by the lower dimensional representations (like t-SNE, PCA, etc.) of single cell mRNA analysis [15, 16]. There are house-keeping proteins which are shared by most of the cells and are not important for the purposes of cell type identification. We look at how this information of cell type and niche is encoded into the glycans from the proteins, and what are the theoretical properties of such an encoding. The advantage of this encoding for the organism might be a compressed representation of its identity that can be modulated on a much faster timescale than proteins, endowing the cell with the ability to adapt to fast changing pathogen properties.

Figure 2.1 shows the schematic of the calculation. The set X represents the set of expression levels for all proteins, each protein expression *word* is a string of bits, which are assumed to be to be binary, representing whether the protein is expressed in the cell-type or not. There is a joint probability associated with the elements of set X and elements of the set of cell-types, Y. Single cell mRNA sequencing data can be used to estimate the joint probability distribution of the cell type and protein expression. The set of glycans T is the bottleneck set of much smaller size than the protein set. The information of the protein word is to be translated to a smaller glycan set while preserving the cell type identification functionality.

In the following section we provide a toy model to generate the joint probability of a protein word and cell type and niches, and study the information bottleneck encoding. We will subsequently use the single cell mRNA sequencing data for estimating the joint probability between the protein expression and cell type.

2.3 A toy model for the joint probability of protein words, and cell types and niches

Denoting the total number of proteins by N_p, a cell-type α can be characterized by the binary string - $x^\alpha = \{s_1^\alpha, s_2^\alpha, \ldots s_{N_p}^\alpha\}$ where $s_j^\alpha \in \{0,1\}, \forall j, \alpha$, represents whether the protein s_j is expressed in cell-type α or not.

The protein expression across the cells of the same cell type can have some variation due to cell to cell variations in the synthesis and transport machinery and other sources of noise. While this distribution can be taken from the m-RNA data across cells of a given cell type, we define a point probability, q, which is the probability of a single bit flipping in the binary string (x^α) associated with a cell type. Given this rule we can associate a conditional probability between a cell type and a binary string of length N_p.

$$\begin{aligned} \text{Prob}(x^\alpha|\alpha) &= 1/Z \\ \text{Prob}(x \in D_1^\alpha|\alpha) &= q/Z \quad D_1^\alpha = \{x : |x - x^\alpha| = 1\} \\ \text{Prob}(x \in D_n^\alpha|\alpha) &= q^n/Z \quad D_n^\alpha = \{x : |x - x^\alpha| = n\} \end{aligned} \tag{2.4}$$

Here $|x - x^\alpha|$ represents the Hamming distance between x and x^α, D_n is the set of all x that are n bit flips away from x^α and Z is a normalization constant. Figure 2.2 show a typical realization of the conditional probability of protein word x given the cell type y, $p(x|y)$, generated by the model. The joint probability, $p(x, y)$, was obtained by assuming uniform distribution over the cell types, $p(y) = 1/N_C$.

We do the optimization of the IB functional (2.3) for the above joint $p(x, y)$ probability distribution and characterize the relevance - compression trade-offs for this system. The goal is to quantitatively capture the notion of complexity for such a system and show how this complexity is affected by various system parameters i.e the variations in the set X, number of cell types N_C etc.

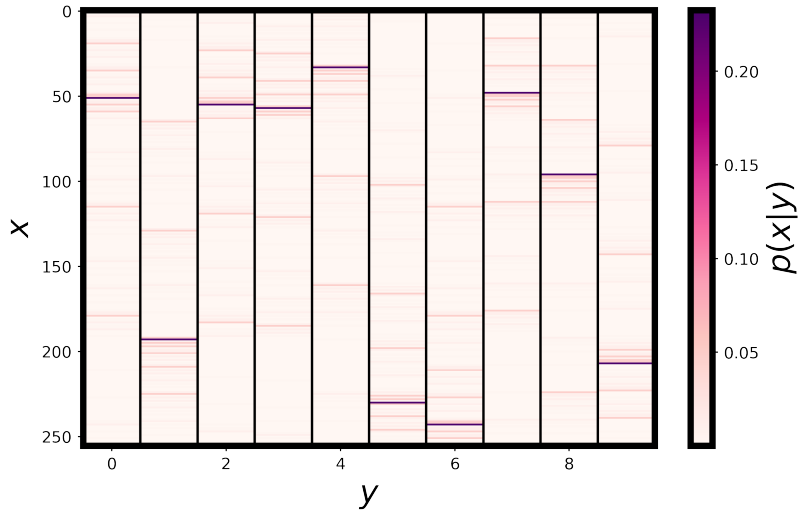


Figure 2.2. A typical realization of the conditional probability of protein word x given the cell type y, $p(x|y)$, generated by the model.

2.3.1 Optimizing the IB functional

In [6], the authors prove that the conditional $p(t|x)$ is a stationary point of the IB functional $\mathcal{L} = I(T,X) - \beta I(T,Y)$ if and only if

$$p(t|x) = \frac{p(t)}{Z(x,\beta)} \exp(-\beta D_{KL}[p(y|x)||p(y,t)]), \quad \forall t \in \mathcal{T}, x \in \mathcal{X} \tag{2.5}$$

where D_{KL} is the KL Divergence (See Chapter 1 (1.8)) between two probability distributions. The above equation along with the fact that the IB variables X, Y, T form a Markov chain represented by $T \leftrightarrow X \leftrightarrow Y$, give rise to an iterative scheme which converges to a stationary point of the IB function with respect to the transitional probabilities $p(t|x)$. This algorithm is sensitive to the initial conditions and only finds the local minimum, and therefore requires many initializations. We combine the iIB algorithm with a deterministic algorithm called aIB [12] into an annealing like procedure. We start with performing the aIB procedure for a very high β. In the limit of high β the transition probabilities become a delta function and the aIB provides the optimal solution. We then decrease β slightly and use the previous optimal $p(t|x)$ as the new initial condition for iIB and repeat this process annealing

β close to zero.

2.3.2 Results

We summarize the results of the calculation below:

- Figure 2.3 shows the relevance compression trade-off for different values of the cardinality of the bottleneck (glycan) set, N_G. Note that the compression decreases as we go from zero to one on the x-axis and the mutual informations $(I(T;X), I(T;Y))$ are normalized such that the range is between zero and one. The area above the curve is the unfeasible region for the system, meaning that above the curve relevance can not be achieved by the system for that value of compression. The points on the curves represent a particular value of β, the β values increase as we go from left to right on the curve, $\beta = 0$ corresponds to the most compressed point on the curve (origin) and $\beta \to \infty$ represents the most relevant point on the curve. As we increase the cardinality of the bottleneck (glycan) set we can achieve more relevance at the cost of compression. The lower cardinalities saturate earlier because of the compression caused by not having enough elements to represent (encode) the relevant information in the set X(proteins).

- Figure 2.4 shows the IB characteristics as a function of the bit flipping probability, q : (a) The IB curve for lower values of q saturate to a lower relevance and the saturation starts at a higher value of compression. Saturating relevance is achieved at a higher compression for low q systems. (b) shows the saturating (maximum) relevance achieved as a function of the cardinality of the bottleneck (glycan) set for various values of q. The shaded region represents ensembles of the same q. Higher q saturates at a larger N_G and at a lower maximum relevance.

- Figure 2.5 shows the IB characteristics as a function of the cardinality of the target set (the number of cell types), N_C (a) For low N_C, higher relevance can be achieved at higher compression, the curves saturates to the best relevance faster and the unfeasible region is smaller. (b) The saturating relevance can be

achieved by a smaller bottleneck (glycan) set for a system with lower number of cell types (N_C).

- Figure 2.6 shows complexity of encoding the protein set to the glycan set: The distribution over the glycan set, $p(t)$ becomes more detailed as we move to more higher relevance on the IB curve.

The results of Figure 2.6 show that the detail in the probability distribution over the glycan set is an indicator of the effectiveness of the glycans as a carrier of cell type information. We use this fact to define glycan complexity of real mass spectrometry glycan profiles. We can fit the glycan probability to a Gaussian Mixture model(GMM) and the number of GMM components required will correspond to the detail and hence the complexity of a glycan distribution. We detail this procedure for real Mass spectrometry glycan data in the following section.

2.4 Estimating complexity of the glycan molecular code

As described in the previous section, a large number of different cell types can be differentiated only if the cells are able to produce a large set of complex glycan profiles. We identified the complexity of a glycan profile with the amount of detail in the profile, which can be quantitatively measured by number of Gaussians needed to fit the profile well. A set of more complex glycan profiles is able to support differentiation of a larger number cell types, or equivalently, a more *complex* organism. In this section we measure the complexity of the mass spectrometry(MS) glycan profile of several cell types.

Before quantifying the complexity of the MS glycan profiles, we first need a consistent way of smoothening or coarse-graining the the raw glycan profiles obtained from MSMS measurements to remove measurement and synthesis noise. Here, we denoise the glycan profile by approximating it by a Gaussian mixture model (GMM) with specified number of components that are supported on a finite set of indices [13]. Consistent with the measure of complexity described in the previous section, we define the complexity of a mixture of Gaussians as the number of components m.

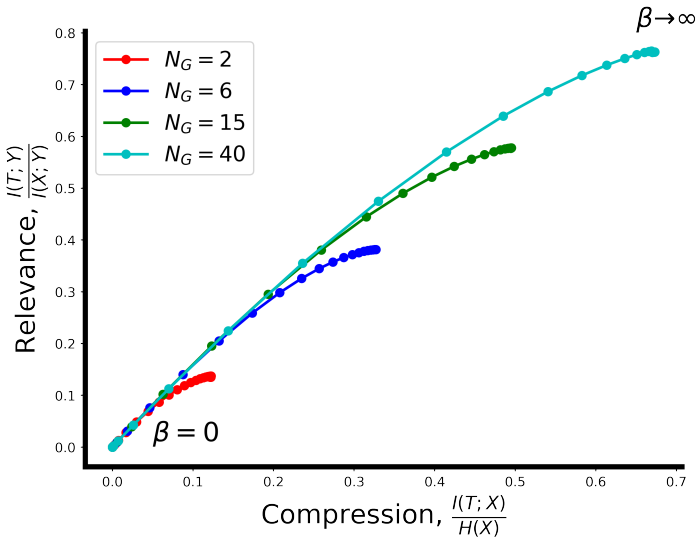


Figure 2.3. Curve showing relevance-compression trade-off for different values of the cardinality of the bottleneck (glycan) set, N_G. Note that the compression decreases as we go from zero to one on the x-axis. The area above the curve is the unfeasible region for the system meaning that relevance above the curve can not be achieved by the system for that value of compression. The points on the curves represent a particular value of β, the β values increase as we go from left to right on the curve, $\beta = 0$ corresponds to the most compressed point on the curve (origin) and $\beta \to \infty$ represents the most relevant point on the curve. As we increase the cardinality of the middle/bottleneck (glycan) set we can achieve more relevance at the cost of compression. The lower cardinalities saturate earlier because of the compression caused by not having enough elements to represent (encode) the relevant information in the set X(proteins).

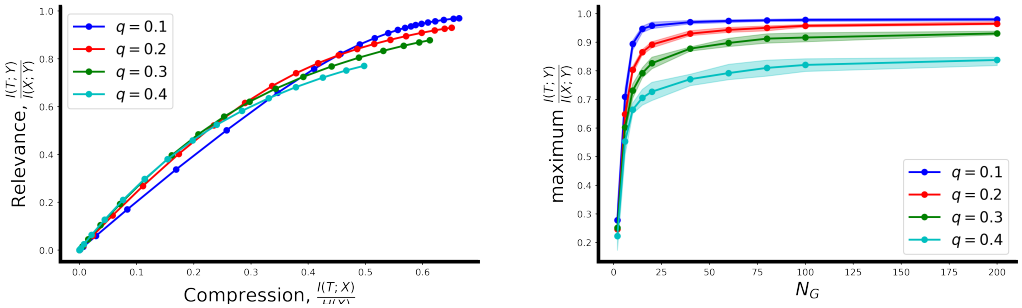


Figure 2.4. IB characteristics as a function of the bit flipping probability, q : (a) The IB curve for lower values of q saturate to a lower relevance and the saturation starts at a higher value of compression. Saturating relevance is achieved at a higher compression for low q systems. (b) shows the saturating (maximum) relevance achieved as a function of the cardinality of the bottleneck (glycan) set for various values of q. The shaded region represents ensembles of the same q. Higher q saturates at a larger N_G and at a lower maximum relevance.

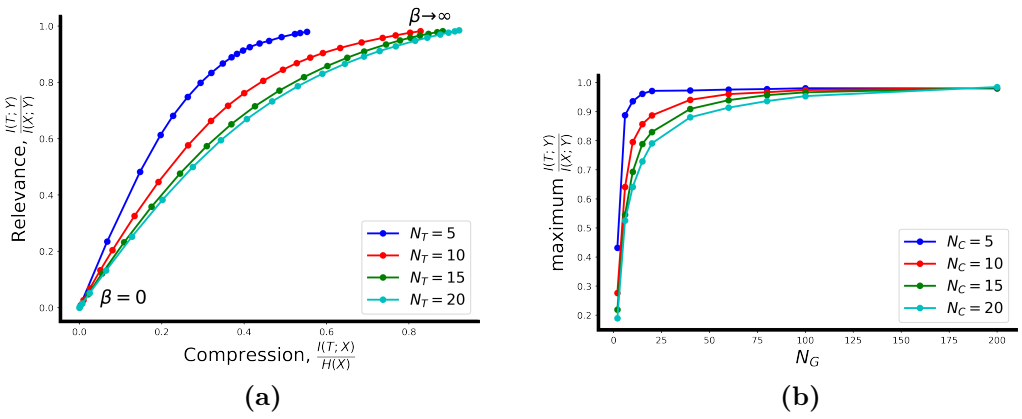


Figure 2.5. IB characteristics as a function of the cardinality of the target set (the number of cell types), N_C (a) For low N_C, higher relevance can be achieved at higher compression, the curves saturates to the best relevance faster and the unfeasible region is smaller. (b) The saturating relevance can be achieved by a smaller bottleneck (glycan) set for a system with lower number of cell types (N_C).

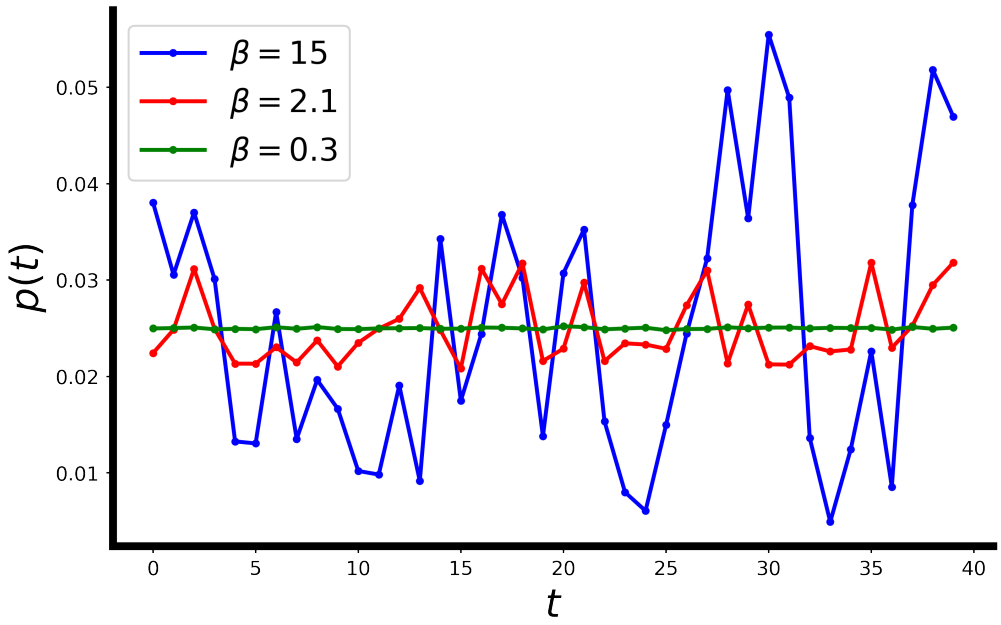


Figure 2.6. Complexity of encoding the protein set to the glycan set: The distribution over the glycan set, $p(t)$ becomes more detailed as we move to higher relevance on the IB curve.

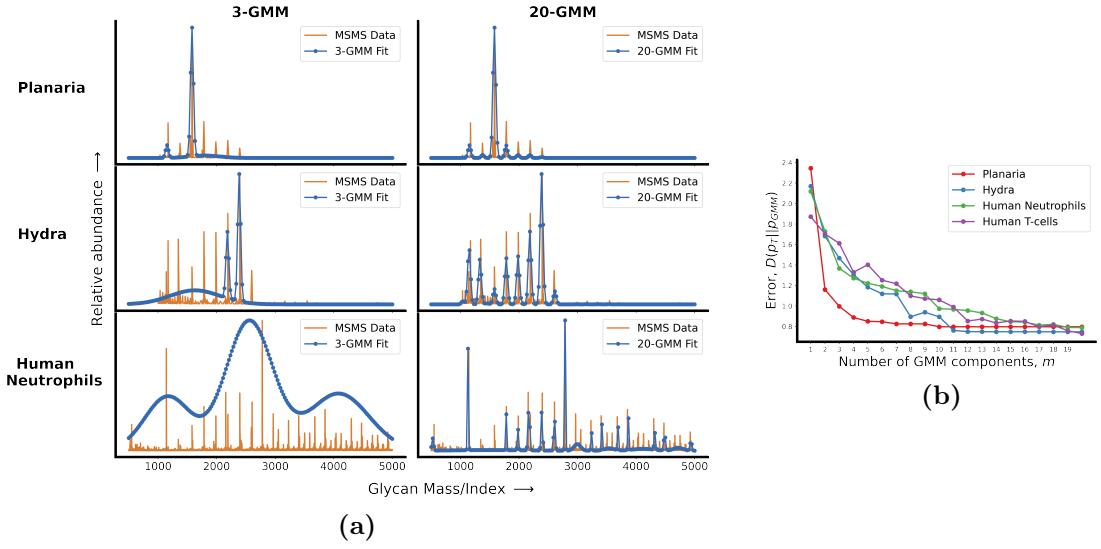


Figure 2.7. Living cells display a complex glycan distribution. (a) 3-GMM and 20-GMM approximation for the relative abundance of glycans taken from MSMS data of planaria *S.mediterranea*, *hydra magnipapillata* and *human* Neutrophils. (b) The change in the KL-divergence $D(p_T \| p^{(m)}_{GMM})$ as a function of the number of GMM components m. The KL-divergence for planaria saturates at $m = 5$, for hydra at $m = 11$, and for *human* cells at $m = 20$. Thus, the number of components required to approximate the glycan profile correlates well with the complexity of the organism.

Figure 1 demonstrates that the value of m at which the m-component GMM approximation of the target profile saturates, is a good measure of complexity. Using this we see that the complexity of the glycan profiles of various organisms correlates well with the number of cell types in an organism

We compare the complexity of glycan profiles of Hydra, Planaria and Humans. The number of cell types in Hydra, Planaria and Humans are around 41 [14], 44 [15] and 103 [16] respectively, based on transcriptome analysis (these are lower bounds based on the main cell types, and especially for Planaria and Hydra, are subject to constant revision). Our analysis of the MSMS data of these organisms suggest that organism with fewer cell types have less complex glycan distribution.

We give the detailed procedure of obtaining the GMM fit from the mass spectrometry data in the following section.

2.4.1 Statistical model for Glycan MS data: GMM

The distribution of the glycans on the cell surface is obtained via mass spectrometry. The x-axis of mass spectroscopy (MS) graphs is mass/charge of the ionized sample molecules and the y-axis is relative intensity corresponding to each mass/charge value, taking the highest intensity as 100%. This relative intensity roughly correlates with the relative abundances of the molecules in the sample.

The raw MS data is noisy and cannot be directly used for further calculations. There are three major sources of noise in the MS data [17]: chemical noise in the sample, the Poisson noise associated with detecting discrete events, and the Nyquist-Johnson noise associated with any charge system. We propose a simple model that accounts for the chemical noise and the Poisson sampling noise. Using this noise model and the available MS data, we generate parametric bootstrap samples of glycan measurements, and fit a Gaussian Mixture Model (GMM) on this sample to approximate the glycan distribution. .

The MS data obtained from [18, 19, 20] had mass ranging between 500 to 5000 Daltons with intensity reported at every 0.0153 Daltons. We first bin this MS data into 180 bins and take the maximum value within each bin as the value of intensity for that bin. Fig. 1 plots the raw MS data and the binned distribution. Next, we describe the parametric bootstrap model that we used to generate the glycan data.

2. Complexity of the glycan code

Let $\bar{I}_k$ represents the relative intensity of the k-th bin in the binned MS graph. We generate a sample population of glycans using the MS data in the following way:

1. Poisson sampling noise: The MS data does not have absolute count information. We assume an arbitrary maximum count $I_{\max}$, and define the intensity $I_k = I_{\max}\bar{I}_k$. The plots in Appendix 2.4 Fig. 2 (a) show that the results are not sensitive to the specific value of $I_{\max}$.

2. Chemical noise: The sample used for MS analysis also contains small amounts of molecules that are not glycans. These appear as the very small peaks in the MS data. We assume that the probability p_k that the peak at index k corresponds to a glycan is given by

$$p_k = 1 - e^{-\frac{I_k}{I_{max}}} = 1 - e^{-\bar{I}_k}$$

   which adequately suppresses this chemical noise.

3. Bootstrapped glycan data: The count n_k at the glycan index k is distributed according to the following distribution:

$$n_k = \begin{cases} 0 & (1 - p_k) \quad n = 0 \\ n & p_k e^{-I_k} \frac{(I_k)^n}{n!} \quad n \geq 1. \end{cases}$$

   We assume that the MS data was generated from N different cells. Thus, the total count at glycan index k is given by the sum of N i.i.d. samples distributed according to the distribution above. We in Appendix 2.4 Fig. 2 (b) show that results are insensitive to N. We normalize the count distribution by the total number of counts across all the bins to obtain the bootstrapped probability mass function p_T.

The bootstrapped distribution p_T is is noisy, and hence can not be used directly as the target distribution. We use a Gaussian Mixture Model (GMM) based approach to de-noise the raw data. The advantage of using a GMM based approach is that it creates an easily interpretable hierarchy of increasingly more detailed distributions to approximate the mass spectrometry profile. We define the *complexity* of a mass spectrometry profile as the minimum number of components (individual Gaussians)

in the GMM model required to approximate it. The details of the GMM calculations are as follows. We fix the number of components m. We want to approximate the bootstrapped probability p_T by the m-component mixture of Gaussian distributions $p_{GMM}(\theta) = \sum_{i=1}^{m} w_i \mathcal{N}_{\eta_i,\Delta_i}$, where $\mathcal{N}_{\eta_i,\Delta_i}$ denotes the Gaussian distribution with mean η_i and variance Δ_i, $w_i \geq 0$ and $\sum_{i=1}^{m} w_i = 1$, the parameter vector $\boldsymbol{\theta} = (\boldsymbol{w}, \boldsymbol{\eta}, \boldsymbol{\Delta})$ We compute the optimal m-component GMM approximation by minimizing the KL-divergence $D(p_T||p_{GMM}(\boldsymbol{\theta}))$ as a function of parameter vector $\boldsymbol{\theta}$. Since

$$\begin{aligned} D(p_T||p_{GMM}(\boldsymbol{\theta})) &:= \sum_{k=1}^{N_s} p_T(k) \log\left(\frac{p_T(k)}{\sum_i^m w_i \mathcal{N}_{\eta_i,\Delta_i}(k)}\right) \\ &= \sum_{k=1}^{N_s} p_T(k) \log p_T(k) - \sum_{k=1}^{N_s} p_T(k) \log\big(\sum_i^m w_i \mathcal{N}_{\eta_i,\Delta_i}(k)\big), \end{aligned}$$

the optimization problem $\min_{\boldsymbol{\theta}} D(p_T||p_{GMM}(\boldsymbol{\theta}))$ is equivalent to

$$\max_{\boldsymbol{\theta}} g(\boldsymbol{\theta}) := \sum_{k=1}^{N_S} p_T(k) \log\left(\sum_{i=1}^{m} w_i \mathcal{N}_{\eta_i,\Delta_i}(k)\right)$$

This is a non-convex optimization. We use an Expectation-Maximization (EM) based iterative heuristic to compute a local maximum. Let $\boldsymbol{\theta}^{(t)}$ denote the current value of the parameters. For each component $i = 1, \ldots, m$, and index $k = 1, \ldots, N_s$, define

$$z_i^{(t)}(k) = \frac{w_i^{(t)} \mathcal{N}_{\eta_i^{(t)},\Delta_i^{(t)}}(k)}{\sum_{j=1}^{m} w_j^{(t)} \mathcal{N}_{\eta_j^{(t)},\Delta_j^{(t)}}(k)}.$$

Then $z_i^{(t)}(k) \geq 0$, and $\sum_{i=1}^{m} z_i^{(t)}(k) = 1$. We interpret $z_i^{(t)}(k)$ as the probability that the count in bin k came from component i. Define

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) = \sum_{k=1}^{N_S} \sum_{i=1}^{m} p_T(k) z_i^{(t)}(k) \log\left(\frac{w_i \mathcal{N}_{\eta_i,\Delta_i}(k)}{z_i^{(t)}(k)}\right)$$

Then we have that

$$Q(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}) = \sum_{k=1}^{N_S} \sum_{i=1}^{m} p_T(k) \hat{z}_i(k) \log\left(\sum_{i=1}^{m} w_i \mathcal{N}_{\eta_i,\Delta_i}(k)\right) = \sum_{k=1}^{N_S} p_T(k) \log\left(\sum_{i=1}^{m} w_i \mathcal{N}_{\eta_i,\Delta_i}(k)\right) = g(\hat{\boldsymbol{\theta}}),$$

and

$$\begin{aligned} g(\boldsymbol{\theta}) &= \sum_{k=1}^{N_S} p_T(k) \log\left(\sum_{i=1}^{m} \frac{w_i \mathcal{N}_{\eta_i,\Delta_i}(k)}{z_i^{(t)}(k)} z_i^{(t)}(k)\right) \\ &\geq \sum_{k=1}^{N_S}\sum_{i=1}^{m} p_T(k) z_i^{(t)}(k) \log\left(\frac{w_i \mathcal{N}_{\eta_i,\Delta_i}(k)}{z_i^{(t)}(k)}\right) = Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}). \end{aligned}$$

Define

$$\theta^{(t+1)} = \arg\max_{\theta}\ Q(\theta, \hat{\theta}) \tag{2.6}$$

Then, we have that

$$g(\theta^{(t+1)}) \geq Q(\theta^{(t+1)}, \theta^{(t)}) \geq Q(\theta^{(t)}, \theta^{(t)}) = g(\theta^{(t)}).$$

Therefore, the iterative algorithm in (2.6) generates a sequence $\{\boldsymbol{\theta}^{(t)} : t \geq 1\}$ with non-decreasing values of g, and the sequence converges to a local maximum. Next, we show that the optimization in (2.6) can be computed efficiently.

