**CHAPTER 4**

**Characterization of highly mutable viruses to identify targets for potent vaccine-induced immune responses**

4.1: Introduction to vaccines

As we noted in Chapter 1, infectious disease-causing microorganisms have plagued humanity since time immemorial. Human ingenuity led to better sanitation, vaccines, and therapies that have considerably mitigated the threat of infectious diseases. Vaccination, in particular, has saved more lives than any other medical procedure. The safe and effective vaccines that we have today are the result of attempts made over many centuries to protect people from infectious diseases.

The history of vaccination can perhaps be divided into four periods. In the first period, it was not known that microbial pathogens cause infectious diseases or that we have an immune system. The variolation procedure developed by the Chinese and Indians to protect against smallpox infection was likely the earliest example of such an effort. In this procedure, material derived from smallpox sores and scabs was stored for a while and then administered to a healthy person. While variolation could be protective, it was a dangerous procedure as it could sometimes cause full-blown disease and local outbreaks. In 1796, Edward Jenner developed a safe and efficacious method to protect against smallpox. His vaccine coupled with global cooperation ultimately eradicated smallpox from the planet in 1979. This was a huge accomplishment as smallpox was a deadly and ancient disease that killed millions of people in the world as recently as the 1950s. Jenner’s vaccine was based on administering material from cowpox sores into humans. Cowpox is a harmless disease for humans, but its proteins are sufficiently similar to those of smallpox that the immune response elicited by inoculation with cowpox conferred protection against smallpox.

The second period of vaccine development began in the 19th century when advances made by many scientists, prominently Robert Koch and Louis Pasteur, led to the discovery that microbial pathogens such as bacteria cause infectious diseases. This allowed for a more principled approach to the development of protective measures. Pasteur’s work led to a paradigm change in procedures to protect people from infection. Rather than inoculating with a related pathogen that was largely harmless as Jenner did, Pasteur understood that administering a live but attenuated form of the disease-causing pathogen can be safe and efficacious. The pathogen had to be attenuated in a way that it was still a live pathogen with its proteins intact, but it did not cause disease. One way to attenuate a pathogen is by repeatedly passing it through animals that are not a natural host for the pathogen. The pathogen would then adapt to survive better in the new animals, and become less virulent for the original hosts (e.g., humans). Pathogens can also be attenuated by repeated passage in cell culture. The vaccine against rabies developed by Pasteur and Emil Roux is an example of a live attenuated vaccine. The Rabies virus was obtained from saliva from a rabid dog, and passaged through rabbits to obtain the material that Pasteur used as the rabies