1. w-update

$$\boldsymbol{w}^{(t+1)} = \arg\max_{\boldsymbol{w}} \sum_{k=1}^{N_S}\sum_{i=1}^{m} p_T k) z_i^{(t)}(k) \log(w_i) \implies w_i^{(t+1)} = \frac{\sum_{k=1}^{N_S} z_i^{(t)}(k) p_T(k)}{\sum_{i=1}^{m}\sum_{k=1}^{N_S} z_i^{(t)}(k) p_T(k)} \tag{2.7}$$

2. η-update

$$\eta_i^{(t+1)} = \arg\min_{\eta} \sum_{k=1}^{N_S} p_T(k) \hat{z}_i(k) |k - \eta_i|^2 \implies \eta_i^{(t+1)} = \frac{\sum_{k=1}^{N_s} z_i^{(t)}(k) k}{\sum_{k=1}^{N_s} z_i^{(t)}(k)}. \tag{2.8}$$

3. Δ-update

$$\begin{aligned} \Delta_i^{(t+1)} &= \text{argmax}_{\Delta \geq \Delta_{cut}} \left\{ -\frac{\sum_{k=1}^{N_s} p_T(k) z_i^{(t)}(k) |k - \eta_i^{t+1}|^2}{2\Delta} - \log(\Delta) \right\} \\ &= \max\left(\sqrt{\frac{\sum_{k=1}^{N_S} p_T(k) z_i^{(t)}(k) |k - \eta_i^{(t+1)}|^2}{\sum_{k=1}^{N_s} p_T(k) z_i^{(t)}(k)}}, \Delta_{cut} \right), \end{aligned} \tag{2.9}$$

where Δ_{cut} is the minimum allowed width of the Gaussians, in our case $\Delta_{cut} = 1$ since glycan index, $k \in \{1, 2, \ldots N_S\}$, takes integer values with spacing 1.

Since this is a heuristic algorithm for a non convex optimization, we performed several initialization of the algorithm to identify the best local maximum. The KL divergence between the true and GMM approximated($D(p_T||p_{GMM})$), shown in Figure 2.7, saturates at some value of number of components, adding components beyond this only increases model complexity without increasing quality of approximation.

2.5 Future extensions

In future, we will use real single cell mRNA data, instead of the toy model, to estimate the joint probability of protein words and cell types. We are getting open source data from Satija's lab [21] and going to characterize the relevance-compression trade-off for this joint probability. We will then compare the MS glycan profile with the profile that we get from the bottleneck set.

Currently, our framework deals with cell type differentiation in presence of only chemical noise. Another future direction is to extended this framework to include the dynamics of host-pathogen interactions which will further increase the complexity of the glycan distribution. Here we add an extra term in the information bottleneck functional which is the mutual information between the glycans and the host. We want to minimize this mutual information. The probabilities are now time dependent because of being given by the dynamics of host pathogen interaction.

2.6 Conclusion

We started with trying to define complexity of glycan distribution in the cell. In general the notion of complexity of a system depends on the function that the system performs. Here we have described the complexity of the glycan distribution in the cell from the context of glycans being the markers of cell type identity. We use the framework of information bottleneck to describe this system. We find that reliable differentiation of many cell types requires a complex glycan distribution in the sense it has more visual detail and requires more number of Gaussians to fit the

distribution reliably. We plan to extend this framework to include the dynamics of host pathogen interaction and how that affects the glycan complexity.

The complexity of glycans arises due to their role in cell type differentiation. They have to be identified as a proper cell type in a proper niche while avoiding recognition from the pathogens. The glycan profiles can be modulated on a short timescale leading to interesting host pathogen interactions which increases the glycan complexity.

We estimated the complexity of glycan profiles of Human neutrophils and t-cells, planaria and hydra. We find that the human cells have more complex glycan distribution than planaria and hydra suggesting complex organisms have complex glycan profiles.

The cellular machinery required for synthesizing a more complex glycan distribution with high fidelity should be more elaborate than a less complex one. In the next chapter we present a simple model for the synthesis machinery. We then explore the trade-offs in various synthesis costs due to the requirement of high fidelity synthesis of a complex distribution.

Bibliography

[1] Hans-Joachim Gabius. The sugar code: why glycans are so important. *BioSystems*, 164:102–111, 2018.

[2] Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

[3] Peter Grassberger. Randomness, information, and complexity. *arXiv preprint arXiv:1208.3459*, 2012.

[4] David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[5] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[6] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[7] Ajit Varki. Biological roles of glycans. *Glycobiology*, 27(1):3–49, 2017.

[8] Prathyush Pothukuchi, Ilenia Agliarulo, Domenico Russo, Riccardo Rizzo, Francesco Russo, and Seetharaman Parashuraman. Translation of genome to glycome: role of the golgi apparatus. *FEBS letters*, 593(17):2390–2411, 2019.

[9] Frederic Bard and Joanne Chia. Cracking the glycome encoder: signaling, trafficking, and glycosylation. *Trends in cell biology*, 26(5):379–388, 2016.

[10] Giovanni D'Angelo, Serena Capasso, Lucia Sticco, and Domenico Russo. Glycosphingolipids: synthesis and functions. *The FEBS journal*, 280(24):6338–6353, 2013.

[11] A Varki et al. *Essentials of Glycobiology*. Cold Spring Harbor Laboratory Press, 2009.

[12] Noam Slonim and Naftali Tishby. Agglomerative information bottleneck. *Advances in neural information processing systems*, 12, 1999.

[13] A. Bacharoglou. Approximation of probability distributions by convex mixtures of gaussian measures. *Proceedings of the American Mathematical Society*, 138(7):2619–2628, 2010.

[14] Stefan Siebert, Jeffrey A Farrell, Jack F Cazet, Yashodara Abeykoon, Abby S Primack, Christine E Schnitzler, and Celina E Juliano. Stem cell differentiation trajectories in hydra resolved at single-cell resolution. *Science*, 365(6451), 2019.

[15] Christopher T Fincher, Omri Wurtzel, Thom de Hoog, Kellie M Kravarik, and Peter W Reddien. Cell type transcriptome atlas for the planarian schmidtea mediterranea. *Science*, 360(6391), 2018.

[16] Xiaoping Han, Ziming Zhou, Lijiang Fei, Huiyu Sun, Renying Wang, Yao Chen, Haide Chen, Jingjing Wang, Huanna Tang, Wenhao Ge, et al. Construction of a human cell landscape at single-cell level. *Nature*, 581(7808):303–309, 2020.

[17] Peicheng Du, Gustavo Stolovitzky, Peter Horvatovich, Rainer Bischoff, Jihyeon Lim, and Frank Suits. A noise model for mass spectrometry based proteomics. *Bioinformatics*, 24(8):1070–1077, 2008.

[18] Richard D. Cummings and Paul Crocker. Functional Glycomics Database, Consortium for Functional Glycomics. `http://www.functionalglycomics.org`, 2020.

[19] Sabarinath Peruvemba Subramanian, Ponnusamy Babu, Dasaradhi Palakodeti, and Ramaswamy Subramanian. Identification of multiple isomeric core chitobiose–modified high-mannose and paucimannose n-glycans in the planarian schmidtea mediterranea. *Journal of Biological Chemistry*, 293(18):6707–6720, 2018.

[20] Sonu Sahadevan, Aristotelis Antonopoulos, Stuart M Haslam, Anne Dell, Subramanian Ramaswamy, and Ponnusamy Babu. Unique, polyfucosylated glycan–receptor interactions are essential for regeneration of hydra magnipapillata. *ACS chemical biology*, 9(1):147–155, 2014.

[21] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck III, Shiwei Zheng, Andrew Butler, Maddie J. Lee, Aaron J. Wilk, Charlotte Darby,

Michael Zagar, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P. Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar B. Fleming, Bertrand Yeung, Angela J. Rogers, Juliana M. McElrath, Catherine A. Blish, Raphael Gottardo, Peter Smibert, and Rahul Satija. Integrated analysis of multimodal single-cell data. *Cell*, 2021.

Chapter 3

Encoder of the glycan code: Chemical synthesis machinery

In previous chapters we discussed the functions of glycans in the cell and particularly focussed on the role of glycans as the carriers of information. We demostrated that for glycans to faithfully represent the cell type and niche identity, the glycan distribution must be 'complex', i.e. have many well separated detailed peaks. In this chapter we look at the how such a 'complex' glycan distribution can be synthesised in the Golgi compartments. We provide a basic mathematical model for the glycan synthesis machinery which captures some of the salient features of the complex biological process of glycan synthesis. We start with a phenomenological summary of the glycan synthesis and the previous attempts at modeling glycosylation. We then introduce our approach to model glycosylation and subsequently discuss the results of the model.

3.1 Glycosylation

3.1.1 Phenomenology

The glycan display at the cell surface is a result of proteins that flux through and undergo sequential chemical modification in the secretory pathway, comprising an array of Golgi cisternae situated between the Endoplasmic Reticulum (ER) and the Plasma Membrane (PM), as depicted in Fig 3.1. Proteins are delivered from the ER to the first cisterna, whereupon they are processed by the resident enzymes in a sequence of steps that constitute the N-glycosylation process [1] (See Figure 1.5). A

generic enzymatic reaction in the cisterna involves the catalysis of a group transfer reaction in which the monosaccharide moiety of a simple sugar donor substrate, e.g. UDP-Gal, is transferred to the acceptor substrate, by a Michaelis-Menten (MM) type reaction [1]

$$\begin{aligned} \text{Acceptor} + \text{glycosyl donor} + \text{Enzyme} &\underset{\omega_b}{\overset{\omega_f}{\rightleftharpoons}} [\text{Acceptor} \cdot \text{glycosyl donor} \cdot \text{Enzyme}] \\ &\xrightarrow{\omega_c} \text{glycosylated acceptor} + \text{nucleotide} + \text{Enzyme} \end{aligned} \tag{3.1}$$

From the first cisterna, the proteins with attached sugars are delivered to the second cisterna at a given inter-cisternal transfer rate, where further chemical processing catalysed by the enzymes resident in the second cisterna occurs. This chemical processing and inter-cisternal transfer continues until the last cisterna, thereupon the fully processed glycans are displayed at the PM [1]. The network of chemical processing and inter-cisternal transfer forms the basis of the physical model that we will subsequently describe.

3.1.2 Previous detailed computational models

Bailey and coworkers [2], in the first attempt to mathematically model glycosylation, modelled a small glycosylation reaction network using mass action enzyme kinetics with transport. The reaction network consisted of 33 reactions that formed 33 oligosacharides of high mannose, hybrid, hybrid-bisected, complex and complex-bisected types. Each reaction of the central reaction network has an enzyme associated with it and they had 7 enzymes in their model. These enzymes are distributed in three Golgi compartments according the the literature on CHO cells and the enzyme parameters are also taken by from a literature review of glycosylation in CHO cells. They look at the steady state distribution of different types of glycan and found it to be in good agreement with the experimental data on the wild type CHO cells. The model then tries to predict the ratio of various types of glycans on perturbations in enzyme parameters from the wild type and compare it with experiments.

Building on this work, [4] extended the mathematical model to include new types of reaction which were not present in [2], like galactosylation, fucosylation, extension

of branches of the oligosacharides and sialytion. This considerably extended the reaction network to generate 7,565 glycan structures in a network of 22,871 reactions. They accomplished this by assigning enzyme rules to the set of 11 enzymes - these rules decide the substrate and product for a particular enzyme. The enzymes are again distributed in three Golgi compartments. Since then the same group has produced more and more detailed models of glycosylation [3].

A recent study by Ungar and coworkers [5, 6] developed a detailed stochastic simulation for glycosylation, which shows how the overall Golgi transit time and cisternal number, can be tuned to engineer a homogeneous glycan distribution.

These attempts of computationally modeling glycosylation by putting in an ever increasing number of rules, in order to get more and more detailed description of glycosylation, are of great practical importance in the controlled manufacture of pharmaceutical glycans. We, however, are trying to make a conceptual point which does not require the level of detail that is present in these models, on the contrary this level of detail can potentially obscure the point. To this end, we provide an simple model of glycosylation in the following section.

3.2 Basic mathematical model of Glycosylation

We provide a fairly general basic model of glycosylation with which we try to capture the salient features of glycosylation and, try and see how these features affect the glycan profile. The elements of this model are summarized below:

- **Reaction network:** The reaction network takes into account that a diverse set of glycan are sequentially produced from a fixed set precursor glycans coming from the ER. The simplest example of a possible reaction network is a linear one with no branching. Note that the glycans in the linear network can still be branched. We can add more glycosylation features to make the reaction network more realistic like addition of branching, pruning and capping. Each of these features will change the topology of the reaction network, e.g. most general glycan branching will add branchesin the reaction network topology, pruning will give rise to cycles and capping will differentially saturate the length of branches of the network.

- **Transport network:** A model for transfer of glycans from ER to Golgi and from one Golgi cisternae to another. The detailed vesicular transport system is too complex for our analysis but at the appropriate scale the complexity of the transport mechanism can be coarse-grained. The probability of a vesicle coming from one compartment to another can be approximated by exponential distribution with some rate.

- **Model for enzyme kinetics:** Glycosylation reactions are carried out by enzymes, both glycosyltranseferase and glycosidases, which show some level of substrate promiscuity. The enzyme kinetics should therefore incorporate this feature.

- **Golgi compartments:** Golgi compartments are homogenized reaction chambers each with a distinct chemical environment, e.g. pH, which affects the efficiency of enzyme reactions in a cisterna. This can be potentially facilitate cisternal localisation of Golgi enzymes [1].

A family of synthesis models can be created with these basic elements. In the following sections we describe one simple model out of this family. We first describe the reaction and transport network and then describe the enzyme kinetics model.

3.2.1 Reaction and transport network

We consider an array of N_C Golgi cisternae, labeled by $j = 1, \ldots, N_C$, between the ER and the PM (Fig. 2). Proteins, denoted as $\mathcal{P}c_1^{(1)}$, are delivered from the ER to cisterna-1 at an injection rate q. It is well established that the concentration of the glycosyl donor in the j-th cisterna is chemostated [1, 7, 8, 9], thus in our model we hold its concentration $c_0^{(j)}$ constant in time for the j-th cisterna. The acceptor $\mathcal{P}c_1^{(1)}$ reacts with $c_0^{(1)}$ to form the glycosylated acceptor $\mathcal{P}c_2^{(1)}$, following an MM-reaction (3.1) catalysed by the appropriate enzyme. The acceptor $\mathcal{P}c_2^{(1)}$ has the potential of being transformed into $\mathcal{P}c_3^{(1)}$, and so on, provided the requisite enzymes are present in that cisterna. This leads to the sequence of enzymatic reactions $\mathcal{P}c_1^{(1)} \to \mathcal{P}c_2^{(1)} \to \ldots \mathcal{P}c_k^{(1)} \to \ldots$, where k enumerates the sequence of glycosylated acceptors, using a consistent scheme (such as in [2]). We assume that this sequence of reactions stops at a finite value, which we call N_s. For the N-glycosylation pathway

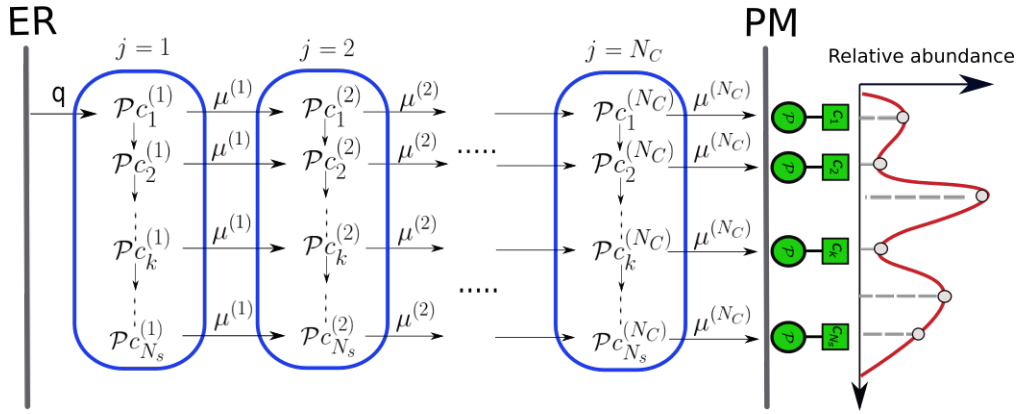


Figure 3.1. Enzymatic reaction and transport network in the secretory pathway. Represented here is the array of Golgi cisternae (blue) indexed by $j = 1, \ldots, N_C$ situated between the ER and PM. Glycan-binding proteins $\mathcal{P}c_1^{(1)}$ are injected from the ER to cisterna-1 at rate q. Superimposed on the Golgi cisternae is transition network of chemical reactions (column) - intercisternal transfer (rows), the latter with rates $\mu^{(j)}$. $\mathcal{P}c_k^{(j)}$ denotes the acceptor substrate in compartment j and the glycosyl donor c_0 is chemostated in each cisterna. This results in a distribution (relative abundance) of glycans displayed at the PM (red curve), that is representative of the cell type.

in a typical mammalian cell, $N_s = 2 \times 10^4$ and N_C = 4-8 [2, 3, 4, 6]. We denote the reaction rate of formation of glycan k in compartment j, $\mathcal{P}c_k^{(j)}$, by $R_{eff}(j,k)$. The glycosylated proteins are transported from cisterna-1 to cisterna-2 at an intercisternal transfer rate $\mu^{(1)}$, whereupon similar enzymatic reactions proceed. The processes of intra-cisternal chemical reactions and inter-cisternal transfer continue to the other cisternae and form a network as depicted in Fig. 3.1. Although, in this thesis, we focus on a sequence of reactions that form a line-graph, the methodology we propose extends to tree-like reaction sequences, and more generally to reaction sequences that form a directed acyclic graphs [10].

We now add to this chemical reaction kinetics, the rates of injection (q) and intercisternal transport $\mu^{(j)}$ from the cisterna j to $j+1$; the complete set of equations that describe the changes in the substrate concentrations $c_k^{(j)}$ with time in the following. These kinetic equations automatically obey the conservation law for the protein

concentration (p).

$$\begin{aligned}
\frac{dc_1^{(1)}}{dt} &= q - R_{eff}(1,1)c_1^{(1)} - \mu^{(1)}c_1^{(1)} \\
\frac{dc_k^{(1)}}{dt} &= R_{eff}(1,k-1)c_{k-1}^{(1)} - R_{eff}(1,k)c_k^{(1)} - \mu^{(1)}c_k^{(1)} \\
\frac{dc_{N_s}^{(1)}}{dt} &= R_{eff}(1,N_s-1)c_{N_s-1}^{(1)} - \mu^{(1)}c_{N_s}^{(1)}
\end{aligned} \tag{3.2}$$

for cisterna-1, and

$$\begin{aligned}
\frac{dc_1^{(j)}}{dt} &= \mu^{(j-1)}c_1^{(j-1)} - R_{eff}(j,1)c_1^{(j)} - \mu^{(j)}c_1^{(j)} \\
\frac{dc_k^{(j)}}{dt} &= \mu^{(j-1)}c_k^{(j-1)} + R_{eff}(j,k-1)c_{k-1}^{(j)} - R_{eff}(j,k)c_k^{(j)} - \mu^{(j)}c_k^{(j)} \\
\frac{dc_{N_s}^{(j)}}{dt} &= \mu^{(j-1)}c_{N_s}^{(j-1)} + R_{eff}(j,N_s-1)c_{N_s-1}^{(j)} - \mu^{(j)}c_{N_s}^{(j)}
\end{aligned} \tag{3.3}$$

for cisternae $j = 2, 3, \ldots, N_C$. These set of dynamical equations (3.2)-(3.3), with initial conditions, can be solved to obtain the concentration $c_k^{(j)}(t)$ for $t \geq 0$. Equations (3.2)-(3.3) automatically obey the conservation law for the protein concentration (p), i.e., the total protein concentration $p^{(j)} = \sum_{k'=1}^{N_s} c_{k'}^{(j)}$ in the j-th cisterna automatically satisfies,

$$\begin{aligned}
\frac{dp^{(1)}}{dt} &= q - \mu^{(1)}p^{(1)} \\
\frac{dp^{(j)}}{dt} &= \mu^{(j-1)}p^{(j-1)} - \mu^{(j)}p^{(j)}
\end{aligned}$$

for $j = 2, 3, \ldots N_C$.

At steady state, the left hand side of in equations (3.2)-(3.3) is set to zero, which after rescaling the kinetic parameters in terms of the injection rate q, i.e. $R_{eff}(j,k) = R_{eff}(j,k)/q$ and $\mu^{(j)} = \mu^{(j)}/q$, gives the following recursion relations for the steady

state concentrations of the glycans in each cisterna. In the first cisterna,

$$\begin{aligned} c_1^{(1)} &= \frac{1}{\mu^{(1)} + R_{eff}(1,1)} \\ c_k^{(1)} &= \frac{R_{eff}(j,k-1)c_{k-1}^{(1)}}{\mu^{(1)} + R_{eff}(j,k)} \\ c_{N_s}^{(1)} &= \frac{R_{eff}(j,N_s-1)c_{N_s-1}^{(1)}}{\mu^{(1)}} \end{aligned} \tag{3.4}$$

and in cisternae $j \geq 2$,

$$\begin{aligned} c_1^{(j)} &= \frac{\mu^{(j-1)}c_1^{(j-1)}}{\mu^{(j)} + R_{eff}(j,1)} \\ c_k^{(j)} &= \frac{\mu^{(j-1)}c_k^{(j-1)} + R_{eff}(j,k-1)c_{k-1}^{(j)}}{\mu^{(j)} + R_{eff}(j,k)} \\ c_{N_s}^{(j)} &= \frac{\mu^{(j-1)}c_{N_s}^{(j-1)} + R_{eff}(j,N_s-1)c_{N_s-1}^{(j)}}{\mu^{(j)}} \end{aligned} \tag{3.5}$$

Equations (3.10)-(3.11) automatically imply that the total steady state glycan concentration in each cisterna $j = 1, \ldots, N_c$ is given by

$$\sum_{k=1}^{N_s} c_k^{(j)} = \frac{1}{\mu^{(j)}}.$$

The steady state mass glycan concentrations depend on the transport rate $\boldsymbol{\mu}$ and the reaction rate $\boldsymbol{R}_{eff}$ where the reaction rate $\boldsymbol{R}_{eff}$ is a result of the enzymes kinetics of glycosylation enzymes. In the following section we provide a model for the kinetics of glycosylation enzymes.

3.2.2 Model for enzyme action

Let the glycosylation reactions in each cisterna j, catalysed by enzymes labeled as $E_\alpha^{(j)}$, with $\alpha = 1, \ldots, N_E$, where N_E is the total number of enzyme species in each cisterna. Since many substrates can compete for the substrate binding site on each enzyme, one expects in general that $N_s \gg N_E$. The configuration space of the network in Fig. 3.1 is $N_s \times N_C$. For the N-glycosylation pathway in a typical

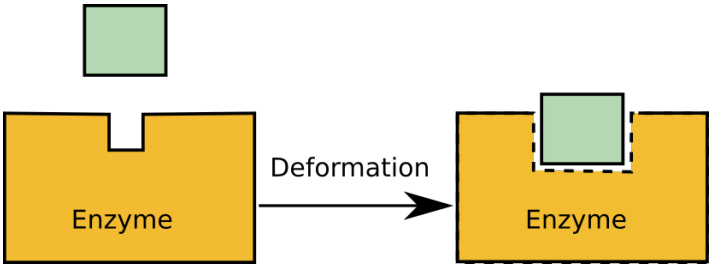


Figure 3.2. Schematic for the enzyme model showing induced fit model for a fixed substrate and flexible enzyme.

mammalian cell, $N_s = 2 \times 10^4$, N_E = 10-20 and N_C = 4-8 [2, 3, 4, 6]. We account for the fact that the enzymes have specific cisternal localisation, by setting their activities to zero in those cisternae where they are not present.

The action of enzyme $E_\alpha^{(j)}$ on the substrate $\mathcal{P}c_k^{(j)}$ in cisterna j is given by

$$\mathcal{P}c_k^{(j)} + E_\alpha^{(j)} \underset{\omega_b(j,k,\alpha)}{\overset{\omega_f(j,k,\alpha)c_0^{(j)}}{\rightleftharpoons}} \left[E_\alpha^{(j)} - \mathcal{P}c_k^{(j)} - c_0^{(j)}\right] \xrightarrow{\omega_c(j,k,\alpha)} \mathcal{P}c_{k+1}^{(j)} + E_\alpha^{(j)} \tag{3.6}$$

where $k = 1, \ldots N_s - 1$. In general, the forward, backward and catalytic rates ω_f, ω_b and ω_c, respectively, depend on the cisternal label j, the reaction label k, and the enzyme label α, that parametrize the MM-reactions [11]. For instance, structural studies on glycosyltransferase-mediated synthesis of glycans [12], would suggest that the forward rate ω_f to depend on the binding energy of the enzyme $E_\alpha^{(j)}$ to acceptor substrate $\mathcal{P}c_k^{(j)}$ and a *physical variable* that characterizes the cisternae.

A potential candidate for such a cisternal variable is pH [13], whose value is maintained homeostatically in each cisterna [14]; changes in pH can affect the shape of an enzyme (substrate) or their charge properties, and in general the reaction efficiency of an enzyme has a pH optimum [11]. Another possible candidate for a cisternal variable is membrane bilayer thickness [15] - indeed both pH [16] and membrane thickness are known to have a gradient across the Golgi cisternae. We take $\omega_f(j,k,\alpha) \propto P^{(j)}(k,\alpha)$, where $P^{(j)}(k,\alpha) \in (0,1)$, is the binding probability of enzyme $E_\alpha^{(j)}$ with substrate $\mathcal{P}c_k^{(j)}$, and define the binding probability $P^{(j)}(k,\alpha)$ using a biophysical model, similar in spirit to the Monod-Wyman-Changeux model of enzyme kinetics [17, 18] that depends on enzyme-substrate induced fit (See Figure 3.2).

3. Encoder of the glycan code: Chemical synthesis machinery

Let $\boldsymbol{\ell}_\alpha^{(j)}$ and $\boldsymbol{\ell}_k$ denote, respectively, the optimal "shape" for enzyme $E_\alpha^{(j)}$ and the substrate $\mathcal{P}c_k^{(j)}$. We assume that the mismatch (or distortion) energy between the substrate k and enzyme $E_\alpha^{(j)}$ is $\|\boldsymbol{\ell}_k - \boldsymbol{\ell}_\alpha^{(j)}\|$, with a binding probability given by,

$$P^{(j)}(k,\alpha) = \exp\left(-\sigma_\alpha^{(j)}\|\boldsymbol{\ell}_k - \boldsymbol{\ell}_\alpha^{(j)}\|\right) \tag{3.7}$$

where $\|\cdot\|$ is a distance metric defined on the space of $\boldsymbol{\ell}_\alpha^{(j)}$ (e.g., the square of the ℓ_2-norm would be related to an elastic distortion model [19]) and the vector $\boldsymbol{\sigma} \equiv [\sigma_\alpha^{(j)}]$ parametrizes enzyme *specificity*. This distortion model captures the above idea that the reaction between the flexible enzyme and fixed substrate is facilitated by an induced fit. A large value of $\sigma_\alpha^{(j)}$ indicates a highly specific enzyme, a small value of $\sigma_\alpha^{(j)}$ indicates a promiscuous enzyme. It is recognized that the degree of enzyme specificity or sloppiness is an important determinant of glycan distribution [1, 20, 21, 22].

Our synthesis model is mean-field, in that we ignore stochasticity in glycan synthesis that may arise from low copy numbers of substrates and enzymes, multiple substrates competing for the same enzymes, and kinetics of inter-cisternal transfer [2, 3, 4]. Then the usual MM-steady state conditions for (3.6), which assumes that the concentration of the intermediate enzyme-substrate complex does not change with time, imply that

$$\left[E_\alpha^{(j)} - \mathcal{P}c_k^{(j)} - c_0^{(j)}\right] = \frac{\omega_f(j,k,\alpha)\, c_0^{(j)}}{\omega_b(j,k,\alpha) + \omega_c(j,k,\alpha)} E_\alpha^{(j)} c_k^{(j)}.$$

where $c_k^{(j)}$ is the *concentration* of the acceptor substrate $\mathcal{P}c_k^{(j)}$ in compartment j.

Together with the constancy of the total enzyme concentration, $\left[E_\alpha^{(j)}\right]_{tot} = E_\alpha^{(j)} + \sum_{k=1}^{N_s}\left[E_\alpha^{(j)} - \mathcal{P}c_k^{(j)} - c_0^{(j)}\right]$, this immediately fixes the kinetics of product formation (not including inter-cisternal transport),

$$\frac{dc_{k+1}^{(j)}}{dt} = \sum_{\alpha=1}^{N_E} \frac{V(j,k,\alpha)P^{(j)}(k,\alpha)c_k^{(j)}}{M(j,k,\alpha)\left(1 + \sum_{k'=1}^{N_s}\frac{P^{(j)}(k',\alpha)c_{k'}^{(j)}}{M(j,k',\alpha)}\right)} \qquad k = 1,\ldots,N_S;\;\; j = 1,\ldots,N_C \tag{3.8}$$

where

$$M(j,k,\alpha) = \frac{\omega_b(j,k,\alpha) + \omega_c(j,k,\alpha)}{\omega_f(j,k,\alpha)c_0^{(j)}} P^{(j)}(k,\alpha)$$

and

$$V(j,k,\alpha) = \omega_c(j,k,\alpha)\left[E_\alpha^{(j)}\right]_{tot} .$$

This reparametrization of the reaction rates $\omega_f, \omega_b, \omega_c$ in terms of $\mathbf{M}, \mathbf{V}$ is convenient, since it relates to experimentally measurable parameters V_{max} and MM-constant K_M, for each (j,k,α) which can be easily read out. As is the usual case, the maximum velocity V_{max} is not an intrinsic property of the enzyme, because it is dependent on the enzyme concentration $\left[E_\alpha^{(j)}\right]_{tot}$; while $K_M(j,k,\alpha) = M(j,k,\alpha)c_0^{(j)}/P^{(j)}(k,\alpha)$ is an intrinsic parameter of the enzyme and the enzyme-substrate interaction. The enzyme catalytic efficiency, the so-called "k_{cat}/K_M" $\propto P^{(j)}(k,\alpha)$ and is high for *perfect* enzymes [23] with minimum mismatch.

$$R_{eff}(j,k) = \sum_{\alpha=1}^{N_E} \frac{V(j,k,\alpha)P^{(j)}(k,\alpha)}{M(j,k,\alpha)\left(1 + \sum_{k'=1}^{N_s} \frac{P^{(j)}(k',\alpha)c_{k'}^{(j)}}{M(j,k',\alpha)}\right)} \qquad k = 1,\ldots,N_S;\;\; j = 1,\ldots,N_C \tag{3.9}$$

Having defined the model for obtaining $\boldsymbol{R}_{eff}$, we are now in a position to obtain the steady state glycan profiles and see the effect of various reaction and transport parameters on the steady state glycan concentration profile.

3.3 Steady state concentrations of glycans

The steady state glycan concentrations are obtained by putting the value of $\boldsymbol{R}_{eff}$ given by Eq. 4.3 into the steady state equations, Eq. 3.4 and 3.5. In the first cisterna,

$$\begin{aligned}
c_1^{(1)} &= \frac{1}{\mu^{(1)} + \sum_{\alpha=1}^{N_E} \frac{V(1,1,\alpha)P^{(1)}(1,\alpha)c_1^{(1)}}{M(1,1,\alpha)\left(1+\sum_{k'=1}^{N_s} \frac{P^{(1)}(k',\alpha)c_{k'}^{(1)}}{M(1,k',\alpha)}\right)}} \\
c_k^{(1)} &= \frac{\sum_{\alpha=1}^{N_E} \frac{V(1,k-1,\alpha)P^{(1)}(k-1,\alpha)c_{k-1}^{(1)}}{M(1,k-1,\alpha)\left(1+\sum_{k'=1}^{N_s} \frac{P^{(1)}(k',\alpha)c_{k'}^{(1)}}{M(1,k',\alpha)}\right)}}{\mu^{(1)} + \sum_{\alpha=1}^{N_E} \frac{V(1,k,\alpha)P^{(1)}(k,\alpha)c_k^{(1)}}{M(1,k,\alpha)\left(1+\sum_{k'=1}^{N_s} \frac{P^{(1)}(k',\alpha)c_{k'}^{(1)}}{M(1,k',\alpha)}\right)}} \\
c_{N_s}^{(1)} &= \frac{\sum_{\alpha=1}^{N_E} \frac{V(1,N_s-1,\alpha)P^{(1)}(N_s-1,\alpha)c_{N_s-1}^{(1)}}{M(1,N_s-1,\alpha)\left(1+\sum_{k'=1}^{N_s} \frac{P^{(1)}(k',\alpha)c_{k'}^{(1)}}{M(1,k',\alpha)}\right)}}{\mu^{(1)}}
\end{aligned} \tag{3.10}$$

and in cisternae $j \geq 2$,

$$\begin{aligned}
c_1^{(j)} &= \frac{\mu^{(j-1)}c_1^{(j-1)}}{\mu^{(j)} + \sum_{\alpha=1}^{N_E} \frac{V(j,1,\alpha)P^{(j)}(1,\alpha)c_1^{(j)}}{M(j,1,\alpha)\left(1+\sum_{k'=1}^{N_s} \frac{P^{(j)}(k',\alpha)c_{k'}^{(j)}}{M(j,k',\alpha)}\right)}} \\
c_k^{(j)} &= \frac{\mu^{(j-1)}c_k^{(j-1)} + \sum_{\alpha=1}^{N_E} \frac{V(j,k-1,\alpha)P^{(j)}(k-1,\alpha)c_{k-1}^{(j)}}{M(j,k-1,\alpha)\left(1+\sum_{k'=1}^{N_s} \frac{P^{(j)}(k',\alpha)c_{k'}^{(j)}}{M(j,k',\alpha)}\right)}}{\mu^{(j)} + \sum_{\alpha=1}^{N_E} \frac{V(j,k,\alpha)P^{(j)}(k,\alpha)c_k^{(j)}}{M(j,k,\alpha)\left(1+\sum_{k'=1}^{N_s} \frac{P^{(j)}(k',\alpha)c_{k'}^{(j)}}{M(j,k',\alpha)}\right)}} \\
c_{N_s}^{(j)} &= \frac{\mu^{(j-1)}c_{N_s}^{(j-1)} + \sum_{\alpha=1}^{N_E} \frac{V(j,N_s-1,\alpha)P^{(j)}(N_s-1,\alpha)c_{N_s-1}^{(j)}}{M(j,N_s-1,\alpha)\left(1+\sum_{k'=1}^{N_s} \frac{P^{(j)}(k',\alpha)c_{k'}^{(j)}}{M(j,k',\alpha)}\right)}}{\mu^{(j)}}
\end{aligned} \tag{3.11}$$

Equations (3.10)-(3.11) automatically imply that the total steady state glycan concentration in each cisterna $j = 1, \ldots, N_c$ is given by

$$\sum_{k=1}^{N_s} c_k^{(j)} = \frac{1}{\mu^{(j)}}.$$

The steady state glycan concentrations, $\boldsymbol{c} \equiv c_k^{(j)}$, as a function of the independent vectors $\boldsymbol{M} \equiv [M(j,k,\alpha)]$, $\boldsymbol{V} \equiv [V(j,k,\alpha)]$, and $\boldsymbol{L} \equiv \|\boldsymbol{\ell}_k - \boldsymbol{\ell}_\alpha^{(j)}\|$, the transport rates $\boldsymbol{\mu} \equiv [\mu^{(j)}]$ and specificity, $\boldsymbol{\sigma} \equiv [\sigma_\alpha^{(j)}]$. While it is possible to numerically solve the above nonlinear recursion relation to obtain the steady state glycan profile for a given value of the parameters $(\boldsymbol{M}, \boldsymbol{V}, \boldsymbol{L}, \boldsymbol{\sigma}, \boldsymbol{\mu})$, it is extremely computationally expensive; making exploration in the high dimension space of the parameters prohibitive expensive. In the following section, we give a different formulation of the steady state glycan profiles in terms of new parameters $(\boldsymbol{R}, \boldsymbol{L}, \boldsymbol{\sigma}, \boldsymbol{\mu})$, in which the concentrations are given by a linear recursion relation. We then prove that both the formulations are equivalent in the sense that the set of all possible glycan profiles in both the formulations is the same. Therefore, instead of exploring in the computationally expensive, $(\boldsymbol{M}, \boldsymbol{V}, \boldsymbol{L}, \boldsymbol{\sigma}, \boldsymbol{\mu})$ space, we can explore in the much cheaper, $(\boldsymbol{R}, \boldsymbol{L}, \boldsymbol{\sigma}, \boldsymbol{\mu})$ space.

3.3.1 Equivalent formulation of the steady state glycan profiles

Define a new set of parameters,

$$R(j,k,\alpha) = \frac{V(j,k,\alpha)}{M(j,k,\alpha)\left(1 + \sum_{k'=1}^{N_s} \frac{P^{(j)}(k',\alpha)c_{k'}^{(j)}}{M(j,k',\alpha)}\right)} \tag{3.12}$$

where $\mathbf{c}$ denotes the steady state glycan concentration corresponding to a specific $(\boldsymbol{M}, \boldsymbol{V}, \boldsymbol{L}, \boldsymbol{\sigma}, \boldsymbol{\mu})$. Define $\mathbf{v}$ by the following set of linear equations:

$$\begin{aligned} v_1^{(1)} =& \frac{1}{\mu^{(1)} + \sum_{\alpha=1}^{N_E} R(1,1,\alpha)P^{(1)}(1,\alpha)} \\ v_k^{(1)} =& \frac{v_{k-1}^{(1)} \sum_{\alpha=1}^{N_E} R(1,k-1,\alpha)P^{(1)}(k-1,\alpha)}{\mu^{(1)} + \sum_{\alpha=1}^{N_E} R(1,k,\alpha)P^{(1)}(k,\alpha)} \\ v_{N_s}^{(1)} =& \frac{v_{N_s-1}^{(1)} \sum_{\alpha=1}^{N_E} R(1,N_s-1,\alpha)P^{(1)}(N_s-1,\alpha)}{\mu^{(1)}} \end{aligned} \tag{3.13}$$

for $j = 1$, and

$$
\begin{aligned}
v_1^{(j)} &= \frac{v_1^{(j-1)}\mu^{(j-1)}}{\mu^{(j)} + \sum_{\alpha=1}^{N_E} R(j,1,\alpha)P^{(j)}(1,\alpha)} \\
v_k^{(j)} &= \frac{v_k^{(j-1)}\mu^{(j-1)}}{\mu^{(j)} + \sum_{\alpha=1}^{N_E} R(j,k,\alpha)P^{(j)}(k,\alpha)} \\
&\quad + \frac{v_{k-1}^{(j)} \sum_{\alpha=1}^{N_E} R(j,k-1,\alpha)P^{(j)}(k-1,\alpha)}{\mu^{(j)} + \sum_{\alpha=1}^{N_E} R(j,k,\alpha)P^{(j)}(k,\alpha)} \\
v_{N_s}^{(j)} &= \frac{v_{N_s}^{(j-1)} \sum_{\alpha=1}^{N_E} R(j,N_s-1,\alpha)P^{(j)}(N_s-1,\alpha)}{\mu^{(j)}} + \frac{v_{N_s}^{(j-1)}\mu^{(j-1)}}{\mu^{(j)}}
\end{aligned}
\tag{3.14}
$$

for $j = 2, \ldots, N_C$. Then, by the definition of $\mathbf{R}$ in (3.12), it trivially follows that the steady state concentration $\mathbf{c}$ corresponding to $(\boldsymbol{M}, \boldsymbol{V}, \boldsymbol{L}, \boldsymbol{\sigma}, \boldsymbol{\mu})$ is a solution for (3.13)-(3.14).

Next, we show that for $\mathbf{v}$ obtained from (3.13)-(3.14) for any parameter $(\boldsymbol{R}, \boldsymbol{L}, \boldsymbol{\sigma}, \boldsymbol{\mu})$, there exists parameter $(\boldsymbol{M}, \boldsymbol{V}, \boldsymbol{L}, \boldsymbol{\sigma}, \boldsymbol{\mu})$ such that (3.10)-(3.11) are automatically satisfied when we set $\mathbf{c} = \mathbf{v}$, i.e. $\mathbf{v}$ is the steady state concentration for $(\boldsymbol{M}, \boldsymbol{V}, \boldsymbol{L}, \boldsymbol{\sigma}, \boldsymbol{\mu})$, and vice versa. Let

$$
\mathcal{A} = \left\{ [c_k^{(j)}]_{j,k} : \begin{array}{l} \mu^{(j)} \geq 0, M(j,k,\alpha) \geq 0, V(j,k,\alpha) \geq 0, l_\alpha^{(j)} \geq 0, \\ [c_k^{(j)}]_{jk} \text{ given by (3.10) and (3.11)}, \end{array} \right\}
$$

and let

$$
\mathcal{B} = \left\{ [v_k^{(j)}]_{j,k} : \begin{array}{l} \mu^{(j)} \geq 0, R(j,k,\alpha) \geq 0, l_\alpha^{(j)} \geq 0 \\ [v_k^{(j)}]_{j,k} \text{ given by (3.13) and (3.14)} \end{array} \right\}.
$$

Then, our task is to show that $\mathcal{A} = \mathcal{B}$. Suppose $[c_k^{(j)}]_{j,k} \in \mathcal{A}$. Let $[M(j,k,\alpha)]$, $[V(j,k,\alpha)]$ and $[l_\alpha^{(j)}]$ be the corresponding parameters. Define

$$
R(j,k,\alpha) = \sum_{\alpha=1}^{N_E} \frac{V(j,k,\alpha)}{M(j,k,\alpha)\left(1 + \sum_{k'=1}^{N_s} \frac{P^{(j)}(k',\alpha)c_{k'}^{(j)}}{M(j,k',\alpha)}\right)} \geq 0
$$

Then $[c_k^{(j)}]_{j,k} \in \mathcal{B}$.

Next, suppose $[v_k^{(j)}]_{j,k} \in \mathcal{B}$. Let $[R(j,k,\alpha)]$, $[l_\alpha^{(j)}]$ denote the corresponding parameters. Since $\sum_{k=1}^{N_s} v_k^{(j)} = 1/\mu^{(j)} < \infty$, it follows that $\sum_{k=1}^{N_s} P^{(j)}(k,\alpha)v_k^{(j)} < 1/\mu^{(j)} <$

∞. Thus, there exists parameters $[M(j,k,\alpha)]$, $[V(j,k,\alpha)]$ and $[l_\alpha^{(j)}]$ such that

$$R(j,k,\alpha) = \frac{V(j,k,\alpha)}{M(j,k,\alpha)\left(1+\sum_{k'=1}^{N_s}\frac{P^{(j)}(k',\alpha)v_{k'}^{(j)}}{M(j,k',\alpha)}\right)} \tag{3.15}$$

Therefore, $[v_k^{(j)}]_{j,k}$ satisfy (3.10) and (3.11), i.e. $[v_k^{(j)}]_{j,k} \in \mathcal{A}$.

Thus, the set of all concentration profiles defined by (3.13)-(3.14) as a function of all possible values of the parameters $(\boldsymbol{R, L, \sigma, \mu})$ is identical to the set defined by (3.10)-(3.11) as function of $(\boldsymbol{M, V, L, \sigma, \mu})$. This is a crucial insight, since it allows us to search the entire parameter space using (3.13)-(3.14), where the concentration is known explicitly in terms of $(\boldsymbol{R, L, \sigma, \mu})$.

3.3.2 Representative glycan steady state concentration profiles of the model

We now compute some representative steady state glycan profiles using (3.13)-(3.14) and additionally making the following simplifying assumptions to the general model:

- We ignore the k dependence in $\mathbf{R}$
- We assume that shape function is a scalar (a length), i.e. $\boldsymbol{l}_\alpha^{(j)} = \ell_\alpha^{(j)}$. It further simplifies the algebra to assume that the length of the substrates are integer multiples of a basic unit (which we take to be 1), i.e. $\boldsymbol{\ell}_k = k$. The norm that appears in (3.7) is taken to be the absolute value difference $|l_k - l_\alpha^{(j)}|$.
- We drop the dependence of the specificity on α and j, and take it to be a scalar σ.

These assumptions greatly simplify the algebraic and numerical complexity while still keeping the important features of the model. In Figure 3.3, we plot the glycan profiles, c_k *vs.* k, for a system of one enzyme, one compartment and two enzymes, two compartments. The profiles are calculated for various values of the parameters enzyme rate, $\mathbf{R}$ and enzyme specificity, σ, while keeping enzyme length, $l_\alpha^{(j)}$, and transport rate, μ, fixed. The results in the plots lead us to the following general observations:

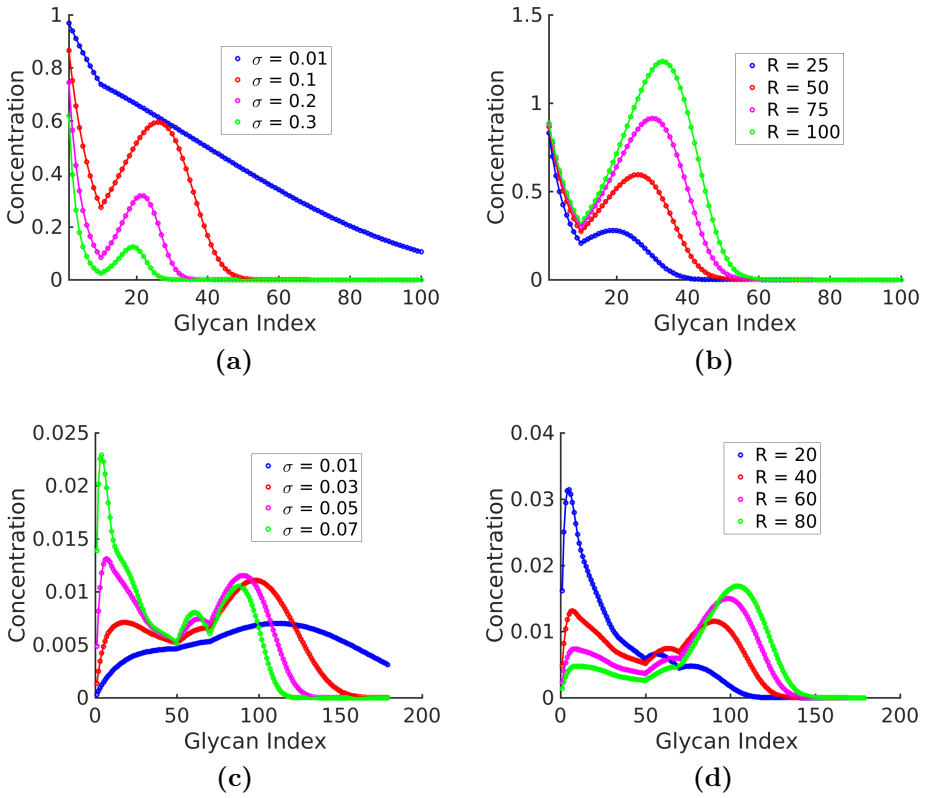


Figure 3.3. Glycan profile $\{c_k : k = 1, \ldots, N_s\}$ as a function of specificity σ ((a) and (c)), and reaction rates R ((b) and (d)).
(a): $N_E = N_C = 1, (R = 50, \mu = 1, l = 10)$. c_k decreases exponentially with k for very low and very high σ; however, the decay rate is lower at low σ. For intermediate values of σ, the distribution has *exactly* two peaks, one of which is at $k = 0$, and eventually decays exponentially. The width of the distribution is a decreasing function of σ.
(b): $N_E = N_C = 1, (\sigma = 0.1, \mu = 1, l = 10)$. At low R, c_k is concentrated at low k. The proportion of higher index glycans in an increasing function of R.
(c): $N_E = N_C = 2, (R = 40, \mu = 1, [l_1^{(1)}, l_2^{(1)}, l_1^{(2)}, l_2^{(2)}] = [10, 30, 50, 70])$. As σ increases, the distribution becomes more complex – from a single peaked distribution at low σ to a maximum of four-peaked distribution at high σ. The peaks gets sharper, and more well defined as σ increases.
(d): $N_E = N_C = 2, (R = 40, \mu = 1, [l_1^{(1)}, l_2^{(1)}, l_1^{(2)}, l_2^{(2)}] = [10, 30, 50, 70])$. As in the plots in (b), increasing R shifts the peaks towards higher index glycans and the proportion of higher index glycan increases.

1. Very low specificity enzymes cannot generate complex glycan distributions. Keeping everything else fixed, intermediate or high specificity enzymes can generate glycan distributions of higher complexity by increasing N_E or N_C (Figures 3.3(a),(c)).

2. Decreasing the specificity σ or increasing the rates R increases the proportion of higher index glycans. Keeping everything else fixed, changes in the rate R have a stronger impact on the relative weights of the higher index glycans to lower index glycans. The relative weight of the higher index glycans increases with increasing N_E and N_C (Figures 3.3(b)-(d)).

3. Keeping everything else fixed, decreasing enzyme specificity increases the spread of the distribution around the peaks. (Figures 3.3(a),(c)).

In the following section we do an analytical calculation, in the limit of large number of glycan species, N_s, to give more intuition about the kinds of glycan profiles that can be generated by the model and the effect of various parameters of the model on the glycan profile.

3.3.3 Analytical calculation in large N_s limit

It is possible to obtain analytical expressions for the steady state glycan distribution in the limit $N_s \gg 1$ when the glycan index k can be approximated by a continuous variable. In this case, (3.10)-(3.11) can be cast as differential equations,

$$\begin{aligned}
&\frac{dc_k^{(1)}}{dk} \approx c_k^{(1)} - c_{k-1}^{(1)} \\
&= \left(\frac{\sum_{\alpha=1}^{N_E} R(1,k-1,\alpha) \exp(-\sigma|k-1-l_\alpha^{(1)}|)}{\mu^{(1)} + \sum_{\alpha=1}^{N_E} R(1,k,\alpha) \exp(-\sigma|k-l_\alpha^{(1)}|)} - 1 \right) c_{k-1}^{(1)} \\
&\approx - \left(\frac{\mu^{(1)} + \frac{d}{dk} \sum_{\alpha=1}^{N_E} R(1,k,\alpha) \exp(-\sigma|k-l_\alpha^{(1)}|)}{\mu^{(1)} + \sum_{\alpha=1}^{N_E} R(1,k,\alpha) \exp(-\sigma|k-l_\alpha^{(1)}|)} \right) c_k^{(1)},
\end{aligned} \tag{3.16}$$

and

$$\begin{aligned}\frac{dc_k^{(j)}}{dk} &\approx c_k^{(j)} - c_{k-1}^{(j)} \\ &= \frac{\mu^{(j-1)}}{\mu^{(j)} + \sum_{\alpha=1}^{N_E} R(j,k,\alpha)\exp(-\sigma|k - l_\alpha^{(j)}|)} c_k^{(j-1)} \\ &\quad - \left(\frac{\mu^{(j)} + \frac{d}{dk}\sum_{\alpha=1}^{N_E} R(j,k,\alpha)\exp(-\sigma|k - l_\alpha^{(j)}|)}{\mu^{(j)} + \sum_{\alpha=1}^{N_E} R(j,k,\alpha)\exp(-\sigma|k - l_\alpha^{(j)}|)} \right) c_k^{(j)}\end{aligned} \tag{3.17}$$

for $j = 2, \ldots, N_C$. In (3.16) and (3.17),

$$\begin{aligned}&\frac{d}{dk}\sum_{\alpha=1}^{N_E} R(j,k,\alpha)\exp(-\sigma|k - l_\alpha^{(j)}|) \\ &\quad = \sum_{\alpha=1}^{N_E} R(j,k,\alpha)\sigma\exp(-\sigma|k - l_\alpha^{(j)}|)(1 - 2\mathbb{I}(k \geq l_\alpha)) + \; R'(j,k,\alpha)\exp(-\sigma|k - l_\alpha^{(j)}|)\end{aligned} \tag{3.18}$$

where the indicator function $\mathbb{I}(\cdot)$ is equal to 1 if the argument is true, and zero otherwise and $R'(j,k,\alpha)$ is the derivative of $R(j,k,\alpha)$ with respect to k.

Define a vector function $C(k) \in \mathbb{R}_c^N$ of the continuous variable k by $C(k) = [c_k^{(1)}, c_k^{(2)}, \ldots c_k^{(N_C)}]$. Then (3.16) and (3.17) can be written as:

$$\frac{dC(k)}{dk} = M(k)C(k) \tag{3.19}$$

where the matrix $M(k)$ is given by

$$M(k) = \begin{bmatrix} A^{(1)}(k) & 0 & 0 & 0 & \ldots 0 \\ B^{(2)}(k) & A^{(2)}(k) & 0 & 0 & \ldots 0 \\ 0 & B^{(3)}(k) & A^{(3)}(k) & 0 & \ldots 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \ldots & 0 & B^{(N_C)}(k) & A^{(N_C)}(k) \end{bmatrix} \tag{3.20}$$

with

$$\begin{aligned} A^{(j)}(k) &= -\frac{\mu^{(j)} + \frac{d}{dk}\sum_{\alpha=1}^{N_E} R(j,k,\alpha)\exp(-\sigma|k - l_\alpha^{(j)}|)}{\mu^{(j)} + \sum_{\alpha=1}^{N_E} R(j,k,\alpha)\exp(-\sigma|k - l_\alpha^{(j)}|)} \\ B^{(j)}(k) &= \frac{\mu^{(j-1)}}{\mu^{(j)} + \sum_{\alpha=1}^{N_E} R(j,k,\alpha)\exp(-\sigma|k - l_\alpha^{(j)}|)} \end{aligned}$$

The functions $A^{(j)}(k)$ and $B^{(j)}(k)$ involve absolute value and indicator functions; therefore, the differential equation has to be solved in a piecewise manner assuming continuity of solution $C(k)$.

The general solution of (3.19)

$$C(k) = C_0 \exp\left(\Omega(k)\right) \tag{3.21}$$

is written in terms of the Magnus Function $\Omega(k) = \sum_{n=1}^{\infty} \Omega(n,k)$, obtained from the Baker-Campbell-Hausdorff formula [24],

$$\begin{aligned} \Omega(1,k) &= \int_0^k M(k_1)dk_1 \\ \Omega(2,k) &= \frac{1}{2}\int_0^k dk_1 \int_0^{k_1} dk_2 \left[M(k_1), M(k_2)\right] \\ \Omega(3,k) &= \frac{1}{6}\int_0^k dk_1 \int_0^{k_1} dk_2 \int_0^{k_2} dk_3 \left[M(k_1), [M(k_2), M(k_3)]\right] + [M(k_3), [M(k_2), M(k_1)]] \\ &\cdots\cdots \end{aligned}$$

where $[M(k_1), M(k_2)] := M(k_1)M(k_2) - M(k_2)M(k_1)$ is the commutator, and the higher order terms in ... contain higher order nested commutators. In Appendix A, we establish conditions under which the series $\sum_{n=1}^{\infty} \Omega(n,k)$ that defines solution $C(k)$ to the differential equation (3.19) converges. Now we solve (3.19) for some special cases.

While in principle we can obtain the glycan profile for any N_E and N_C with arbitrary accuracy, assuming $R(j,k,\alpha) = R_\alpha^{(j)}$, we provide explicit formulae for a few representative cases : (i) $(N_E = 1, N_C = 1)$ and (ii) $(N_E = 1, N_C = 2)$.

(i) $N_E = 1, N_C = 1$: The solution of the differential equation is given by

$$c(k) = \begin{cases} c_0 e^{-k} \left(\frac{\mu + R\exp(-\sigma(l-k))}{\mu + R\exp(-\sigma l)} \right)^{(1/\sigma)-1} & k \leq l \\ c(l) e^{-(k-l)} \left(\frac{\mu + R}{\mu + R\exp(-\sigma(k-l))} \right)^{(1/\sigma)+1} & k > l \end{cases} \tag{3.22}$$

A representative concentration profile is plotted in Figure 3.4(a). The concentration profile consists of two distinct components: an initial exponential decay, and then an exponential rise and fall concentrated around l. The relative weight of these two components is controlled by the sensitivity σ and the rate R. Such explicit formulae can be obtained for any $N_E > 1$, as long as $N_C = 1$.

(ii) $N_E = 1, N_C = 2$: The concentration profile $c^{(2)}$ in cisterna 2 can be obtained from the following calculation. Let $l^{(j)}$ denote the "length" of the enzyme in cisterna $j = 1, 2$. For $k \leq \min\{l^{(1)}, l^{(2)}\}$

$$\begin{aligned} c^{(2)}(k) = c_0 \mu^{(1)} e^{-k} \Big(\frac{\mu^{(2)} + R^{(2)} \exp(-\sigma(l^{(2)} - k))}{\mu^{(1)} + R^{(1)} e^{-\sigma l^{(1)}}} \Big)^{(1/\sigma)-1} \\ \int_0^k \frac{(\mu^{(1)} + R^{(1)} \exp(-\sigma(l^{(1)} - k)))^{(1/\sigma)-1}}{(\mu^{(2)} + R^{(2)} \exp(-\sigma(l^{(2)} - k)))^{1/\sigma}} dk \\ + c^{(2)}(0) e^{-k} \left(\frac{\mu^{(2)} + R^{(2)} e^{-\sigma(l^{(2)} - k)}}{\mu^{(2)} + R^{(2)} e^{-\sigma l^{(2)}}} \right)^{(1/\sigma)-1} \end{aligned}$$

Next, consider the case where $l^{(1)} \leq l^{(2)}$. Then, for $l^{(1)} < k \leq l^{(2)}$

$$\begin{aligned} c^{(2)}(k) &= c^{(1)}(l^{(1)}) \mu^{(1)} e^{-(k-l^{(1)})} (\mu^{(1)} + R^{(1)})^{(1/\sigma)+1} (\mu^{(2)} + R^{(2)} \exp(-\sigma(l^{(2)} - k)))^{(1/\sigma)-1} \\ &\quad \int_{l^{(1)}}^k \frac{(\mu^{(2)} + R^{(2)} \exp(-\sigma(l^{(2)} - k)))^{-1/\sigma}}{(\mu^{(1)} + R^{(1)} \exp(-\sigma(k - l^{(1)})))^{(1/\sigma)+1}} dk \\ &+ c^{(2)}(l^{(1)}) e^{-(k-l^{(1)})} \left(\frac{\mu^{(2)} + R^{(2)} e^{-\sigma(l^{(2)} - k)}}{\mu^{(2)} + R^{(2)} e^{-\sigma(l^{(2)} - l^{(1)})}} \right)^{(1/\sigma)-1} \end{aligned} \tag{3.23}$$

and for $l^{(1)} \leq l^{(2)} < k$,

$$\begin{aligned} c^{(2)}(k) &= c^{(1)}(l^{(1)})\mu^{(1)}e^{-(k-l^{(1)})}\left(\frac{\mu^{(1)}+R^{(1)}}{\mu^{(2)}+R^{(2)}\exp(-\sigma(k-l^{(2)}))}\right)^{(1/\sigma)+1} \\ &\quad \int_{l^{(2)}}^{k}\frac{(\mu^{(2)}+R^{(2)}\exp(-\sigma(k-l^{(2)})))^{1/\sigma}}{(\mu^{(1)}+R^{(1)}\exp(-\sigma(k-l^{(1)})))^{(1/\sigma)+1}}dk \\ &+ c^{(2)}(l^{(2)})e^{-(k-l^{(2)})}\left(\frac{\mu^{(2)}+R^{(2)}}{\mu^{(2)}+R^{(2)}e^{-\sigma(k-l^{(2)})}}\right)^{(1/\sigma)+1} \end{aligned} \tag{3.24}$$

Next, the case where $l^{(1)} \geq l^{(2)}$. For $l^{(2)} < k \leq l^{(1)}$,

$$\begin{aligned} c^{(2)}(k) &= c_0\mu^{(1)}e^{-k}\frac{(\mu^{(1)}+R^{(1)}e^{-\sigma l^{(1)}})^{1-(1/\sigma)}}{(\mu^{(2)}+R^{(2)}\exp(-\sigma(k-l^{(2)})))^{(1/\sigma)+1}} \\ &\quad \int_{l^{(2)}}^{k}\frac{(\mu^{(1)}+R^{(1)}\exp(-\sigma(l^{(1)}-k)))^{(1/\sigma)-1}}{(\mu^{(2)}+R^{(2)}\exp(-\sigma(k-l^{(2)})))^{-1/\sigma}}dk \\ &+ c^{(2)}(l^{(2)})e^{l^{(2)}-k}\left(\frac{\mu^{(2)}+R^{(2)}}{\mu^{(2)}+R^{(2)}e^{-\sigma(k-l^{(2)})}}\right)^{(1/\sigma)+1} \end{aligned} \tag{3.25}$$

For ${}^{(2)} \leq l^{(1)} < k$,

$$\begin{aligned} c^{(2)}(k) &= c^{(1)}(l^{(1)})\mu^{(1)}e^{-(k-l^{(1)})}\left(\frac{\mu^{(1)}+R^{(1)}}{\mu^{(2)}+R^{(2)}\exp(-\sigma(k-l^{(2)}))}\right)^{(1/\sigma)+1} \\ &\quad \int_{l^{(2)}}^{k}\frac{(\mu^{(2)}+R^{(2)}\exp(-\sigma(k-l^{(2)})))^{1/\sigma}}{(\mu^{(1)}+R^{(1)}\exp(-\sigma(k-l^{(1)})))^{(1/\sigma)+1}}dk \\ &+ c^{(2)}(l^{(1)})e^{-(k-l^{(1)})}\left(\frac{\mu^{(2)}+R^{(2)}e^{-\sigma(l^{(1)}-l^{(2)})}}{\mu^{(2)}+R^{(2)}e^{-\sigma(k-l^{(2)})}}\right)^{(1/\sigma)+1} \end{aligned} \tag{3.26}$$

The integrals in (3.23) to (3.26) can evaluated numerically. The result of the numerical computation is shown in Fig. 3.4.

3.4 Extensions of the Glycan synthesis model

- **Retrograde transport:** The transport network that we used was uni-directional but in reality there is backward transport of cargo, called retrograde transport, in the Golgi. We can allow retrograde transport in our model. This increases the complexity of the model considerably and also has some interesting impli-

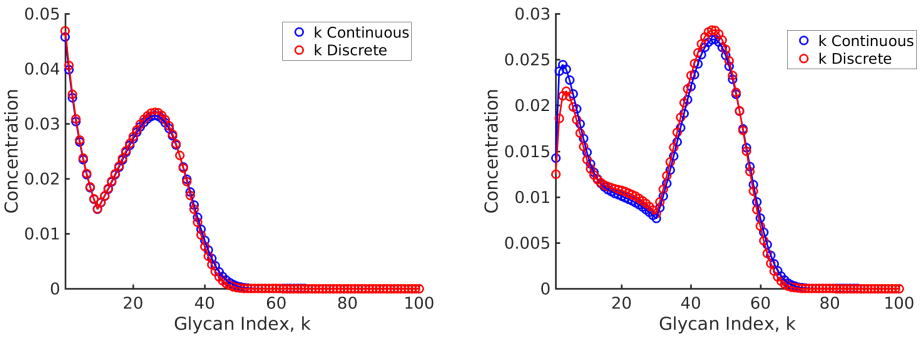


Figure 3.4. Glycan concentration profile calculated from the model using (a) formula (3.22) for $N_E = N_C = 1$ and (b) formulae (3.23)-(3.26) for $N_E = 1, N_C = 2$.

cations. Cargo specificity of the transport in general and retrograde transport in particular. How do the number of peaks in the steady state distribution change on the inclusion of retrograde transport. Combination with capping can be interesting. Error correction and glycan profile sculpting in time varying environment

- **Branching, Pruning and Capping:** Adding features like branching, pruning and capping to the reaction network and observing the effect of these on the steady state glycan profile.

- **Extended enzyme distortion model:** The distortion model we are using for calculating the binding probabilities of substrates with enzymes allows every enzyme α to in principle catalyze any reaction in any cisterna. This allowed for the ideal enzyme length $l_\alpha^{(j)}$ in (3.7) to vary across the cisternae in an unconstrained manner, leading to simplification in the calculation. This is unrealistic and a more reasonable model for the ideal enzyme length is given by

$$\ell_\alpha^{(j)} = \ell_\alpha^{(0)} + \delta\ell_\alpha^{(j)}, \qquad \delta\ell_\alpha^{(j)} \in [-\ell_b^{(j)}, \ell_b^{(j)}],$$

i.e., the nominal length $\ell_\alpha^{(0)}$ can be distorted in a cisterna by a correction $\delta\ell_\alpha^{(j)}$ but within a specified bound $\ell_b^{(j)}$ that is not subject to optimization. One can render some enzymes inactive in certain cisternae by choosing appropriate

values of $\ell_\alpha^{(0)}$ and $\delta\ell_\alpha^{(j)}$. We will see the effect of this extension in Sect. 4.2.1.

3.5 Conclusion

In this chapter we provided a fairly general basic model of glycan synthesis machinery which captures many of its salient features. The elements of this model are - the reaction network, the transport network, a model of enzyme kinetics and chemically distinct compartments of the Golgi complex. This framework allows us to create a family of models of which we describe in detail one of the simplest. In this model, we take a linear reaction network with uni-directional transport (from the ER to PM). The enzyme kinetics is based on the induced fit model of interaction with a deformable enzyme and fixed substrate. The deformability of the enzyme is a parameter which is related to the substrate specificity of the enzyme.

The steady state glycan concentration profile of this model depends on the enzyme parameters - the enzyme rate and enzyme specificty, the transport parameters - the transport rate and chemical environment of the Golgi cisternae. These parameters affect the glycan profile differently - e.g the enzyme specificity controls the typical width of the peaks in the glycan profile.

In the next chapter, we will use this synthesis model to generate glycan profiles similar to those found in real cells and analyze the trade-offs between various components of the synthesis machinery.

We plan to extend the synthesis model in future to include more features of the glycan synthesis like retrograde transport, capping and pruning enzymes.

Bibliography

[1] A Varki et al. *Essentials of Glycobiology*. Cold Spring Harbor Laboratory Press, 2009.

[2] Pablo Umaña and James E Bailey. A mathematical model of n-linked glycoform biosynthesis. *Biotechnology and bioengineering*, 55(6):890–908, 1997.

[3] Frederick J Krambeck, Sandra V Bennun, Someet Narang, Sean Choi, Kevin J Yarema, and Michael J Betenbaugh. A mathematical model to derive n-glycan structures and cellular enzyme activities from mass spectrometric data. *Glycobiology*, 19(11):1163–1175, 2009.

[4] Frederick J Krambeck and Michael J Betenbaugh. A mathematical model of n-linked glycosylation. *Biotechnology and Bioengineering*, 92(6):711–728, 2005.

[5] Peter Fisher, Hannah Spencer, Jane Thomas-Oates, A Jamie Wood, and Daniel Ungar. Modeling glycan processing reveals golgi-enzyme homeostasis upon trafficking defects and cellular differentiation. *Cell reports*, 27(4):1231–1243, 2019.

[6] Peter Fisher and Daniel Ungar. Bridging the gap between glycosylation and vesicle traffic. *Frontiers in cell and developmental biology*, 4:15, 2016.

[7] Carlos B Hirschberg, Phillips W Robbins, and Claudia Abeijon. Transporters of nucleotide sugars, atp, and nucleotide sulfate in the endoplasmic reticulum and golgi apparatus, 1998.

[8] Carolina E Caffaro and Carlos B Hirschberg. Nucleotide sugar transporters of the golgi apparatus: from basic science to diseases. *Accounts of chemical research*, 39(11):805–812, 2006.

[9] Patricia M Berninsone and Carlos B Hirschberg. Nucleotide sugar transporters of the golgi apparatus. *Current opinion in structural biology*, 10(5):542–547, 2000.

[10] Nenad Trinajstic. *Chemical graph theory*. Routledge, 2018.

[11] NC Price and L Stevens. *Fundamentals of Enzymology: The cell and molecular biology of catalytic proteins*. Oxford University Press, 1999.

[12] Kelley W Moremen and Robert S Haltiwanger. Emerging structural insights into glycosyltransferase-mediated synthesis of glycans. *Nature chemical biology*, 15(9):853–864, 2019.

[13] Sakari Kellokumpu. Golgi ph, ion and redox homeostasis: How much do they really matter? *Frontiers in cell and developmental biology*, 7:93, 2019.

[14] Joseph R Casey, Sergio Grinstein, and John Orlowski. Sensors and regulators of intracellular ph. *Nature reviews Molecular cell biology*, 11(1):50, 2010.

[15] Serge Dmitrieff, Madan Rao, and Pierre Sens. Quantitative analysis of intra-golgi transport shows intercisternal exchange for all cargo. *Proceedings of the National Academy of Sciences*, 110(39):15692–15697, 2013.

[16] Juan Llopis, J Michael McCaffery, Atsushi Miyawaki, Marilyn G Farquhar, and Roger Y Tsien. Measurement of cytosolic, mitochondrial, and golgi ph in single living cells with green fluorescent proteins. *Proceedings of the National Academy of Sciences*, 95(12):6803–6808, 1998.

[17] Jacque Monod, Jeffries Wyman, and Jean-Pierre Changeux. On the nature of allosteric transitions: a plausible model. *J Mol Biol*, 12(1):88–118, 1965.

[18] Jean-Pierre Changeux and Stuart J Edelstein. Allosteric mechanisms of signal transduction. *Science*, 308(5727):1424–1428, 2005.

[19] Yonatan Savir and Tsvi Tlusty. Conformational proofreading: the impact of conformational changes on the specificity of molecular recognition. *PloS one*, 2(5):e468, 2007.

[20] Saul Roseman. Reflections on glycobiology. *Journal of Biological Chemistry*, 276(45):41527–41542, 2001.

[21] Patrick Hossler, Bhanu Chandra Mulukutla, and Wei-Shou Hu. Systems analysis of n-glycan processing in mammalian cells. *PloS one*, 2(8), 2007.

[22] Min Yang, Charlie Fehl, Karen V Lees, Eng-Kiat Lim, Wendy A Offen, Gideon J Davies, Dianna J Bowles, Matthew G Davidson, Stephen J Roberts, and Benjamin G Davis. Functional and informatics analysis enables glycosyltransferase activity prediction. *Nature chemical biology*, 14(12):1109–1117, 2018.

[23] Arren Bar-Even, Ron Milo, Elad Noor, and Dan S Tawfik. The moderately efficient enzyme: futile encounters and enzyme floppiness. *Biochemistry*, 54(32):4969–4977, 2015.

[24] Sergio Blanes, Fernando Casas, JA Oteo, and José Ros. The magnus expansion and some of its applications. *Physics Reports*, 470(5-6):151–238, 2009.

Chapter 4

Cellular trade-offs in high fidelity glycan encoding

We demonstrated, in Chapter 2 of the thesis, that complex oragnism have complex glycan distribution, and in Chapter 3 we provided a basic mathematical model of glycosylation. In this chapter, we combine these two together to ask what constraints does the requirement of producing a complex distribution with high fidelity put on the Golgi synthesis machinery. We start with defining fidelity of synthesis in the current context and then discuss the optimization for maximizing fidelity. We follow it up with the results of the optimization showing trade-offs between the fidelity of synthesis, complexity of the glycan profile, number and specificity of enzymes and the number of compartments. We later discuss how does the requirement of producing a diverse repertoire of glycans affect the synthesis machinery.

4.1 Fidelity of synthesis

As discussed in the Chapter 2, glycans are the markers of cell type identity and niche, and distinct cell types (in distinct niche) have distinct glycan profile. We say that there is a "target" glycan profile associated with a cell type which the synthesis machinery of a cell of that cell type should achieve. There will be some variation in the glycan profiles of cells of the same cell type as a result of various kinds of noise in the synthesis machinery and cell to cell variations. Therefore the "target" glycan profile of a cell type is the average profile over many cells of the same cell type.

We obtain the target glycan distribution from glycan profiles for real cells using Mass Spectrometry coupled with determination of molecular structure (MSMS)

measurements [6] on a sample prepared from many cells of the same cell type. The raw MSMS data, however, is not suitable as a target distribution. This is because it is very noisy, with chemical noise in the sample and Poisson noise associated with detecting discrete events being the most relevant [7] as described in Sect. **??**. This means that many of the small peaks in the raw data are not part of the signal, and one has to "smoothen" the distribution to remove the impact of noise.

We use MSMS data from *human* T-cells [6] for our analysis. As discussed in Sect. 2.4, the Gaussian mixture models (GMM) are often used to approximate distributions with a mixed number of modes or peaks [2], or in our setting, a given fixed complexity. Here, we use a variation of the Gaussian mixture models (see Sect. 2.4 for details) to create a hierarchy of increasingly complex distributions to approximate the MSMS raw data. Thus, the 3-GMM and 20-GMM approximations represent the low and high complexity benchmarks, respectively. In Sect. 2.4, we show that the likelihood for the glycan distribution of the *human* T-cell saturates at 20 peaks. Thus, statistically the *human* T-cell glycan distribution is accurately approximated by 20 peaks.

This hierarchy allows us to study the trade-off between the complexity of the target distribution and the complexity of the synthesis model needed to generate the distribution as follows. Let $\mathbf{T}^{(i)}$ denote the i-component GMM approximation for the *human* T-cell MSMS data. We sample this target distribution at indices $k = 1, \ldots, N_s$, that represent the glycan indices, and then renormalize to obtain the discrete distribution $\{T_k^{(i)}, k = 1, \ldots, N_s\}$. To highlight the role of target distribution complexity, we focus on the 3-GMM $\boldsymbol{T}^{(3)}$ (low complexity) and 20-GMM approximation $\boldsymbol{T}^{(20)}$ (high complexity) in the describing our results.

Now we define the fidelity of synthesis using the "target" distribution obtained from MSMS data of real cells. Let $\boldsymbol{c}^*$ denote the "target" concentration distribution, normalized the distribution so that $\sum_{k=1}^{N_s} c_k^* = 1$, for a particular cell type, i.e. the goal of the sequential synthesis mechanism described in Sect. 3.3 is to approximate $\boldsymbol{c}^*$. Let $\bar{\mathbf{c}}$ denote the normalized steady state glycan concentration distribution displayed on the PM. Then (3.14) implies that $\bar{c}_k = \mu^{(N_C)} c_k^{(N_C)}$, $k = 1, \ldots, N_s$. We measure the *fidelity* $F(\boldsymbol{c}^* \| \bar{\boldsymbol{c}})$ between the $\mathbf{c}^*$ and $\bar{\mathbf{c}}$ by the ratio of the Kullback-

Leibler divergence $D(\boldsymbol{c}^*\|\boldsymbol{c})$ [1, 2] to the entropy $H(\boldsymbol{c}^*)$

$$F(\boldsymbol{c}^*\|\bar{\mathbf{c}}) := \frac{D(\boldsymbol{c}^*\|\bar{\mathbf{c}})}{H(\boldsymbol{c}^*)} = \frac{\sum_{k=1}^{N_s} c_k^* \ln\left(\frac{c_k^*}{\bar{c}_k}\right) = \sum_{k=1}^{N_s} c_k^* \ln\left(\frac{c_k^*}{c_k^{(N_C)}\mu^{(N_C)}}\right)}{\sum_{k=1}^{N_s} c_k^* \ln(1/c_k^*)} \tag{4.1}$$

The reason why we divide the KL-divergence by the entropy of the target distribution is to enable comparison of the fidelity of the mechanism across target distributions of different complexity. Note that high fidelity corresponds to low values of $F(\boldsymbol{c}^*\|\bar{\mathbf{c}})$, vice versa.

Thus, the problem of designing a sequential synthesis mechanism that approximates $\mathbf{c}^*$ for a given enzyme specificity $\boldsymbol{\sigma}$, number of enzymes N_E, and number of cisternae N_C is given by

$$\begin{aligned} \textit{Optimization A}: \quad \bar{D}(\boldsymbol{\sigma}, N_E, N_C, \boldsymbol{c^*}) \quad &:= \quad \min_{\boldsymbol{\mu},\ \boldsymbol{R},\ \boldsymbol{L}} F(\boldsymbol{c}^*\|\bar{\mathbf{c}}) \\ &\text{s.t.} \quad \mu_{min} \le \mu^{(j)} \le \mu_{max} \\ &\qquad R_{min} \le R(j,\alpha) \le R_{max} \\ &\qquad 1 \le l_\alpha^{(j)} \le N_s \end{aligned} \tag{4.2}$$

where we emphasize that the optimum fidelity $\bar{D}(\boldsymbol{\sigma}, N_E, N_C, \boldsymbol{c^*})$ is a function of $(\sigma, N_E, N_C, \boldsymbol{c^*})$. The physical bounds on the reaction and transport rates - (R_{min}, R_{max}, μ_{min}, μ_{max}) are estimated from the literature on glycosylation enzymes and Golgi transport times (See Appendix C for details). Note that there is separation of time scales implicit in Optimization A – the chemical kinetics of the production of glycans and their display on the PM happens over cellular time scales, while the issues of trade-offs and changes of parameters are related to evolutionary timescales. In Appendix B, we describe the variant of the Sequential Quadratic Programming (SQP) [3], that we use to numerically solve the optimization problem.

The dimension of the optimization search space is extremely large $\approx O(N_s \times N_E \times N_C)$. To make the optimization search more manageable, we make the following simplifying assumptions on the synthesis model described in Sect. 3.3:

1. We ignore the k-dependence of the vectors $(\mathbf{M}, \mathbf{V})$, or alternatively of $\mathbf{R}$.

2. The enzyme-substrate binding probability $P^{(j)}(k,\alpha)$ is still dependent on the substrate k. We assume that shape function is a scalar (a length), i.e. $\boldsymbol{l}_\alpha^{(j)} = \ell_\alpha^{(j)}$. It further simplifies the algebra to assume that the length of the substrates are integer multiples of a basic unit (which we take to be 1), i.e. $\boldsymbol{\ell}_k = k$. The norm that appears in (3.7) is taken to be the absolute value difference $|l_k - l_\alpha^{(j)}|$. Other metrics, such as $|l_k - l_\alpha^{(j)}|^2$, corresponding to the elastic distortion model [4], do not pose any computational difficulties, and we see that the results of our optimization remain qualitatively unchanged.

3. We drop the dependence of the specificity on α and j, and take it to be a scalar σ. We will explore the effects of enzyme and compartment dependent enzyme specficity in the next chapter.

These restrictions significantly reduce the dimension of the optimization search, making the problem tractable while still retaining the essential features of the model. In Sect. 3.3.3 we show that (3.10)-(3.11) can be solved analytically in the limit $N_s \gg 1$, since the glycan index k can be approximated by a continuous variable, and the recursion relations for the steady state glycan concentrations (3.10)-(3.11) can be cast as a matrix differential equation. This allows us to obtain an *explicit* expression for the steady state concentration in terms of the parameters $(\mathbf{R}, \mathbf{L})$. This helps us obtain some useful heuristics (Sect. 3.3.3) on how to tune the parameters, e.g. N_E, N_C, σ, and others, in order to generate glycan distributions $\mathbf{c}$ of a given complexity. These heuristics inform our more detailed optimization using "realistic" target distributions.

The calculations in Sect. 3.3.3 imply, as one might expect, that the synthesis model needs to be more elaborate, i.e., needs a larger number of cisternae N_C or a larger number of enzymes N_E, in order to produce a more complex glycan distribution. For a real cell type in a niche, the specific elaboration of the synthesis machinery, would depend on a variety of control costs associated with increasing N_E and N_C. While an increase in the number of enzymes would involve genetic and transcriptional costs, the costs involved in increasing the number of cisternae could be rather subtle.

Notwithstanding the relative control costs of increasing N_E and N_C, it is clear from the special case, that increasing the number of cisternae achieves the goal of obtaining an accurate representation of the target distribution. Suppose the target

distribution $c_k^* = \delta(k - M)$ for a fixed $M \gg 1$, i.e. $c_k^* = 1$ when $k = M$, and 0 otherwise, and that the N_E enzymes that catalyse the reactions are highly specific. In this limit, *Optimization A* reduces to a simple enumeration exercise [5]: clearly, one needs $N_E = M$ enzymes, one for each $k = 1, \ldots, M$ reactions, in order to generate $\mathcal{P}c_M$. For a single Golgi cisterna with a finite cisternal residence time (finite μ), the chemical synthesis network will generate a significant steady state concentration of lower index glycans $\mathcal{P}c_k$ with $k < M$, contributing to a low fidelity. To obtain high fidelity, one needs multiple Golgi cisternae with a specific enzyme partitioning $(E_1, E_2, \ldots, E_M)$ with E_j enzymes in cisterna $j = 1, \ldots, N_c$. This argument can be generalized to the case where the target distribution is a finite sum of delta-functions. The more general case, where the enzymes are allowed to have variable specificity, needs a more detailed study, to which we turn to next.

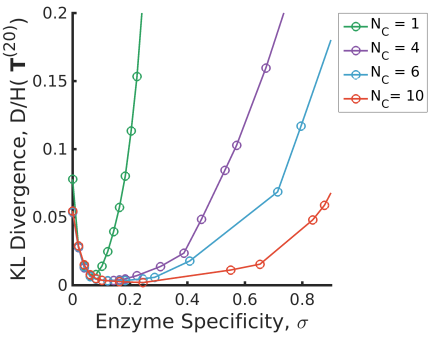
4.2 Trade-offs between fidelity, number and specificity of enzyme, and number of compartments

We summarize the main results that follow from an optimization of the parameters of the glycan synthesis machinery to a given target distribution in Figs. 4.1-4.2.

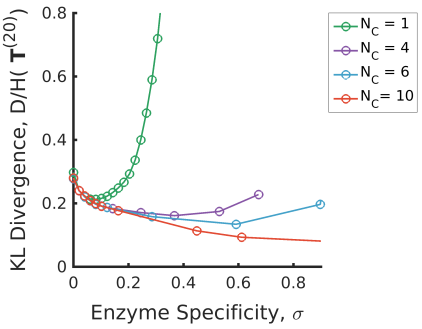
1. The optimal fidelity $\bar{D}(\sigma, N_E, N_C, \mathbf{c}^*)$ is a convex function of σ for fixed values for other parameters (see Fig. 4.1), i.e. it first decreases with σ and then increases beyond a critical value of $\sigma_{\min}$.

   The fidelity $\bar{D}(\sigma, N_E, N_C, \mathbf{c}^*)$ is decreasing in N_C and N_E for fixed values of the other parameters, and increasing in the complexity of $\mathbf{c}^*$ for fixed (σ, N_C). The marginal contribution of N_C and N_E in improving fidelity $\bar{D}$ is approximately equal (see Figs. 4.2a, 4.2b). We discuss the origin of this symmetry later in this section.

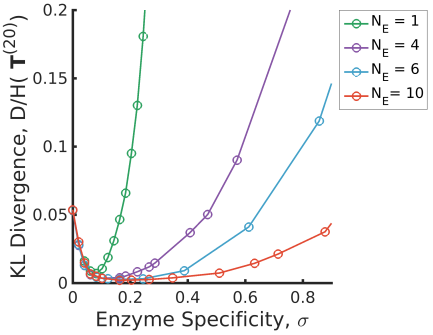
   The lower complexity distributions can be synthesized with high fidelity with small (N_E, N_C), whereas higher complexity distributions require significantly larger (N_E, N_C) (see Figs. 4.2a, 4.2b). For a typical mammalian cell, the number of enzymes in the N-glycosylation pathway are in the range $N_E = 10 -$



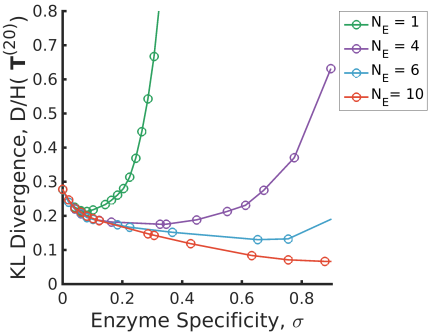
(a) Less complex target, 3-GMM approximation



(b) More complex target, 20-GMM approximation



(c) Less complex target, 3-GMM approximation



(d) More complex target, 20-GMM approximation

Figure 4.1. Trade-offs amongst the glycan synthesis parameters, enzyme specificity σ, cisternal number N_C and enzyme number N_E, to achieve a complex target distribution $\mathbf{c}^*$). (a)-(b) Normalised Kullback-Leibler distance $\bar{D}(\sigma, N_E, N_C, \mathbf{c}^*)$ as function of σ and N_C (for fixed $N_E = 3$), (c)-(d) $\bar{D}(\sigma, N_E, N_C, \mathbf{c}^*)$ as function of σ and N_E (for fixed $N_C = 3$), with the target distribution $\mathbf{c}^*$ set to the 3-GMM (less complex) and 20-GMM (more complex) approximations for the *human* T-cell MSMS data. $\bar{D}(\sigma, N_E, N_C, \mathbf{c}^*)$ is a convex function of σ for each $(N_E, N_C, \mathbf{c}^*)$, decreasing in N_C, N_E for each $(\sigma, \mathbf{c}^*)$, increasing in the complexity of $\mathbf{c}^*$ for fixed (σ, N_E, N_C). The specificity $\sigma_{\min}(\mathbf{c}^*, N_E, N_C) = \text{argmin}_\sigma\{\bar{D}(\sigma, N_E, N_C, \mathbf{c}^*)\}$ that minimises the error for given $(N_E, N_C, \mathbf{c}^*)$ is an increasing function of N_C, N_E and the complexity of the target distribution $\mathbf{c}^*$.

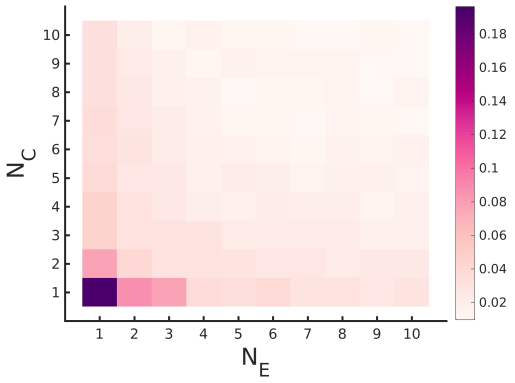
20 [8, 9, 10, 11], Fig. 4.2b would then suggest that the optimal cisternal number would range from $N_C = 3 - 8$ [12].

2. The optimal enzyme specificity $\sigma_{\min}(\mathbf{c}^*, N_C) = \text{argmin}_\sigma\{\bar{D}(\sigma, \bar{N}_E, N_C, \mathbf{c}^*)\}$, that minimises the error as function of $(N_C, \mathbf{c}^*)$ with N_E fixed at $\bar{N}_E$, is an increasing function of N_C and the complexity of the target distribution $\mathbf{c}^*$ (Figs. 4.1a, 4.1b, 4.2c, 4.2d). This is consistent with the results in Appendix 3.3.3 where we established that the width of the synthesized distribution is inversely dependent on the specificity σ: since a GMM approximation with fewer peaks has wider peaks, $\sigma_{\min}$ is low, and vice versa. Similar results hold when N_C is fixed at $\bar{N}_C$, and N_E is varied (see Figs. 4.1c, 4.1d, 4.2c, 4.2d).

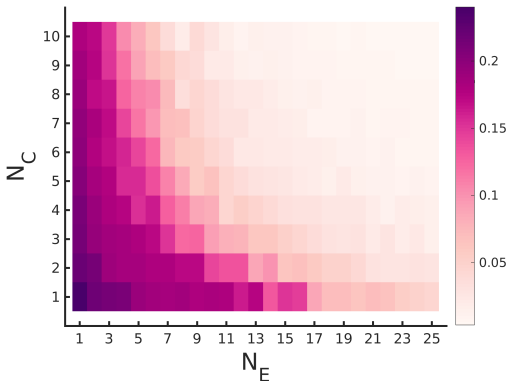
Our results are consistent with those in [13]. They optimize incoming glycan ratio, transport rate and effective reaction rates in order to synthesize a narrow target distribution centred around a desired glycan. The ability to produce specific glycans without much heterogeneity is an important goal in pharmaceutical industry. They define heterogeneity as the total number of glycans synthesized, and show that increasing the number of compartments N_C decreases heterogeneity, and increases the concentration of the specific glycan. They also show that the effect of compartments in reducing heterogeneity cannot be compensated by changing the transport rate. Our results are entirely consistent with theirs - we have shown that $\bar{D}$ decreases as we increase N_C. Thus, if the target distribution has a single sharp peak, increasing N_C will reduce the heterogeneity in the distribution.

4.2.1 Symmetry of the N_E, N_C space

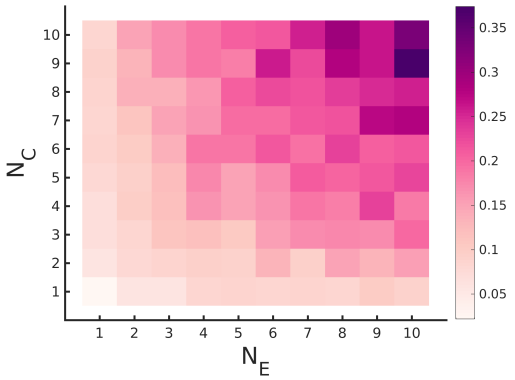
We insert an important cautionary note here. It would seem that the results in Fig. 4.2 imply that there is an approximate $N_E - N_C$ symmetry in the model, i.e. increasing either N_E or N_C affects the fidelity, optimal enzyme specificity and the sensitivity in approximately the same way. This would be an erroneous inference, and is a consequence of the distortion model we have used for calculating the binding probabilities of substrates with enzymes. The root cause for this apparent symmetry is that we have allowed for all enzymes to catalyse reactions in all cisternae (albeit with different efficiencies). This symmetry is violated by simply restricting the



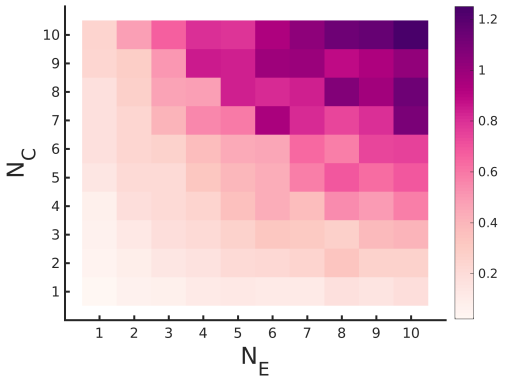
(a) Fidelity for less complex target, $\mathbf{c}^* = 3$-GMM approximation



(b) Fidelity for more complex target $\mathbf{c}^* = 20$-GMM approximation

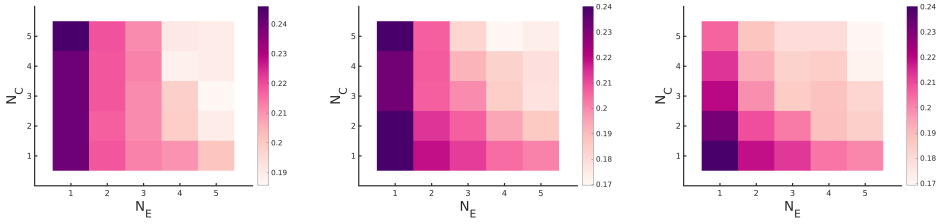


(c) Optimal enzyme specificity for less complex target, $\mathbf{c}^* = 3$-GMM approximation



(d) Optimal enzyme specificity for more complex target $\mathbf{c}^* = 20$-GMM approximation

Figure 4.2. Fidelity of glycan distribution and optimal enzyme properties to achieve a complex target distribution. The target $\mathbf{c}^*$ is taken from 3-GMM (less complex) and 20-GMM (more complex) approximations of the *human* T-cell MSMS data. (a)-(b) Optimum fidelity $\min_\sigma\{\bar{D}(\sigma, N_C, N_E, \mathbf{c}^*)\}$ as a function of (N_E, N_C). More complex distributions require either a larger N_E or N_C. The marginal impact of increasing N_E and N_C on the fidelity $\bar{D}$ is approximately equal. (c)-(d) Enzyme specificity $\sigma_{\min}$ that achieves $\min_\sigma\{\bar{D}(\sigma, N_C, N_E, \mathbf{c}^*)\}$ as a function of (N_E, N_C). $\sigma_{\min}$ increases with increasing N_E or N_C. To synthesize the more complex 20 GMM approximation with high fidelity requires enzymes with higher specificity $\sigma_{\min}$ compared to those needed to synthesize the broader, less complex 3-GMM approximation.



(a) $l_b/N_S = 0.01$ **(b)** $l_b/N_S = 0.1$ **(c)** $l_b/N_S = 0.5$

Figure 4.3. Optimum fidelity $\bar{D}_{KL}$ as a function of (N_E, N_C) for different values of ℓ_b/N_s, where ℓ_b bounds the deformation in the ideal length $\ell_\alpha^{(0)}$ of an enzyme $\alpha = 1, \ldots, N_E$. Small values of ℓ_b restricts all enzymes from working in all cisternae and all substrates, where large value of ℓ_b removes this constraint.

activity of the enzymes to be dependent on the cisternae. We describe a simple realisation of this below.

The distortion model we are using for calculating the binding probabilities of substrates with enzymes allows every enzyme α to in principle catalyse any reaction in any cisterna. This allowed for the ideal enzyme length $l_\alpha^{(j)}$ in Equation 3.7 to vary across the cisternae in an unconstrained manner, leading to simplification in the calculation. We now find that by changing this aspect of the model, the apparent symmetry between $N_E - N_C$ is lifted. A more reasonable model for the ideal enzyme length is given by

$$\ell_\alpha^{(j)} = \ell_\alpha^{(0)} + \delta\ell_\alpha^{(j)}, \qquad \delta\ell_\alpha^{(j)} \in [-\ell_b^{(j)}, \ell_b^{(j)}],$$

i.e., the nominal length $\ell_\alpha^{(0)}$ can be distorted in a cisterna by a correction $\delta\ell_\alpha^{(j)}$ but within a specified bound $\ell_b^{(j)}$ that is not subject to optimization. One can render some enzymes inactive in certain cisternae by choosing appropriate values of $\ell_\alpha^{(0)}$ and $\delta\ell_\alpha^{(j)}$. For small values for the bound ℓ_b, e.g $l_b/N_s \leq 0.2$ (here $N_s - 1$ is the number of enzymatic reactions), the decrease in $\bar{D}$ on increasing N_C is small compared to increasing N_E (see Fig. 4.3). On the other hand for large ℓ_b, e.g. $l_b/N_S \geq 0.3$, there is an approximate symmetry between N_E and N_C (see Fig. 4.3). Here we have taken the bounds to be compartment independent, i.e. $\ell_b^{(j)} = \ell_b$.

4.2.2 Optimal partitioning of enzymes in cisternae

Having studied the optimum N_E, N_C, σ to attain a given target distribution with high fidelity, we ask what is the optimal partitioning of the N_E enzymes in these N_C cisternae? Answering this within the context of our chemical reaction model (Sect. 3.3) requires some care, since it incorporates the following enzymatic features: (a) enzymes with a finite specificity σ can catalyse several reactions, although with an efficiency that varies with both the substrate index k and cisternal index j, and (b) every enzyme appears in each cisternae; however their reaction efficiencies depend on the enzyme levels, the enzymatic reaction rates and the enzyme matching function $\mathbf{L}$, all of which depend on the cisternal index j.

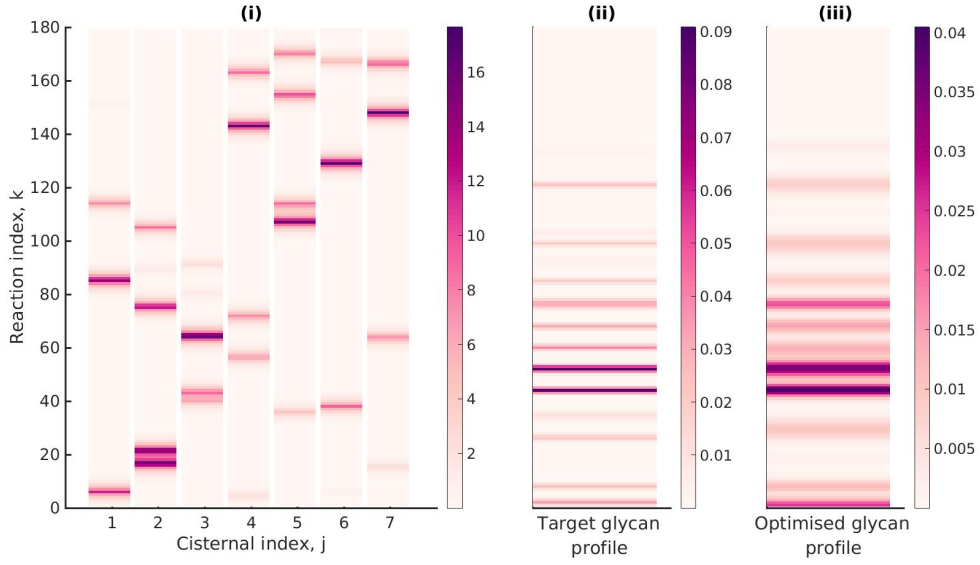
Therefore, instead of focusing on the cisternal partitioning of enzymes, we identify the chemical reactions that occur with high propensity in each cisternae. For this we use the effective reaction rate $R_{eff}(j,k)$ for $\mathcal{P}c_k \to \mathcal{P}c_{k+1}$ in the j-th cisterna which can be written as

$$R_{eff}(j,k) = \sum_{\alpha=1}^{N_E} R_\alpha^{(j)} P^{(j)}(k,\alpha). \tag{4.3}$$

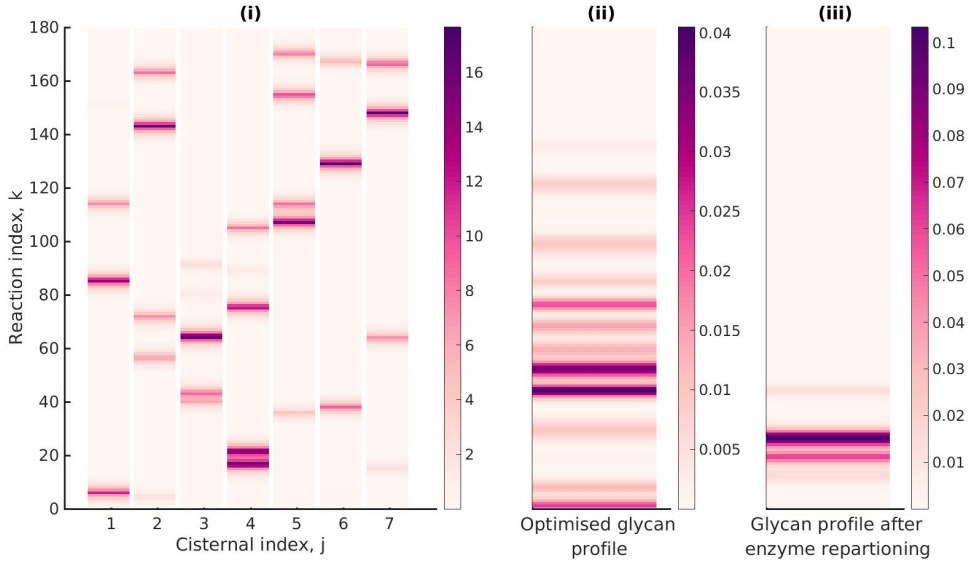
According to our model presented in Sect. 3.3, the list of reactions with high effective reaction rates in each cisterna, corresponds to a cisternal partitioning of the perfect enzymes.

Figure 4.4 (a) (i) shows the heat map of the effective reaction rates in each cisterna for the optimal N_E, N_C, σ that minimises the normalised KL-distance to the 20 GMM target distribution $\mathbf{T}^{(20)}$ (see Fig. 4.4 (a) (ii)). The optimized glycan profile displayed in Fig. 4.4 (a) (iii) is very close to the target. An interesting observation from Fig. 4.4 (a) (i) is that the same reaction can occur in multiple cisternae.

Keeping everything else fixed at the optimal value, we ask whether simply repartitioning the optimal enzymes amongst the cisternae, alters the displayed glycan distribution. In Fig. 4.4 (b) (i), we have exchanged the enzymes of the fourth and second cisterna. The glycan profile after enzyme partitioning (see Fig. 4.4 (b) ((iii)) is now completely altered (compare Fig. 4.4 (b) (ii) with Fig. 4.4 (b) (iii)). Thus, one can generate different glycan profiles by repartitioning enzymes amongst the same number of cisternae [5].



(a)



(b)

Figure 4.4. Optimal enzyme partitioning in cisternae. (a) Heat map of the effective reaction rates in each cisterna (representing the optimal enzyme partitioning) and the steady state concentration in the last compartment ($\mathbf{c}^{(N_C)}$) for the 20-GMM target distribution. Here $N_E = 5$, $N_C = 7$, normalised $D(\mathbf{T}^{(20)} \| \mathbf{c}^{(N_C)})/H(\mathbf{T}^{(20)}) = 0.11$. (b) Effective Reaction rates after swapping the optimal enzymes of the fourth and second cisternae. The displayed glycan profile is considerably altered from the original profile.

4.3 Robustness of the optimal solution

Here we analyze the change in fidelity on small perturbations in $\mathbf{R}$, $\boldsymbol{\mu}$, $\mathbf{L}$ and σ around the optimal solution. This allows us to determine where the cell needs to develop a tighter control mechanism (*stiff* directions) and where it has more leeway around the optimal values (*sloppy* directions). We do this by analyzing the eigenvalues and eigenvectors of the Hessian around the optimal point. We find that small perturbations around the optimal values in σ, change the glycan profile a lot more compared to perturbations in the other parameters and this stiffness in σ generally decreases on increasing N_E, N_C (Fig. 4.5a-4.5c). Small perturbations in $\boldsymbol{\mu}$ and some $\boldsymbol{L}$ directions around the optimum also significantly alter the glycan profile and the stiffness increases on increasing N_C, N_E, eventually becoming comparable to σ. The glycan profile is robust to perturbations in most $\boldsymbol{R}$ and some $\boldsymbol{L}$ directions (Fig. 4.5b). The total average stiffness of the optimization parameters, defined by the mean of all eigenvalues of the hessian, decreases on increasing N_E, N_C (Fig. 4.5d). We now describe the procedure in detail, starting with calculating the Hessian:

$$H(i,j) = \frac{\partial^2}{\partial X_i \partial X_j} F \Big|_{X_{\min}} \tag{4.4}$$

Here $X = [\boldsymbol{\mu}, \boldsymbol{R}, \boldsymbol{L}, \sigma]$ denotes the entire set of optimization variables (note that the enteries in X are normalized by their respective range and do not carry physical dimensions). We calculated the eigenvalues, denoted by λ_i, and eigenvectors, denoted by $\boldsymbol{V_i}$ of the Hessian matrix to identify the stiff and sloppy directions [14, 15] in the optimization space. The eigenvectors of the Hessian matrix can be grouped in $\boldsymbol{R}, \boldsymbol{L}, \boldsymbol{\mu}$ and σ directions by looking for the most dominant component in the eigenvector. We find that most of the eigenvectors have significant entries along the direction of only one of the optimization variables $\boldsymbol{\mu}, \boldsymbol{R}, \boldsymbol{L}, \sigma$, e.g.. in Fig. 4.5a, the eigenvectors $21 - 36$ have significant entries only in the $\boldsymbol{L}$ directions. There is however a small number of eigenvectors that have entries over more than one optimization direction, e.g., the eigenvector with σ dominant direction has some $\boldsymbol{\mu}$ component as well (Fig. 4.5a).

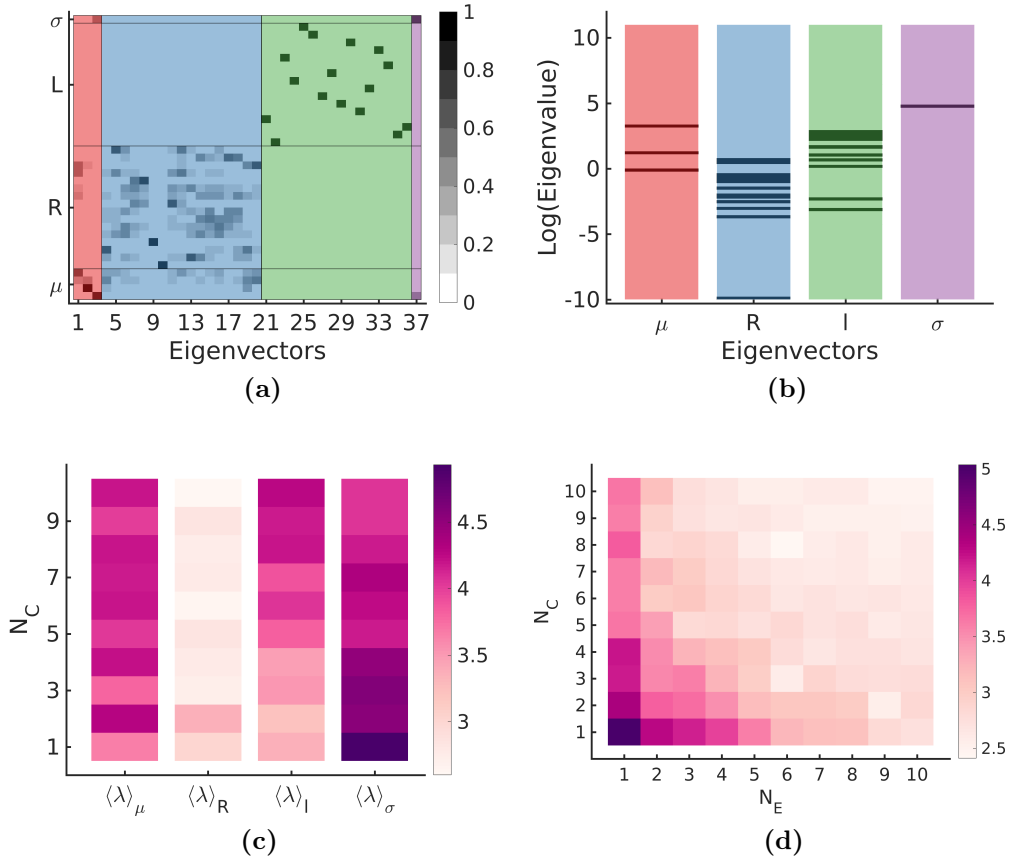


Figure 4.5. (a) Eigenvectors of the Hessian matrix $\frac{\partial^2}{\partial X_i \partial X_j} F\Big|_{X_{\min}}$ for $(N_E, N_C) = (4, 4)$. The x-axis indexes the $N_C + 2N_E N_C + 1 = 37$ eigenvectors, the y-axis indexes the $N_C + 2N_E N_C + 1$ *components* of the eigenvectors, and the grayscale denotes the absolute value of the component in the range $[0, 1]$. The componenets are grouped according to $(\boldsymbol{\mu}, \boldsymbol{R}, \boldsymbol{L}, \sigma)$ and the eigenvectors are ordered according to the most dominant component in the eigenvector($\boldsymbol{\mu}$ (orange), $\boldsymbol{R}$ (blue), $\boldsymbol{L}$ (green), σ (purple)). There is some mixing of the different components ($\mathbf{R}$ and $\boldsymbol{\mu}$ or σ and $\boldsymbol{\mu}$) but this is usually small. (b) The distribution of eigenvalues λ_i of the Hessian matrix $\frac{\partial^2}{\partial X_i \partial X_j} F\Big|_{X_{\min}}$. Each stripe represents an eigenvalue and the location of the stripe on the x-axis represents whether the dominant component of the associated eigenvector belongs to $\boldsymbol{\mu}$, $\boldsymbol{R}$, $\boldsymbol{L}$ or σ direction. (c) The average stiffness along $\boldsymbol{\mu}$, $\boldsymbol{R}$, $\boldsymbol{L}$ or σ directions, defined by the log of average of eigenvalues corresponding to the eigenvectors in the respective group, as a function of N_C for fixed $N_E = 4$. (d) Total average stiffness $\langle\lambda\rangle = \log\left(\frac{\sum \lambda_i}{N_C + 2N_E N_C + 1}\right)$ as a function of N_E, N_C.

Stiff and sloppy directions:

We find that the eigenvalues of the eigenvectors dominated by σ and some $\boldsymbol{\mu}$, $\mathbf{L}$ directions are orders of magnitude higher than for those dominated by the $\boldsymbol{R}$ directions (See Fig. 4.5b). This suggests $\bar{D}$ has a valley-like structure around the optimal, with $\boldsymbol{R}$ and some $\boldsymbol{L}$ being the flat or sloppy directions.

The fact that enzyme specificity σ and some of the $\mathbf{L}$ directions are stiff should not be surprising, since the typical width and position of peaks in the synthesized distribution is primarily controlled by σ and $\mathbf{L}$. We have already shown that $\bar{D}$ is a sharp convex function of σ for low values of (N_E, N_C) (see Fig. 3 of the paper), which gradually flattens out as we increase (N_E, N_C).

The fact that transport rate $\boldsymbol{\mu}$ is a stiff direction is surprising! The stiffness in $\boldsymbol{\mu}$ is due to the fact that the optimal $\boldsymbol{\mu}$ is always at the lower bound, and with even slight increase in $\boldsymbol{\mu}$, the transport becomes too fast for the reactions to be able to produce the intermediate products. For the $(\boldsymbol{R}, \boldsymbol{L})$-dominated eigenvectors, there are bands of sloppy direction and stiff directions. We define the average stiffness in $\boldsymbol{\mu}, \boldsymbol{R}, \boldsymbol{L}$ and σ by a weighted average of eigenvalues, where the weight is given by the strength of the corresponding components of the eigenvector.

$$\langle\lambda\rangle_\mu = \ln\left(\textstyle\sum_i w_i^{(\mu)}\lambda_i\right),\ \langle\lambda\rangle_R = \ln\left(\textstyle\sum_i w_i^{(R)}\lambda_i\right),$$
$$\langle\lambda\rangle_L = \ln\left(\textstyle\sum_i w_i^{(L)}\lambda_i\right),\ \langle\lambda\rangle_\sigma = \ln\left(\textstyle\sum_i w_i^{(\sigma)}\lambda_i\right)$$

Here $w_i^{(\mu)} = \sum_{j\in\boldsymbol{\mu}} |V_{i,j}| / \sum_j |V_{i,j}|$, $w_i^{(R)} = \sum_{j\in\boldsymbol{R}} |V_{i,j}| / \sum_j |V_{i,j}|$, $w_i^{(L)} = \sum_{j\in\boldsymbol{L}} |V_{i,j}| / \sum_j |V_{i,j}|$ and $w_i^{(\mu)} = \sum_{j\in\sigma} |V_{i,j}| / \sum_j |V_{i,j}|$.

Fig 4.5c shows $\langle\lambda\rangle_\mu$, $\langle\lambda\rangle_R$, $\langle\lambda\rangle_L$ and $\langle\lambda\rangle_\sigma$ as a function of N_C for fixed $N_E = 4$. The average stiffness in $\boldsymbol{R}$ directions, $\langle\lambda\rangle_R$, is considerably lower than the average stiffness in σ, $\boldsymbol{\mu}$ and $\boldsymbol{L}$ directions. σ is the stiffest direction but the stiffness decreases on increasing the N_C. Interestingly, the stiffness along $\boldsymbol{L}$ directions increases on increasing N_C.

We now define the total average stiffness $\langle\lambda\rangle = \log(\frac{\sum\lambda_i}{N_C+2N_EN_C+1})$, i.e. log of the sum of the eigenvalues divided by the dimension of the optimization problem, in the space of N_E, N_C . We find that the average stiffness is higher for low values of (N_E, N_C) as compared to higher values of (N_E, N_C), with a few exceptions; and

eventually, the average stiffness settles to a fixed low value (Fig. 4.5d)

4.4 Non-convexity of the optimization

The synthesis model is highly degenerate, in the sense that many combinations of parameters give rise to the same glycan profile. This makes the optimization non convex as there are many equally good minima. These degeneracies are both discrete and continuous. The continuous degeneracies correspond to regions in reaction rate ($\mathbf{R}$) -transport rate ($\boldsymbol{\mu}$) space moving along which does not change the concentration profile. The discrete degeneracies are disconnected regions in the parameter space which correspond to the same glycan profile. The number of discrete degeneracies increases exponentially with increase in (N_E, N_C). We also find that the fraction of initial conditions converging to a solution close to the global minima increases on increasing (N_E, N_C).

Validation of the numerical optimization scheme.

In order to test whether our numerical optimization procedure is able to converge to the global minimum we run the following test. We generate 100 random values of $(\boldsymbol{\mu}, \boldsymbol{R}, \boldsymbol{L}, \sigma)$ within their respective ranges for a problem instance with $(N_E = 2, N_C = 2)$. The sampled value for $(\boldsymbol{\mu}, \boldsymbol{R}, \boldsymbol{L}, \sigma)$ is used to generate concentration profiles that are then used as the target distribution for the optimization. Since the target distribution is achievable, the optimal value of the constrained *Optimization B* for these sampled targets is $\bar{D} = 0$. We solve the constrained *Optimization B* using our numerical scheme. The average optimal value $\bar{D}$ across all sampled values was 9.1835e-07, 30 out of 100 values were exactly zero, and the highest $\bar{D}$ was 1.1761e-05. Therefore, the optimization scheme was able to recover the concentration profiles almost exactly. Next, we ask whether the optimization problem recovers the value of $(\boldsymbol{\mu}, \boldsymbol{R}, \boldsymbol{L}, \sigma)$ that was used to create the particular target distribution. We were able to recover σ exactly, except in cases where the concentration profile was almost a delta function at the first glycan (see Fig. 4.6). This is because σ decides the typical width of the empirical distribution, and hence the optimal σ is determined by the typical width of the target distribution, except in the pathological case of

a concentration profile that is almost a delta function at the first glycan – such a concentration profile can be made produced for any value σ by simply making transport μ very fast as compared to the reaction rates.

We note that the optimization in $(\boldsymbol{\mu}, \boldsymbol{R}, \boldsymbol{L})$ is not convex, and leads to many equally good minimas corresponding to different values of $(\boldsymbol{\mu}, \boldsymbol{R}, \boldsymbol{L})$. The resulting redundancies in the model and their importance are discussed next.

Degeneracy in the model

Recall that in equation 4.3 , we defined

$$R_{eff}(j,k) = \frac{\sum_\alpha R_\alpha^{(j)} \exp(-\sigma|k - l_\alpha^{(j)}|)}{\mu^{(j)}}$$

In terms of these renormalised rates, the steady glycan concentration can be written as

$$\bar{c}_k^{(j)} = \frac{R_{eff}(j,k)\bar{c}_{k-1}^{(j)} + \bar{c}_k^{j-1}}{1 + R_{eff}(j,k)}, \tag{4.5}$$

i.e. the concentration is only a function of $R_{eff}(k,j)$. Thus, any combination of $(\boldsymbol{\mu}, \boldsymbol{R}, \boldsymbol{L}, \sigma)$ that maps to the same value of R_{eff} will result in the same concentration profile, and will be indistinguishable from the perspective of the objective function. Additionally, the mapping from $\mathbf{R}_{eff}$ to the concentration profile $\bar{\boldsymbol{c}}$ also has degeneracy. We show these redundancies in the schematic below, which shows a systematic reduction in dimension to 1 (scalar) which is the quantity we optimize,

$$\begin{array}{ccccccc} \mathbf{R},\ \mathbf{L},\ \mu,\ \sigma & \longrightarrow & \mathbf{R}_{eff} & \longrightarrow & \bar{\boldsymbol{c}} & \longrightarrow & F \\ N_C + 2N_E N_C + 1 & & N_C(N_S - 1) & & N_S - 1 & & 1 \end{array}$$

Since $F(\boldsymbol{c}_T || \bar{\boldsymbol{c}}) = 0$ if, and only if, $\boldsymbol{c}_T = \bar{\boldsymbol{c}}$, it follows that the last mapping does not have redundancy. Some of the sources of degeneracies in the mapping from $(\boldsymbol{\mu}, \boldsymbol{R}, \boldsymbol{L}, \sigma)$ to $\boldsymbol{R}_{eff}$ are as follows:

1. For fixed $(\sigma, \boldsymbol{L})$, setting $R_\alpha^{(j)} \leftarrow \gamma R_\alpha^{(j)}$ and $\mu_j \leftarrow \gamma^{-1}\mu_j$ leaves R_{eff} invariant.

2. Permutations in the α index leaves R_{eff} invariant. Thus, there are at least $(N_E!)^{N_C}$ distinct minima that map to the same value of R_{eff}, and therefore,

the same concentration $\bar{\boldsymbol{c}}$.

Additionally, there are degeneracies coming from the optimization which depend on the target distribution $\boldsymbol{c}_T$. Having discussed the sources of degeneracies of the optimized solution, we now discuss the distribution of the optimized solutions.

Distribution of minima:

To study the behaviour of the optimization algorithm for different initial points, we numerically investigate the distribution of function values at different local minima. Since the dimension of the optimization problem is $N_C + 2N_E N_C + 1$, which is large, we divide the optimization space into a grid of $I = n_p^{(N_C+2N_E N_C+1)}$ points. We did this numerical experiment for $(N_E, N_C) \in \{(1,1),(1,2),(2,1),(2,2)\}$. The value of $n_p = 3$ for $(N_E, N_C) = (2,2)$ and $n_p = 4$ for the rest. The target distribution for all the cases is a single Gaussian with mean 20, standard deviation 5, with support on $1 \leq k \leq 20$. The results of this numerical experiment are summarized in Fig. 4.7 and Table 4.1, from which we deduce the following:

1. A large fraction of the initial starting points converge to a set of degenerate minima with objective function value exactly equal to the global minimum. These minima are a result of the degeneracies of the optimization problem.

2. There are other local minima with objective value very close to (but not equal) to the global minimum.
   Most initial points converge to one of these two sets of minima.

3. Finally, there are a small set of local minima with significantly higher objective values. These correspond to minima with $\sigma = 0$. The fraction of initial points that converge to such minima reduces as the dimension of the optimization space increases.

4.4.1 Implications for robustness to parametric noise

Since the synthesized glycan distribution displayed by the cell marks its identity, it must be robust to noise intrinsic to the synthesis machinery. The degeneracy

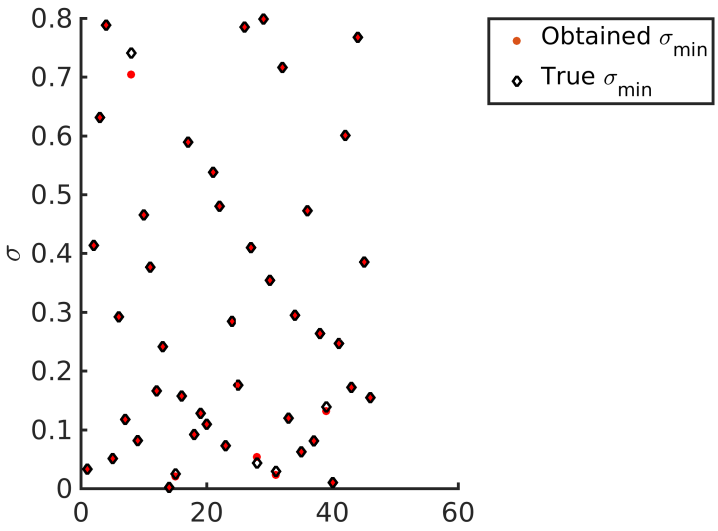
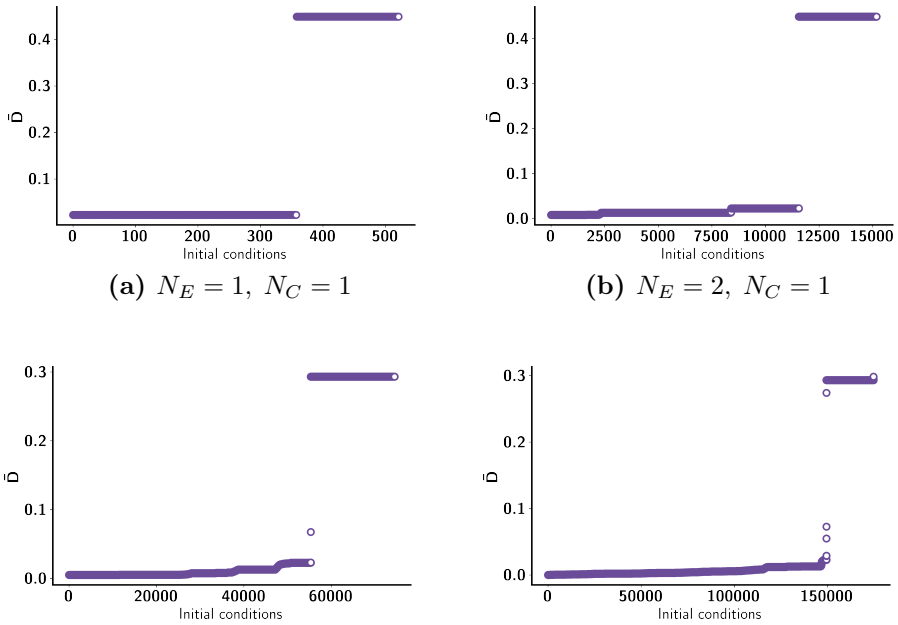


Figure 4.6. Recovering the σ values for different target distribution. Note that barring 4 data points, all other optimized σ values (red dots) exactly overlap with the corresponding target σ (diamonds).



(a) $N_E = 1,\ N_C = 1$ **(b)** $N_E = 2,\ N_C = 1$

(c) $N_E = 1,\ N_C = 2$ **(d)** $N_E = 2,\ N_C = 2$

Figure 4.7. $\bar{D}$ for various initial conditions, sorted in increasing order for clarity. This clearly shows the fraction of initial conditions for which the optimised $\bar{D}$ is small (see Table 4.1).

N_E	N_C	$\min \bar{D}_{KL}$	$\max \bar{D}_{KL}$	Fraction of initial conditions within $\bar{D}_{KL} \leq 0.0228$
1	1	0.0228	0.44	0.56
2	1	0.0081	0.44	0.73
1	2	0.0051	0.29	0.70
2	2	1.17e-4	0.29	0.84

Table 4.1. Distribution of local minima

of solutions and sloppy directions in the fidelity landscape makes the glycan distribution robust to intrinsic noise in the synthesis and cell to cell variations in the kinetic parameters. We find that the number of degeneracies increases on increasing (N_E, N_C), and the average stiffness of the optimized parameters decreases on increasing (N_E, N_C) making the synthesis more robust to parameter fluctuations. Further, while the parameter space is high dimensional, the dimension of *controllable* parameters (measured by the stiff directions) is low dimensional. We find this dimensional reduction a compelling idea which can be explored further.

4.5 Diversity

So far we have studied how the complexity of the target glycan distribution places constraints on the evolution of Golgi cisternal number and enzyme specificity. We now take up another issue, namely, how the physical properties of the Golgi cisternae, namely cisternal number and inter-cisternal transport rate, may drive diversity of glycans [16, 17]. There is substantial correlative evidence to support the idea that cell types that carry out extensive glycan processing employ larger numbers of Golgi cisternae. For example, the salivary Brunner's gland cells secrete mucous that contains heavily O-glycosylated mucin as its major component [18]. The Golgi complex in these specialized cells contain $9 - 11$ cisternae per stack. Additionally, several organisms such as plants and algae secrete a rather diverse repertoire of large, complex glycosylated proteins, for a variety of functions [19, 20, 21, 22, 23, 24, 25, 26, 27, 28]. These organisms possess enlarged Golgi complexes with multiple cisternae per stack [29, 30, 31, 32, 33].

We define *diversity* as the total number of glycan species produced above a specified threshold abundance c_{th}. This last condition is necessary because very small

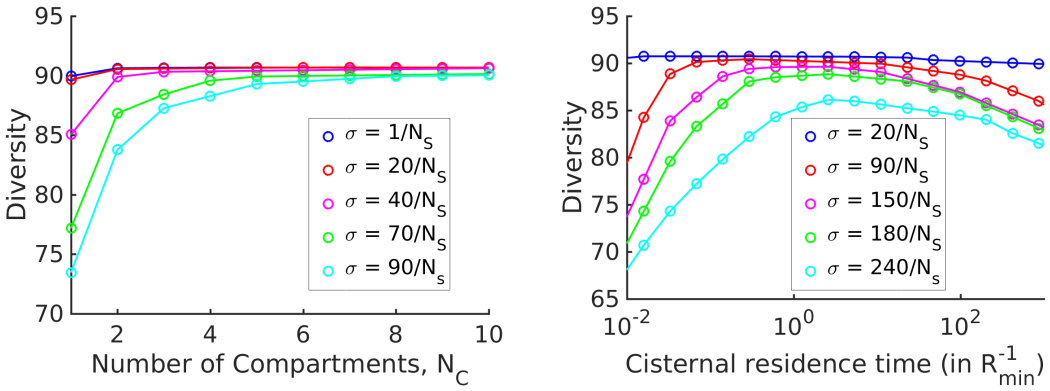


Figure 4.8. Strategies for achieving high glycan diversity. Diversity versus N_C and transport rate μ at various values of specificity σ for fixed $N_E = 3$. (a) Diversity vs. N_C at optimal transport rate μ. Diversity initially increases with N_C, but eventually levels off. The levelling off starts at a higher N_C when σ is increased. These curves are bounded by the $\sigma = 0$ curve. (b) Diversity vs. cisternal residence time (μ^{-1}) in units of the reaction time ($R_{\min}^{-1}$) at various value of σ, for fixed $N_C = 4$ and $N_E = 10$.

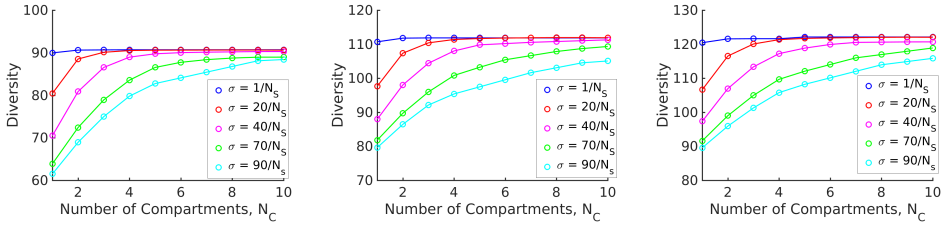
peaks will not be distinguishable in the presence of noise. In computing the diversity from our chemical synthesis model, we have chosen the threshold to be $c_{th} = 1/N_s$, where N_s is the total number of glycan species. We have checked that the qualitative results do not depend on this choice (see Fig. 4.9).

We use the sigmoid function $(1+e^{-x/\tau})^{-1}$ as a differentiable approximation to the Heaviside function $\Theta(x)$, define the following optimization to maximize diversity for a given set of parameter values, N_E, N_C, σ:

$$\begin{aligned} \text{Diversity}(\sigma, N_C, N_E) := \max_{\boldsymbol{\mu},\mathbf{R},\mathbf{L}} \quad & \textstyle\sum_{i=1}^{N_s} \left(1 + e^{-N_s(c_i - c_{th})}\right)^{-1} \\ \text{s.t.} \quad & R_{\min} \le R_\alpha^{(j)} \le R_{\max}, \\ & \mu_{\min} \le \mu^{(j)} \le \mu_{\max}, \end{aligned}$$

where, as before, $(\mu_{\max}, \mu_{\min}) = (1, 0.01)/\text{min}$, and $(R_{\max}, R_{\min}) = (20, 0.018)/\text{min}$, and $c_{th} = 1/N_s$ is the threshold. See Appendix C for details on the parameter estimation.

The results displayed in Fig. 4.8 (a), show that for a fixed specificity σ, the diversity at first increases with the number of cisternae N_C, and then saturates at a



(a) $c_{th} = \frac{1}{N_s}$ **(b)** $c_{th} = \frac{1}{2N_s}$ **(c)** $c_{th} = \frac{1}{4N_s}$

Figure 4.9. Diversity vs. N_C for different values of σ keeping $N_E = 1$ fixed, for three different values of the threshold, $c_{th} = \frac{1}{N_s}, \frac{1}{2N_s}, \frac{1}{4N_s}$. Changing the value of the threshold c_{th}, only changes the saturation value of the diversity curve.

value that depends on σ. For very high specificity enzymes, one can achieve very high diversity by appropriately increasing N_C. This establishes the link between glycan diversity and cisternal number. However, this link is correlational at best, since there are many ways to achieve high glycan diversity – notably by increasing the number of enzymes.

On the other hand, one of the goals of glycoengineering is to produce a particular glycan profile with low heterogeneity [5, 13]. For low specificity enzymes, the diversity remains unchanged upon increasing the cisternal residence time. For enzymes with high specificity, the diversity typically shows a non-monotonic variation with the cisternal residence time. At small cisternal residence time, the diversity decreases from the peak because of early exit of incomplete oligomers. At large cisternal residence time the diversity again decreases as more reactions are taken to completion. Note that the peak is generally very flat, this is consistent with the results in [13]. To get a sharper peak, as advocated for instance by [5], one might need to increase the number of high specificity enzymes N_E further.

4.6 Conclusions

In summary,

- We say that every cell type is associated with a “target" glycan distribution which is complex for complex multicellular organisms.
- The glycan synthesis machinery tries to achieve a glycan profile close to the

"target" glycan profile for that cell type. This puts constraints on the glycan synthesis machinery.

- We find that increasing number of glycosylation enzyme and number of Golgi compartments increases the fidelity of synthesis for a complex "target" profile. Increasing the number of enzymes requires an elaborate gentic cost; increasing the number of Golgi compartments maybe cheaper for the cell
- There is an optimal enzyme specificity for achieving a particular "target", which says that glycosylation enzymes should show a degree of substrate promiscuity.
- The syntheis model is degenerate and has parameters which do not need tight control. This makes the glycan distribution robust to variations in the parameters in cells of the same cell type.

A major implication of this work is that the control of Golgi cisternal number must involve a *coupling* between the non-equilibrium self assembly of Golgi cisternae with the enzyme reactions kinetics happening inside the cisternae [**? ?**].

Bibliography

[1] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[2] David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[3] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[4] Yonatan Savir and Tsvi Tlusty. Conformational proofreading: the impact of conformational changes on the specificity of molecular recognition. *PloS one*, 2(5):e468, 2007.

[5] Anjali Jaiman and Mukund Thattai. Algorithmic biosynthesis of eukaryotic glycans. *BioRxiv*, page 440792, 2018.

[6] Richard D. Cummings and Paul Crocker. Functional Glycomics Database, Consortium for Functional Glycomics. `http://www.functionalglycomics.org`, 2020.

[7] Peicheng Du, Gustavo Stolovitzky, Peter Horvatovich, Rainer Bischoff, Jihyeon Lim, and Frank Suits. A noise model for mass spectrometry based proteomics. *Bioinformatics*, 24(8):1070–1077, 2008.

[8] Pablo Umaña and James E Bailey. A mathematical model of n-linked glycoform biosynthesis. *Biotechnology and bioengineering*, 55(6):890–908, 1997.

[9] Frederick J Krambeck, Sandra V Bennun, Someet Narang, Sean Choi, Kevin J Yarema, and Michael J Betenbaugh. A mathematical model to derive n-glycan structures and cellular enzyme activities from mass spectrometric data. *Glycobiology*, 19(11):1163–1175, 2009.

[10] Frederick J Krambeck and Michael J Betenbaugh. A mathematical model of n-linked glycosylation. *Biotechnology and Bioengineering*, 92(6):711–728, 2005.

[11] Peter Fisher and Daniel Ungar. Bridging the gap between glycosylation and vesicle traffic. *Frontiers in cell and developmental biology*, 4:15, 2016.

[12] Debrup Sengupta and Adam D Linstedt. Control of organelle size: the golgi complex. *Annual review of cell and developmental biology*, 27:57–77, 2011.

[13] Peter Fisher, Hannah Spencer, Jane Thomas-Oates, A Jamie Wood, and Daniel Ungar. Modeling glycan processing reveals golgi-enzyme homeostasis upon trafficking defects and cellular differentiation. *Cell reports*, 27(4):1231–1243, 2019.

[14] Ryan N Gutenkunst, Joshua J Waterfall, Fergal P Casey, Kevin S Brown, Christopher R Myers, and James P Sethna. Universally sloppy parameter sensitivities in systems biology models. *PLoS computational biology*, 3(10):e189, 2007.

[15] Benjamin B Machta, Ricky Chachra, Mark K Transtrum, and James P Sethna. Parameter space compression underlies emergent theories and predictive models. *Science*, 342(6158):604–607, 2013.

[16] Ajit Varki. Evolutionary forces shaping the golgi glycosylation machinery: why cell surface glycans are universal to living cells. *Cold Spring Harbor perspectives in biology*, 3(6):a005462, 2011.

[17] James W Dennis, Ivan R Nabi, and Michael Demetriou. Metabolism, cell surface organization, and disease. *Cell*, 139(7):1229–1241, 2009.

[18] Herman van Halbeek, Gerrit J Gerwig, Johannes FG Vliegenthart, Henk L Smits, Peter JM Van Kerkhof, and Mebius F Kramer. Terminal α (1$\rightarrow$ 4)-linked n-acetylglucosamine: A characteristic constituent of doudenal-gland mucous glycoproteins in rat and pig: A high-resolution 1h-nmr study. *Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology*, 747(1-2):107–116, 1983.

[19] Heather E McFarlane, Anett Döring, and Staffan Persson. The cell biology of cellulose synthesis. *Annual review of plant biology*, 65:69–94, 2014.

[20] Bjørn EV Koch, Jens Stougaard, and Herman P Spaink. Keeping track of the growing number of biological functions of chitin and its interaction partners in biomedical research. *Glycobiology*, 25(5):469–482, 2015.

[21] Malcolm A O'Neill, Tadashi Ishii, Peter Albersheim, and Alan G Darvill. Rhamnogalacturonan ii: structure and function of a borate cross-linked cell wall pectic polysaccharide. *Annu. Rev. Plant Biol.*, 55:109–139, 2004.

[22] Takahisa Hayashi and Rumi Kaida. Functions of xyloglucan in plant cells. *Molecular Plant*, 4(1):17–24, 2011.

[23] Pardeep Kumar, Meng Yang, Brian C Haynes, Michael L Skowyra, and Tamara L Doering. Emerging themes in cryptococcal capsule synthesis. *Current opinion in structural biology*, 21(5):597–602, 2011.

[24] Neil AR Gow and Bernhard Hube. Importance of the candida albicans cell wall during commensalism and infection. *Current opinion in microbiology*, 15(4):406–412, 2012.

[25] Melani A Atmodjo, Zhangying Hao, and Debra Mohnen. Evolving views of pectin biosynthesis. *Annual review of plant biology*, 64, 2013.

[26] Stephen J Free. Fungal cell wall organization and biosynthesis. In *Advances in genetics*, volume 81, pages 33–82. Elsevier, 2013.

[27] Markus Pauly, Sascha Gille, Lifeng Liu, Nasim Mansoori, Amancio de Souza, Alex Schultink, and Guangyan Xiong. Hemicellulose biosynthesis. *Planta*, 238(4):627–642, 2013.

[28] Rachel A Burton and Geoffrey B Fincher. Evolution and development of cell walls in cereal grains. *Frontiers in plant science*, 5:456, 2014.

[29] Burkhard Becker and Michael Melkonian. The secretory pathway of protists: spatial and functional organization and evolution. *Microbiol. Mol. Biol. Rev.*, 60(4):697–721, 1996.

[30] Alexander A Mironov, Irina S Sesorova, Elena V Seliverstova, and Galina V Beznoussenko. Different golgi ultrastructure across species and tissues: Implications under functional and pathological conditions, and an attempt at classification. *Tissue and Cell*, 49(2):186–201, 2017.

[31] Bryon S Donohoe, Byung-Ho Kang, and L Andrew Staehelin. Identification and characterization of copia-and copib-type vesicle classes associated with plant and algal golgi. *Proceedings of the National Academy of Sciences*, 104(1):163–168, 2007.

[32] Soren Mogelsvang, Natalia Gomez-Ospina, Jon Soderholm, Benjamin S Glick, and L Andrew Staehelin. Tomographic evidence for continuous turnover of golgi cisternae in pichia pastoris. *Molecular biology of the cell*, 14(6):2277–2291, 2003.

[33] Mark S Ladinsky, Christine C Wu, Shane McIntosh, J Richard McIntosh, and Kathryn E Howell. Structure of the golgi and distribution of reporter molecules at 20 c reveals the complexity of the exit compartments. *Molecular biology of the cell*, 13(8):2810–2825, 2002.

Chapter 5

Thermodynamic control of the enzyme specificity

In the previous chapter, we have shown that enzyme specificity is an important parameter that controls the glycan distribution of the cell. We had taken the enzyme specificity to be a scalar that is independent of the enzyme or the compartment for the previous calculation. Indeed, the enzyme specificity towards substrates of fixed shapes in a induced fit model of enzyme substrate interaction is a coarse grained property dependent on the molecular structure and folding of the enzyme [1, 2]. Here, we study one possible mechanism of putting a thermodynamic control on the enzyme specificity which can make the enzyme specificity dependent on the compartment and the consequent energy-accuracy trade-offs. In a broader biological context, the ability of mimicking changes in a molecular property, which is genetically controlled and changes over evolutionary timescales, by processes which can cause changes over a much shorter timescales may be evolutionarily important to the organism by making them more adaptable to sudden changes in the environment [3].

We start with showing that compartment dependent enzyme specificity increases the fidelity of synthesis for the same number of enzymes, number of compartments. We generalize the induced fit enzyme-substrate interaction model that we used in Chapter 3 to include an external non-equilibrium drive. We then show that this generic non-equilibrium drive can increase the effective enzyme promiscuity. Finally, we calculate the energetic cost of implementing this drive by calculating the entropy production by the system.

5.1 Enzyme and compartment dependent enzyme specificity

In Chapter 4, we had used the following distortion relation to model the induced fit binding of an enzyme α with a substrate k in a compartment j

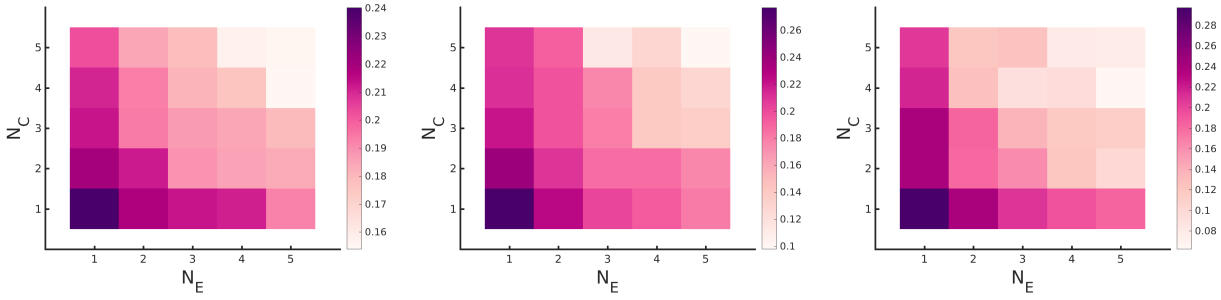
$$P^{(j)}(\alpha, k) = \exp(-\sigma|l_\alpha^{(j)} - l_k|) \tag{5.1}$$

where σ represents the enzyme specificity, $l_\alpha^{(j)}$ denotes the ideal enzyme length and l_k denotes the substrate length. σ was assumed to be a scalar quantity independent of the enzyme or the compartment. In this section, we relax this assumption by making enzyme specificity dependent on both the enzyme and the compartment. In our enzyme model the enzyme specificity was related to the enzyme elasticity which is a coarse grained molecular property of the enzyme, and therefore it should be dependent on the enzyme. The compartment dependence of enzyme specificity requires some external mechanism that can change the specificity. We will explore the detailed mechanisms in the following sections. Here we show that making enzyme specificity enzyme and compartment dependent results in better (lower) fidelity of synthesis for the same number of enzymes and compartments. The distortion relation for enzyme and compartment dependent enzyme specificity can be written as

$$P^{(j)}(\alpha, k) = \exp(\sigma_\alpha^{(j)}|l_\alpha^{(j)} - l_k|) \tag{5.2}$$

The steady state glycan concentrations are still given by (3.13)-(3.14) of Chapter 3 with $P^{(j)}(\alpha, k)$ now given by (5.2). The optimization of fidelity of synthesis can now be written as follows

$$\begin{aligned} \textit{Optimization}: \quad \bar{D}(N_E, N_C, \boldsymbol{c^*}) \quad &:= \quad \min_{\boldsymbol{\sigma},\, \boldsymbol{\mu},\, \boldsymbol{R},\, \boldsymbol{L}} F(\boldsymbol{c}^*\|\bar{\boldsymbol{c}}) \\ &\text{s.t.} \quad \sigma_\alpha^{(j)} \geq 0 \\ &\quad\;\; \mu_{min} \leq \mu^{(j)} \leq \mu_{max} \\ &\quad\;\; R_{min} \leq R(j,\alpha) \leq R_{max} \\ &\quad\;\; 1 \leq l_\alpha^{(j)} \leq N_s \end{aligned} \tag{5.3}$$



(a) **(b)** **(c)**

Figure 5.1. Synthesis error (fidelity) in the space of N_E, N_C in the case of (a) enzyme and compartment independent enzyme specificity, (b) enzyme dependent but compartment independent enzyme specificity, (c) enzyme and compartment dependent enzyme specificity. The system in which enzyme specificity depends on the enzyme or/and compartment does better than the system in which enzyme specificity in independent of enzymes and compartments.

Here, $F(\boldsymbol{c^*}\|\bar{\boldsymbol{c}}) = D_{KL}(c^*\|\bar{c})/H(\boldsymbol{c^*})$, measures the fidelity of synthesis same as (4.1) in Chapter 4. The bound of rates are taken from literature detailed in Appendix C. The optimization is performed by same algorithm as detailed in Chapter 4. The results of this optimization (see Figure 5.1 and 5.2) show that better fidelity can be achieved for the same value of N_E, N_C if enzyme specificity has a compartment or/and enzyme dependence. Compartment dependent enzyme specificity further increases the power of compartments making an increase in the number of compartments more effective way of achieving better synthesis fidelity (See Figure 5.2). Another interesting find of this optimization is that, in the case of enzyme and compartment dependent specificity, the optimal partitioning of enzymes, in the sense of R_{eff} defined in (4.3), is the one in which the enzymes in the first compartment are much more promiscuous than the enzymes in the rest of the compartments (See Figure 5.3).

These results provide one incentive for the cell to control enzyme specificity by some mechanism separately in each of the compartments. In the rest of this chapter, we propose one such mechanism involving a mechanical non-equilibrium drive to change the enzyme specificity and calculate the thermodynamic cost of setting it up.

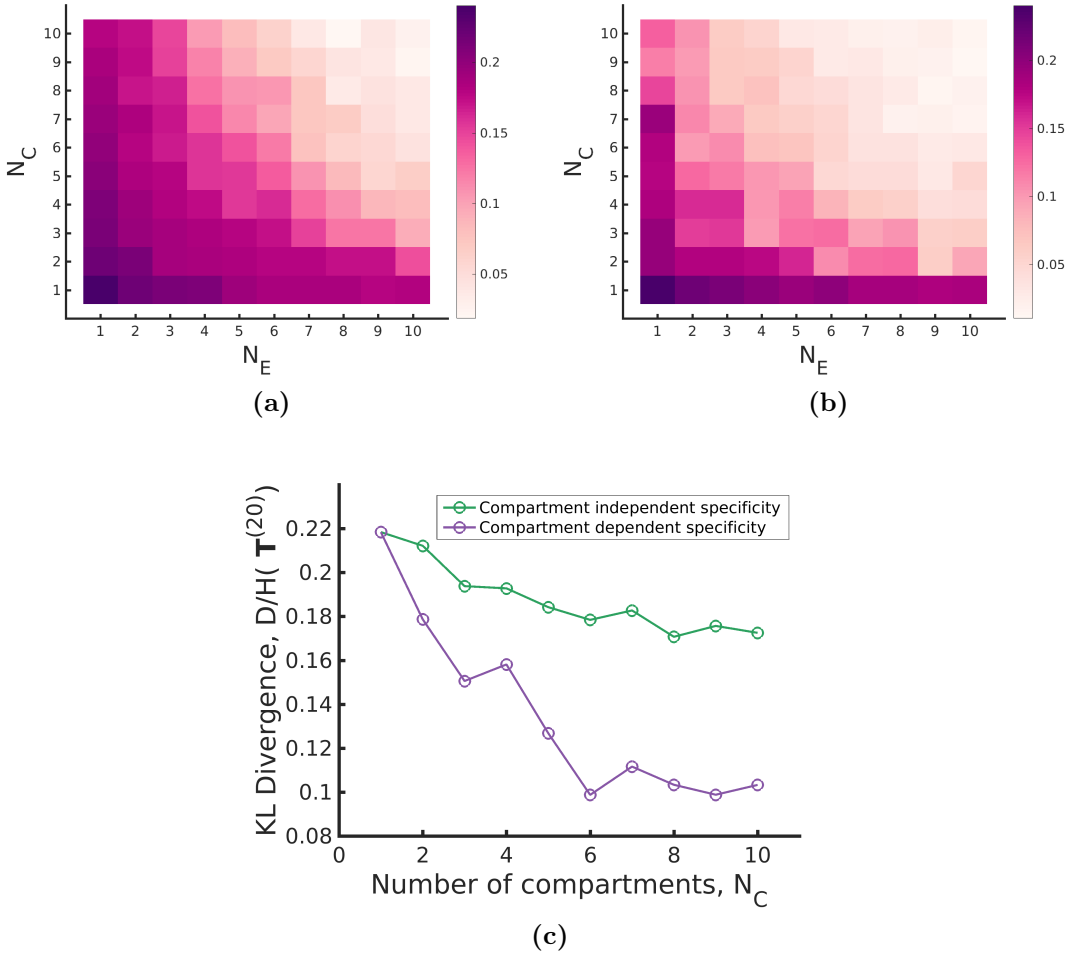
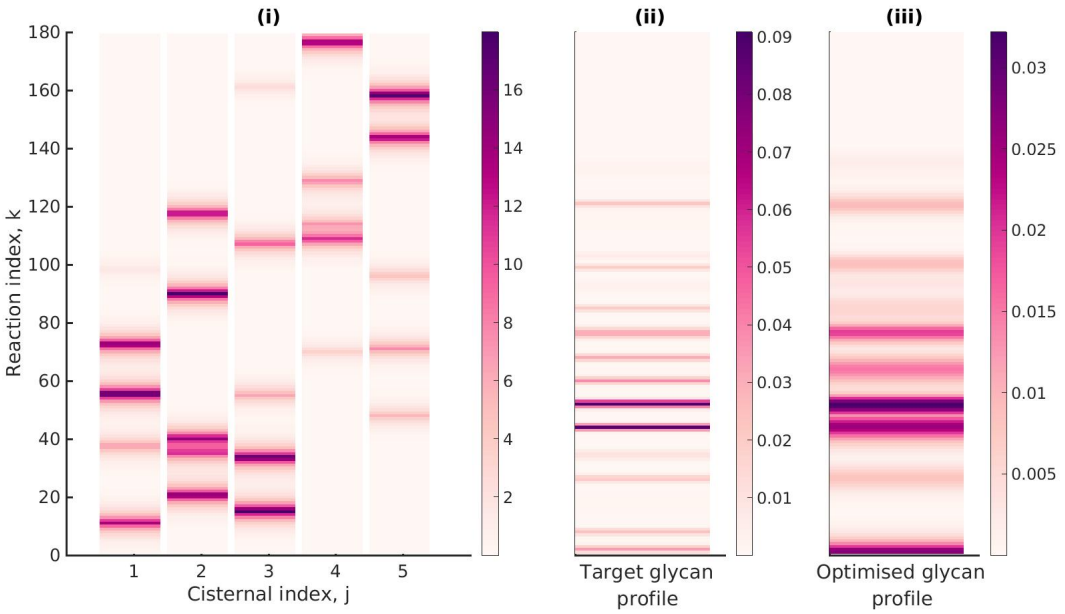
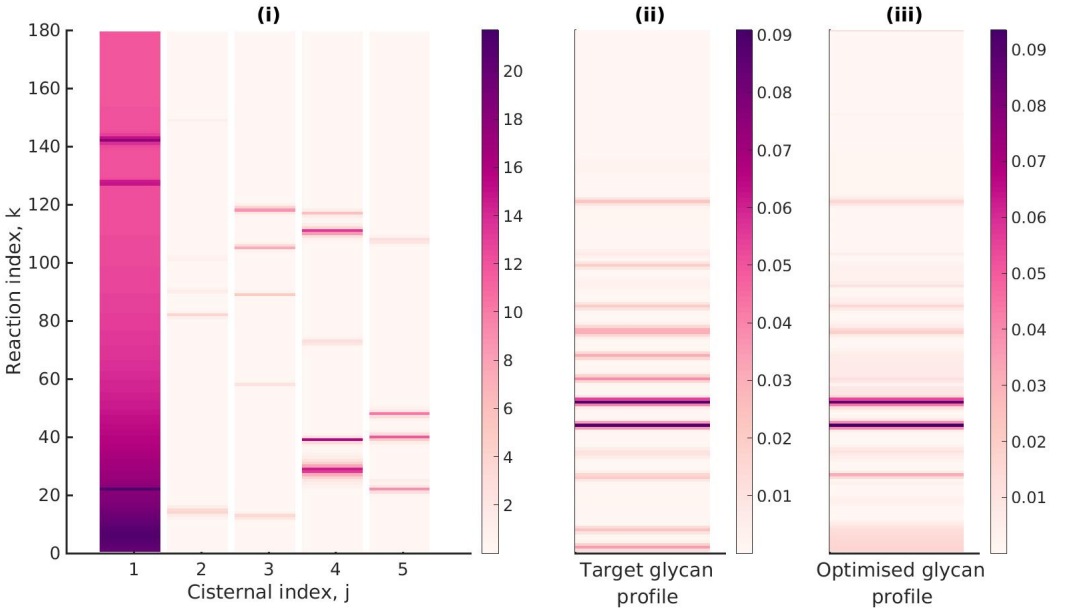


Figure 5.2. Effect of compartment dependent enzyme specificity on the fidelity of synthesis. (a) Fidelity in the space of N_E, N_C in the case of compartment independent enzyme specificity (b) Fidelity in the space of N_E, N_C in the case of compartment dependent enzyme specificity (c) Fidelity as a function of N_C for fixed $N_E = 2$. The figures show that the system with compartment dependent enzyme specificity achieves better(lower) synthesis fidelity.



(a)



(b)

Figure 5.3. $R_{eff}(j,k)$ in the space of compartment, j, and reaction index, k, for a system of $N_E = 5, N_C = 5$ in case of (a) enzyme and compartment independent specificity, (b) enzyme and compartment dependent specificity. Note that the enzymes in the first compartment are a lot more promiscuous and the optimized glycan profile is closer to the target profile in the case of enzyme and compartment dependent enzyme specificity.

5.2 An elastic model for induced fit enzyme sunstrate binding

We model the enzyme substrate binding by induced fit, similar to [4, 5], of a flexible enzyme and fixed substrate, like in the previous chapters, but now in presence of a thermal bath and a non equilibrium drive. The model is schematically displayed in Figure 5.4. The shape of the enzyme in the model is a dynamical quantity which is affected by (i) elastic deformation of the enzyme [6] modeled by an over-damped spring, (ii) thermal noise of the bath and, (iii) the non equilibrium external mechanical drive. In this model, binding between the enzyme and substrate occurs when the shape of the enzyme is close to the shape of the substrate for an extended period of time, τ_b (See Figure 5.4). We assume that the binding energy and typical binding time, τ_b, of all the substrates is the same and substrates are discriminated only on the basis of mismatch in shape between the enzyme and the substrate. This is not a very strong restriction because (i) it has been observed that the binding energy affects only the unbinding rate of the enzyme [7], and (ii) in our model of glycosylation, glycans are polymers of the same monomers, therefore discrimination between different glycans is likely to be because of different shapes rather than chemical binding energy.

In the interest of clarity, we describe the model for enzyme substrate binding for a one dimensional enzyme in a thermal bath with N_S different one dimensional substrates of length $l_1, \ldots l_{N_S}$. The model can easily be extended to higher dimensions. The dynamics of the enzyme length, l, is described by the following equation

$$\frac{dl(l)}{dt} = -\frac{K}{\gamma}(l(t) - l_0) + \frac{f_d(t)}{\gamma} + \frac{D}{\gamma}\eta(t) \tag{5.4}$$

where l_0 represents the resting length of the enzyme, K represents the elastic constant of the enzyme, γ is the dissipation constant, f_d is the external mechanical drive, D is the diffusion constant and $\eta(t)$ is Gaussian white noise with $\langle \eta(t) \rangle = 0$ and $\langle \eta(t)\eta(t') \rangle = \delta_{t,t'}$. Here we have assumed the enzyme to be over-damped, the first term is the elastic force, the second term is the force due to an external driving protocol and the third term is the noisy force arising due to the random collisions of

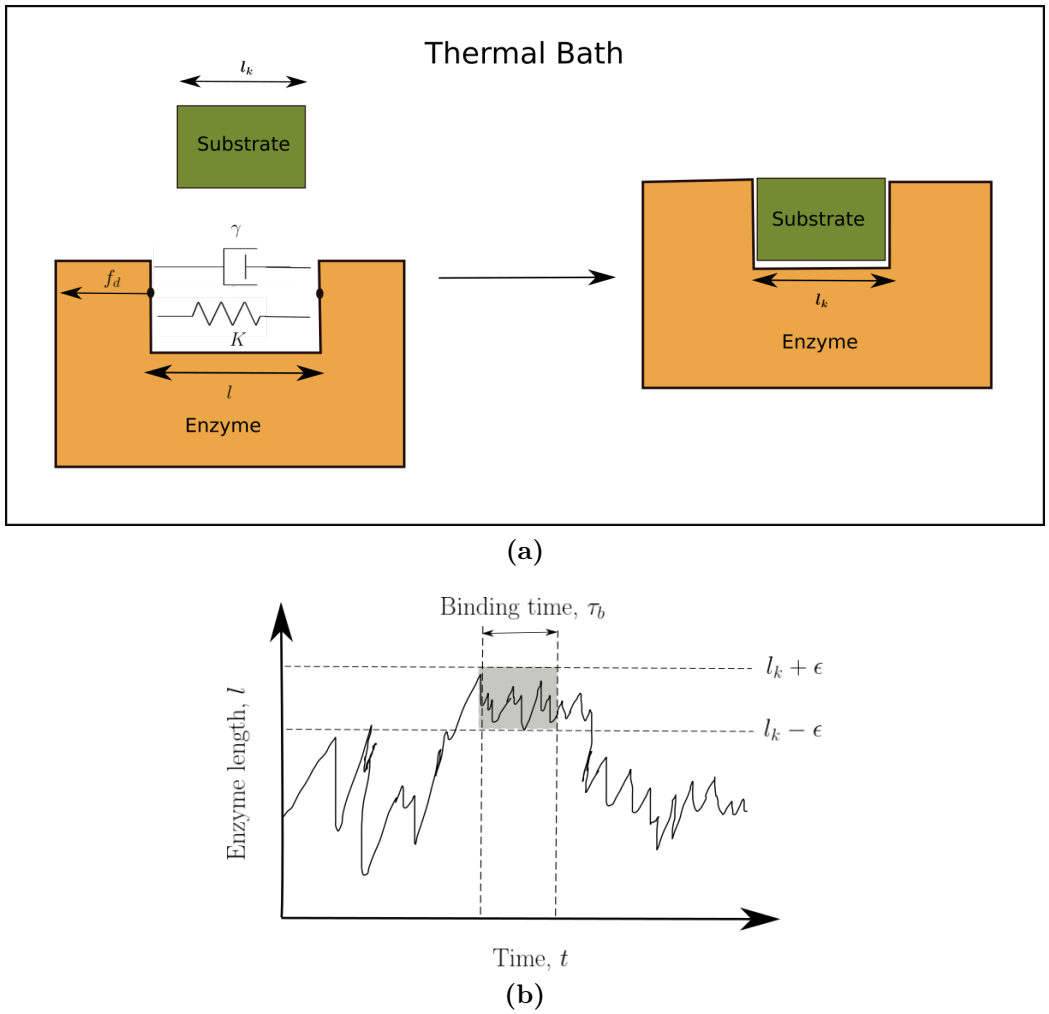


Figure 5.4. (a) The enzyme substrate binding model based on induced fit of a flexible enzyme and a fixed substrate in a thermal bath. The shape (length in 1D) of the binding site of an enzyme, denoted by l, is assumed to be deformable. The dynamics of l is given by an elastic part, K, a damping part, γ, noise from the thermal bath, and the external oscillatory mechanical drive, f_d. The shape of substrate k is denoted by l_k and the resting shape of enzyme α in compartment j is given by l_j^α. (b) Binding of substrates to enzyme occurs when the shape of the enzyme is close to the shape of the substrate, $l_k - \epsilon \leq l(t') \leq l_k + \epsilon$, for an extended period of time, $t \leq t' \leq t + \tau_b$. The shaded region in the figure represents a binding event.

particles in the thermal bath. A similar equation in higher dimensions can be written for the "shape" of the elastic enzyme. The equivalent Fokker-Planck equation describing the probabilities of length of the enzyme, $l(t)$, can be written as follows.

$$\frac{\partial p(l,t)}{\partial t} = -\frac{\partial}{\partial l}J(l,t) = -\frac{\partial}{\partial l}\left[(-\frac{K}{\gamma}(l-l_0) + \frac{f_d(t)}{\gamma})p(l,t) - \frac{D}{\gamma}\frac{\partial}{\partial l}p(l,t)\right] \tag{5.5}$$

where J is the probability current.

Given this description of the elastic enzyme in a thermal bath, we now describe one way of formalizing the notion of probability of binding between the enzyme and the substrate as described in the Figure 5.4. A binding event between the enzyme and a substrate k occurs at time t if the length of the enzyme, $l(t)$, stays around the substrate length, l_k, for an extended period of time, τ_b; i.e. $l_k - \epsilon \leq l(t') \leq l_k + \epsilon$ for $t - \tau_b \leq t' \leq t$. The probability of this happening can be calculated by the following

$$P(k,t) = \prod_{t_i=t-\tau_b}^{t} \int_{l_k-\epsilon}^{l_k+\epsilon} p(l,t_i)dl \tag{5.6}$$

where we have discretized the time window, $(t-\tau_b, t)$ into N_d points separated by $\Delta t = \tau_b/N_d$. We now do the calculation in the limit $\tau_b = 0$, we will subsequently relax this assumption and see the effect of a finite non-zero τ_b. In this case the steady state binding probability can then be written as:

$$P(k,t) = \int_{l_k-\epsilon}^{l_k+\epsilon} p(l,t)dl \tag{5.7}$$

We now show that in equilibrium (absence of external drive, $f_d(t) = 0$), the current description of binding probabilities is equivalent to our earlier model for the binding probabilities in (5.1). We solve the following equation to obtain the equilibrium ($J = 0$) solution of (5.5) in absence of external force

$$\frac{K}{\gamma}(l-l_0)p(l,t) + \frac{D}{\gamma}\frac{\partial}{\partial l}p(l,t) = 0$$

in one dimension in the region $l \in (0,\infty)$ with boundary conditions $p(0,t) =$

$p(\infty, t) = 0$. The steady state equilibrium ($J = 0$) solution is given by

$$p^{ss}(l) = \frac{1}{\mathcal{N}} \exp(-\frac{K}{2D}(l - l_0)^2)$$

The binding probabilities from (5.7) can be written in terms of this steady state probability leading to the following expression for binding probabilities.

$$P(k) = \frac{1}{\mathcal{N}} \exp\left(-\frac{K}{2D}(l_k - l_0)^2\right) \tag{5.8}$$

which is similar to the distortion function (5.1) used in Chapter 3 but with the distance between the two lengths defined by the $\mathcal{L}_2$ norm rather than the $\mathcal{L}_1$ norm. As described earlier in Chapter 3, this change of distance metric from $\mathcal{L}_1$ norm to $\mathcal{L}_2$ norm does not affect the qualitative results of Chapter 3 and 4. The ratio, $K/2D$, in the current description plays the same role as enzyme specificity, σ, of the earlier description. The enzyme elastic constant, K, is a molecular property of the enzyme [1] and the the diffusion constant, D, is related to the physiological temperature of the bath in the compartment. In the next section we show that the enzyme specificity can be lowered by a generic periodic non-equilibrium drive. This provide a mechanism for mimicking promiscuous enzymes without the requiring a change in the molecular property of the enzymes, and making the enzyme specificity compartment dependent.

5.3 Effect of an oscillatory mechanical drive on enzyme specificity

Here we analyze the effect of an external oscillatory mechanical drive that changes the resting length of the enzyme on the specificity of the enzyme. The origin of this oscillatory drive could be due to changes in the membrane properties of the Golgi cisternae, since glycosylation enzymes are embedded in the membrane of the Golgi cisternae [8]. Indeed, it has been shown that membrane structure and composition affect the activities of membrane proteins [9], e.g. lipid bilayer of different thickness deforms membrane proteins differently to ensure good hydrophobic matching to the surrounding lipid bilayer [9]. In this scenario the dynamics of enzyme length (or

"shape"), l, can be described by the following Langevin equation:

$$\frac{dl(t)}{dt} = -\frac{K}{\gamma}(l(t) - l_0(1 + r\cos(\omega t))) + \frac{D}{\gamma}\eta(t) \tag{5.9}$$

Here, the external drive is given by $f_d(t) = -\frac{K}{\gamma} l_0 r \cos(\omega t)$ where $r \in [0, 1)$ is the strength of the drive, and ω is the frequency. The corresponding Fokker-Planck equation can be written as

$$\frac{\partial p(l,t)}{\partial t} = -\frac{\partial}{\partial l}\left[-\frac{K}{\gamma}\left\{l - l_0(1 + r\cos(\omega t))\right\} p(l,t) - \frac{D}{\gamma}\frac{\partial}{\partial l} p(l,t)\right] \tag{5.10}$$

We solve this equation with the following initial condition

$$p(l, 0) = \delta(l - l_0) \tag{5.11}$$

in the region $l > 0$ with periodic boundary conditions. We make the following variable changes

$$x = l - l_0 \quad \mu = \frac{K}{\gamma} \quad A = -\frac{K}{\gamma} l_0 r \quad D = \frac{D}{\gamma} \tag{5.12}$$

The equation in terms of the new variables can be written as

$$\frac{\partial p(x,t)}{\partial t} = -\frac{\partial}{\partial x}\left[-(\mu x + A\cos(\omega t))p(x,t) - D\frac{\partial}{\partial x} p(x,t)\right] \tag{5.13}$$

with initial condition

$$p(x, 0) = \delta(x) \tag{5.14}$$

In order to solve this equation analytically, we assume $l_0 >> 1L$ where L is a unit of length to measure the enzyme length. This assumption makes the region of integration, $x \in (-\infty, \infty)$ and the boundary conditions $p(-\infty, t) = p(\infty, t) = 0$. We first take the Fourier transform of (5.11) in space and then solve the resulting first order Fourier partial differential equation (pde) using method of characteristics. We define the fourier transform as:

$$\tilde{p}(k,t) = \int_{-\infty}^{\infty} p(x,t)\exp(ikl)dx$$

The Fourier transform of (5.13) is given by

$$\frac{\partial \tilde{p}(k,t)}{\partial t} = -k\mu \frac{\partial \tilde{p}(k,t)}{\partial k} - (ikA\cos(\omega t) + k^2 D)\tilde{p}(k,t)$$

This is a first order linear pde that can be solved by the method of characteristic. The characteristic equations for this pde can be obtained from the following Lagrange-Charpit equations

$$\frac{dt}{1} = \frac{dk}{\mu k} = -\frac{d\tilde{p}}{ikA\cos(\omega t + k^2 D)\tilde{p}(k,t)}$$

resulting in the following two odes

$$\frac{dk}{dt} = \mu k \implies k = C_1 \exp(\mu t) \tag{5.15}$$

$$\frac{d\tilde{p}}{dk} = -\frac{iA\cos(\omega t) + kD}{\mu}\tilde{p} \implies \tilde{p} = C_2 \exp(-\frac{iAk\cos(\omega t)}{\mu} - \frac{k^2 D}{2\mu}) \tag{5.16}$$

where C_1 and C_2 are integration constants for the characteristic curve and are related to each other by an unknown function f

$$C_2 = f(C_1) = f(k\exp(-\mu t)) \tag{5.17}$$

We determine the function f using the initial condition $\tilde{p}(k,0) = 1$ obtained by taking the fourier transform of the initial condition in (5.11). The resulting f is given by

$$f(x) = \exp(\frac{iAx}{\mu} + \frac{x^2 D}{2\mu}) \tag{5.18}$$

$\tilde{p}(k,t)$ can now be obtained by putting the value of C_2 from equations (5.17) and (5.18) in (5.16).

$$\tilde{p}(k,t) = \exp\left[-\frac{iAk}{\mu}(\cos(\omega t) - \exp(-\mu t)) - \frac{k^2 D}{2\mu}(1 - \exp(-2\mu t))\right]$$

Finally, we take the inverse Fourier transform of $\tilde{p}(k,t)$ to obatin $p(x,t)$

$$p(x,t) = \sqrt{\frac{\mu}{2\pi D(1-\exp(-\mu t))}}\exp\left[-\frac{\mu(x + \frac{A}{\mu}(\cos\omega t - \exp(-\mu t)))^2}{2D(1-\exp(-2\mu t))}\right]$$

We can now transform this equation back to the original variables to obtain the probability density of enzyme length, l

$$p(l,t) = \sqrt{\frac{\mu}{2\pi D(1-\exp(-\mu t))}} \exp\left[-\frac{\mu(l - l_0(1 + r\cos\omega t) - \exp(-\mu t)))^2}{2D(1-\exp(-2\mu t))}\right] \tag{5.19}$$

Figure 5.5 shows the probability density, $p(l,t)$, in the space of the enzyme length, l, and time, t, for two cases: Figure 5.5(a) shows $p(l,t)$ in the case of no external drive, $r = 0$ and Figure 5.5(b) shows $p(l,t)$ in the case of an oscillatory external drive with amplitude $r = 0.1$. Width of the distribution in both the cases is decided by the ratio of elastic constant to the diffusion constant $K/2D$ which we have taken to be $0.25L^2$, where L is the unit of the enzyme length. The amplitude of oscillation is decided by the drive amplitude, r, and the frequemcy is decided by the drive frequency, ω. At long time, $t >> 1/\mu$, the probability density becomes time periodic, $p^{ss}(l,t) \approx p^{ss}(l, t + 2\pi/\omega)$ and can be written as

$$p^{ss}(l,t) = \sqrt{\frac{K}{2\pi D}} \exp\left[-\frac{K}{2D}(l - l_0(1 - r\cos\omega t))^2\right]$$

The time averaged probability over a time period ($T = 2\pi/\omega$) is defined as

$$\bar{p}^{ss}(l) = \frac{1}{T}\int_0^T p^{ss}(l,t)dt = \frac{1}{T}\sqrt{\frac{K}{2\pi D}}\int_0^T \exp\left[-\frac{K}{2D}(l - l_0(1 - r\cos\omega t))^2\right] dt$$

We can further write down the enzyme substrate binding probability from (5.7) using $\bar{p}^{ss}(l)$ as

$$P(k) = \frac{1}{\mathcal{N}}\int_0^T \exp\left[-\frac{K}{2D}(l_k - l_0(1 - r\cos\omega t))^2\right] dt \tag{5.20}$$

The integral in the equation for $\bar{p}^{ss}(l)$ and $P(k)$ is a gaussian in l and l_k respectively with mean l_0 and a width that depends on r. We plot this in Figure 5.6 for different values of the drive amplitude, r (See Figure 5.2(a)), and for different values of the drive frequency, ω (See Figure 5.2(b)). The width of the distribution increases on increasing r making the enzyme becomes more promiscuous. In contrast, changing the drive frequency has no effect on the enzyme promiscuity for the case of instant

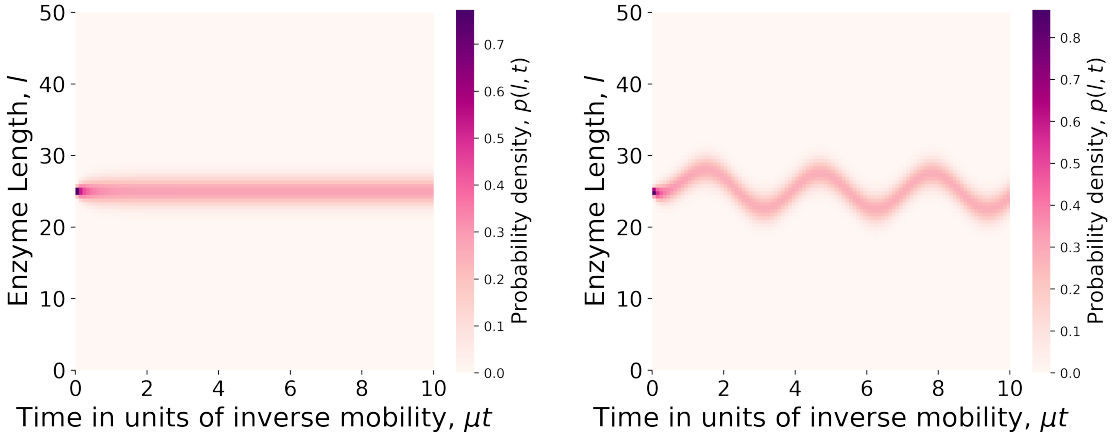


Figure 5.5. The probability density, $p(l,t)$, of enzyme length, l, at time, t. (a) $p(l,t)$ in the case of no external drive, $r = 0$. (b) $p(l,t)$ in the case of an oscillatory external drive with amplitude, $r = 0.1$. Width of the distribution in both the cases is decided by the ratio of elastic constant to the diffusion constant $K/2D$ which here is ..., the amplitude of oscillation is decided by the drive amplitude, r, and the frequency is decided by the drive frequency, ω.

binding of substrate to the enzyme ($\tau_b = 0$). The case of finite binding time τ_b might give a dependence on ω. We will explore this further in future.

This calculation shows that the enzyme specificity can be modulated by a non-equilibrium drive without requiring changes in the molecular structure of the enzyme. For membrane enzymes, like the glycosylation enzymes [8], this drive can be implemented by changing the membrane properties like membrane tension [9]. The implementation of this drive will inevitably cost energy which we can calculate by the total entropy produced by the system.

The entropy production in the system is based on the trajectory thermodynamics [10] where thermodynamic quantities are consistently associated with a trajectory of the system. Here trajectory corresponds to the "shape" or length of the enzyme as a function of time and we ignore the chemical interactions between the substrate and the enzyme in calculation of the entropy production. The entropy of a trajectory($l(t)$) is defined as

$$s(t) = -\log(p(l(t), t)) \tag{5.21}$$

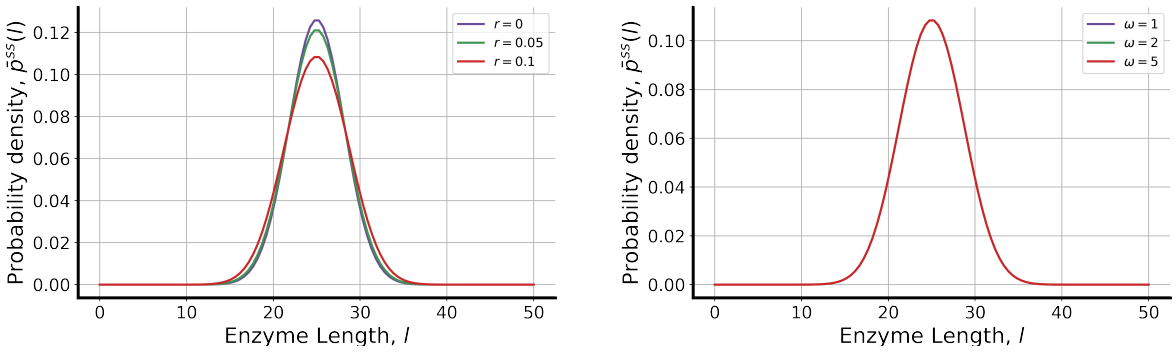


Figure 5.6. The long time probability density, $p^{ss}(l,t)$, of enzyme length averaged over a time period, $2\pi/\omega$, denoted by $\bar{p}^{ss}(l)$ vs the enzyme length l (a) $\bar{p}^{ss}(l)$ for different values of the drive amplitude, r. (b) $\bar{p}^{ss}(l)$ vs l for different values of drive frequency, ω. The enzyme becomes more promiscuous on increasing the drive amplitude whereas changing the drive frequency has no effect on the enzyme promiscuity.

The heat dissipated in the bath by the system associated with increase in entropy of the bath

$$\delta s^m[l(t)] = q[l(t)]/T \tag{5.22}$$

where $q[l(t)]$ is defined as $q[l(t)] = \int_0^t d\tau F(l,\tau)\dot{l} = -\int_0^t d\tau \mu(l - l_0(1 + r\cos\omega\tau))\dot{l}$. The average total entropy production can be written as

$$\begin{aligned} \dot{S}(t) &= \int_l dl(\dot{s}^m(t) + s(t))p(l,t) \\ &= \int_l dl \frac{j(l,t)^2}{Dp(l,t)} = \frac{\langle(\nu(l,t)^2\rangle}{D} \geq 0 \end{aligned} \tag{5.23}$$

where $\nu(l,t) = j(l,t)/p(l,t)$. The current for our enzyme system $j(l,t)$ can be calculated by solving the Fokker Planck equation in (5.5) for a finite region, $(l \in [L_{\min}, L_{\max}])$, with periodic boundary conditions, $(p(L_{\min},t) = p(L_{\max},t))$.

5.4 Future work

In future, we will explore other generic ways of driving the system of enzyme and substrate out of equilibrium. One particularly interesting idea is to manipulate the

statistics of work done by biasing the trajectories towards consuming more energy like done in [11]. Effectively using multiplicative noise instead of white noise.

We also plan to explore the biological consequences of the ability to increase the enzyme promiscuity without changing the molecular structure of the enzyme in a broader context. We will do this by looking at the evolutionary advantages of such a system in a fast changing environment and how this makes the system more adaptable [3, 12].

5.5 Conclusion

- We started with showing that the compartment dependent enzyme specificity results in increased fidelity of synthesis of glycans.
- We extended the enzyme substrate binding model that we used in previous chapters model the enzyme shape as an over-damped elastic spring in presence of thermal bath and external mechanical drive.
- In equilibrium (no external drive), this new model returns the same binding probability distortion relation that we previously used in Chapter 3 and 4.
- In presence of an oscillatory mechanical drive that changes the resting length of the spring, the effective specificity of the enzymes decreases i.e., the enzymes becomes more promiscuous.
- We calculate the energetic cost of increasing the promiscuity of the enzyme by a mechanical oscillatory drive by calculating the total entropy production by the system.

Bibliography

[1] Chakra Chennubhotla, AJ Rader, Lee-Wei Yang, and Ivet Bahar. Elastic network models for understanding biomolecular machinery: from enzymes to supramolecular assemblies. *Physical biology*, 2(4):S173, 2005.

[2] Olivier Rivoire. Geometry and flexibility of optimal catalysts in a minimal elastic model. *The Journal of Physical Chemistry B*, 124(5):807–813, 2020.

[3] Marc Kirschner and John Gerhart. Evolvability. *Proceedings of the National Academy of Sciences*, 95(15):8420–8427, 1998.

[4] Jacque Monod, Jeffries Wyman, and Jean-Pierre Changeux. On the nature of allosteric transitions: a plausible model. *J Mol Biol*, 12(1):88–118, 1965.

[5] Jean-Pierre Changeux and Stuart J Edelstein. Allosteric mechanisms of signal transduction. *Science*, 308(5727):1424–1428, 2005.

[6] Yonatan Savir and Tsvi Tlusty. Conformational proofreading: the impact of conformational changes on the specificity of molecular recognition. *PloS one*, 2(5):e468, 2007.

[7] John J Hopfield. Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity. *Proceedings of the National Academy of Sciences*, 71(10):4135–4139, 1974.

[8] A Varki et al. *Essentials of Glycobiology*. Cold Spring Harbor Laboratory Press, 2009.

[9] Anthony G Lee. How lipids affect the activities of integral membrane proteins. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1666(1-2):62–87, 2004.

[10] Udo Seifert. Stochastic thermodynamics, fluctuation theorems and molecular machines. *Reports on progress in physics*, 75(12):126001, 2012.

[11] Alexandra Lamtyugina, Yuqing Qiu, Étienne Fodor, Aaron R Dinner, and Suriyanarayanan Vaikuntanathan. Thermodynamic control of activity patterns in cytoskeletal networks. *Physical Review Letters*, 129(12):128002, 2022.

[12] Olivier Rivoire and Stanislas Leibler. A model for the generation and transmission of variations in evolution. *Proceedings of the National Academy of Sciences*, 111(19):E1940–E1949, 2014.

Chapter 6

Overview and tasks for the future

In this thesis we look at the process of synthesis of glycans, called glycosylation, in the Golgi complex from the perspective of information theory and explore the how does the glycan function constrains the glycan synthesis machinery consisting of the Golgi complex and glycosylation enzymes.

In the introduction we give a brief overview of information theory and discuss how it can be a useful framework to analyze biological systems. We briefly outline the thermodynamics of out of equilibrium meso-scale systems and describe the phenomenology of glycans, glycosylation and the Golgi complex. In Chapter 2 we explore the notion of complexity for glycan distributions on the cell surface in the context of identification of a cell type in a 'niche'. We find that for reliable identification of many cell types and niches the glycan distribution on the cell surface should be 'detailed' with a lot of peaks. We then look at real glycan data of various organisms like humans, hydra, and quantify the 'complexity' or 'detail' in the glycan profile. We find that complex multicellular organism have complex glycan profiles.

In Chapter 3 we give a basic mathematical model of glycosylation consisting of Golgi compartments, glycosylation enzymes, linear reaction network and unidirectional transport. In Chapter 4 we study the constraints put on the synthesis machinery by the requirement of creating 'complex' glycan distributions. We find that high fidelity synthesis of complex glycan distributions requires a large number of Golgi compartments and glycosylation enzymes. Since increasing the number of enzymes requires an elaborate genetic mechanism this calculation provides a functional motivation for a multi cisternal system. We also find that the glycosylation enzymes have an optimal enzyme specificity, i.e they should not be too specific and

have some promiscuity. The geometry of the fidelity landscape in the multidimensional space of the number and specificity of enzymes, inter-cisternal transfer rates, and number of cisternae provides a measure for robustness and identifies stiff and sloppy directions. In Chapter 5 we start with showing that compartment dependent enzyme specificity can improve the fidelity of synthesis. We then discuss a possible mechanism to have thermodynamic control on the enzyme specificity.

This thesis provides a novel way of looking at the process of Glycosylation and provides an arguement for the need of multi-compartment Golgi complex. This work implies that there should be a coupling between the glycosylation enzyme kinetics and the mechanism of non-equilibrium self assembly of the Golgi complex.

Possible future directions coming out of the thesis:

- Recycling as a cellular strategy for error correction in glycans and adaptation to niche: The surface glycan distribution synthesized by the Golgi machinery is further sculpted by the environment by differential internalization rates based on interaction with the extracellular matrix (ECM). We can write a mechanical model for the interaction between the glycans and the ECM. The introduction of this recycling of the glycans may help in adaptation of the cell to a fast changing environment. The slow response over evolutionary timescales by changing the synthesis machinery and the fast response by recycling and sculpting the synthesized distribution.

- The synthesis model can be extended to include more chemical features of glycosylation like pruning and capping enzymes, and branching in the reaction network. These extension can have interesting effects on the fidelity of synthesis due to the possibility of error correction by these enzymes.

- Extension of the information bottleneck framework to include the effect of host-pathogen interactions on the complexity of glycan profiles.

Chapter A

Convergence of the Magnus sereis

Here, we establish conditions under which the series $\sum_{n=1}^{\infty}\Omega(n,k)$ that defines solution $C(k)$ to the differential equation (3.19) converges.

The commutator

$$[M(k_1),M(k_2)]=\begin{bmatrix} 0 & 0 & 0 & 0 & \dots & 0 \\ a_{21} & 0 & 0 & 0 & \dots & 0 \\ a_{31} & a_{32} & 0 & 0 & \dots & 0 \\ 0 & a_{42} & a_{43} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & \dots & & a_{n,n-2} & a_{n,n-1} & 0 \end{bmatrix}$$

where

$$\begin{aligned} a_{i,i-1} &= A^{(i-1)}(k_2)B^{(i)}(k_1)+A^{(i)}(k_1)B^{(i)}(k_2)-A^{(i-1)}(k_1)B^{(i)}(k_2)+A^{(i)}(k_2)B^{(i)}(k_1) \\ a_{i,i-2} &= B^{(i-1)}(k_2)B^{(i)}(k_1)-B^{(i-1)}(k_1)B^{(i)}(k_2) \end{aligned}$$

The general form of $\Omega(n,k)$ is given by [**?**]

$$\Omega(n,k)=\frac{z_n}{n!}\int_0^k dk_1\int_0^{k_1}dk_2\dots\int_0^{k_{n-2}}dk_{n-1}\int_0^{k_{n-1}}dk_n\sum_l W_l M(k_{p_1^l})M(k_{p_2^l})\dots M(k_{p_n^l}) \tag{A.1}$$

where $(p_1^{(l)},p_2^{(l)}\dots p_n^{(l)})$ is a permutation of $(1,2,3,\dots n)$, $W_l\in\{-1,1\}$, and $z_n\in 1,\dots n$.

Let $\bar{A} = \max_{k,l,m} |M_{l,m}(k)|$. Define

$$\bar{M} = \begin{bmatrix} \bar{A} & 0 & 0 & 0\dots & 0 \\ \bar{A} & \bar{A} & 0 & 0\dots & 0 \\ 0 & \bar{A} & \bar{A} & 0\dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & \dots & 0 & \bar{A} & \bar{A} \end{bmatrix}$$

We can bound all the matrix elements of $\Omega(n,k)$ in the following way

$$\begin{aligned} \Omega_{lm}(n,k) &\leq z_n \bar{M}^n_{l,m} \int_0^k dk_1 \int_0^{k_1} dk_2 \dots \int_0^{k_{n-1}} dk_n \\ &= z_n \bar{M}^n \Big|_{lm} \frac{k^n}{n!} \end{aligned} \tag{A.2}$$

The matrix

$$\bar{M}^n = \begin{bmatrix} a_{11} & 0 & 0 & 0 & \dots & 0 \\ a_{21} & a_{22} & 0 & 0 & \dots & 0 \\ a_{31} & a_{32} & a_{33} & 0 & \dots & 0 \\ a_{41} & a_{42} & a_{43} & a_{44} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ a_{n1} & \dots & & a_{n,n-2} & a_{n,n-1} & a_{nn} \end{bmatrix}$$

where $a_{lm} = S_{lm}(n)\bar{A}^n$ for appropriately defined polynomials $S_{l,m}(n)$. Thus, it follows that $\Omega_{lm} \leq z_n S_{lm}(n)(A^*)^n \frac{k^n}{n!}$ and $\Omega_{l,m}(k) \leq \sum_{n=1}^{\infty} z_n S_{l,m}(n)(A^*)^n \frac{k^n}{n!}$. Consequently, the series will converge if $\bar{A}k < 1$, i.e. $k \leq \frac{1}{\bar{A}}$. Assuming $\mu^{(j)} = \mu \ \forall j$, we can bound $\bar{A}$ as

$$\begin{aligned} \bar{A} \leq \ & \max_{j,k} \left(\frac{\mu + \sigma \sum_{\alpha=1}^{N_E} R(j,k,\alpha) \exp(-\sigma|k - l_\alpha^{(j)}|)}{\mu + \sum_{\alpha=1}^{N_E} R(j,k,\alpha) \exp(-\sigma|k - l_\alpha^{(j)}|)} \right. \\ & \left. + \frac{\sum_{\alpha=1}^{N_E} R'(j,k,\alpha) \exp(-\sigma|k - l_\alpha^{(j)}|)}{\mu + \sum_{\alpha=1}^{N_E} R(j,k,\alpha) \exp(-\sigma|k - l_\alpha^{(j)}|)} \right) \end{aligned} \tag{A.3}$$

Since the parameters μ, σ, $R(j,k,\alpha)$, $l_\alpha^{(j)}$ and N_E are finite and positive, and $R'(j,k,\alpha)$ is finite, $\bar{A}$ has a finite upper bound, implying k is always greater than zero, and the series has finite radius of convergence.

Chapter B

Numerical scheme for performing the optimization

We solve Optimization C using the numerical scheme detailed below. The optimization problem consists of minimising a non-convex objective with linear box constraints. We use the MATLAB FMINCON function to solve this optimization. We use Sequential Quadratic Programming (SQP), a gradient based iterative optimization scheme for solving optimizations with non-linear differentiable objective and constraints. Since our problem is non-convex and SQP only gives local minima, we initialise the algorithm with many random initial points. We use SOBOLSET function of MATLAB to generate space filling pseudo random numbers. We have taken 1000 initialisations for each N_E, N_C and σ value. We have taken 50 equally spaced points between 0 and 1 to explore the σ-space for Fig. 4.1. Some minor fluctuations in D due to non-convexity of the objective function in the final results were smoothed out by taking the convex hull of the D vs. σ graph. The results for $\sigma_{min}(N_E, N_C)$ and $D(\sigma_{min}, N_E, N_C)$ (Fig. 4.2) were obtained by adding σ to the optimization vector and then performing the optimization.

A similar numerical scheme was used to optimize diversity.

Chapter C

Parameter estimation

The typical transport time of glycoproteins across the Golgi complex is estimated to be in the range 15-20 mins. [1], which corresponds to the transport rate $\mu = 0.18/\text{min}$. We bound the transport rate for our optimization between 0.01/min and 1/min.

Next, we estimate the range of values for the chemical reaction rates. The injection rate q is in the range $100 - 1500$ pmol/10^6 cell 24 h [1, 2]. For our calculation we set $q = 387.30$ pmol/10^6 cells 24 hr $= 0.27$ pmol/10^6 cells min, where 387.30 is the geometric mean of 100 and 1500. We set the range for the enzymatic rate R to be

$$R_{\min} = \min_{\alpha} \left\{ \frac{V_{\max}^{(\alpha)}/\nu}{K_M^{(\alpha)} + \frac{1}{\nu}\frac{q}{\mu}} \right\} \leq R \leq R_{\max} = \max_{\alpha} \left\{ \frac{V_{\max}^{(\alpha)}/\nu}{K_M^{(\alpha)}} \right\}.$$

where $K_M^{(\alpha)}$ and $V_{\max}^{(\alpha)}$ denote the Michaelis constants and $V_{\max}$ of the α-th enzyme. The conversion from 1 pmoles/10^6 cells to concentration can be obtained by taking cisternal volume (ν) to be $2.5\mu m^3$ [1, 2]. This gives

$$1 \text{ pmoles}/10^6 \text{ cells} = \frac{10^{-12}\text{moles}}{10^6 \times 2.5 \times 10^{-18} \times 10^3\text{litre}} = 400\mu M. \qquad \text{(C.1)}$$

In Table C.1 we report the parameters for the 8 enzymes taken from Table 3 in [1].

From these parameters it follows that

$$\begin{aligned} R_{\min} &= \min_{\alpha}\left\{\frac{V_{\max}^{(\alpha)}/\nu}{K_M^{(\alpha)}+\frac{1}{\nu}\frac{q}{\mu}}\right\} \\ &= \frac{V_{\max}^{(7)}/\nu}{K_M^{(7)}+\frac{1}{\nu}\frac{q}{\mu}} = \frac{.16\times 400\mu M/\min}{3400\mu M+149.4\mu M} = 0.018\min^{-1} \\ R_{\max} &= \max_{\alpha}\left\{\frac{V_{\max}^{(\alpha)}/\nu}{K_M^{(\alpha)}}\right\} \\ &= \frac{V_{\max}^{(1)}/\nu}{K_M^{(1)}} = \frac{5\times 400\mu M/\min}{100\mu M} = 20\min^{-1} \end{aligned}$$

α	$K_M^{(\alpha)}$ (μmol)	$V_{\max}^{(\alpha)}$ (pmol/10^6 cell-min)
1	100	5
2	260	7.5
3	200	5
4	100	5
5	190	2.33
6	130	.16
7	3400	.16
8	4000	9.66

Table C.1. Enzyme parameters taken from Table 3 in [1] that we use to calculate the bounds on the reaction rate R. Here $K_M^{(\alpha)}$ and $V_{\max}^{(\alpha)}$ denote the Michaelis constant and $V_{\max}$ of the α-th enzyme.

Bibliography

[1] Pablo Umaña and James E Bailey. A mathematical model of n-linked glycoform biosynthesis. *Biotechnology and bioengineering*, 55(6):890–908, 1997.

[2] Frederick J Krambeck, Sandra V Bennun, Someet Narang, Sean Choi, Kevin J Yarema, and Michael J Betenbaugh. A mathematical model to derive n-glycan structures and cellular enzyme activities from mass spectrometric data. *Glycobiology*, 19(11):1163–1175, 2009.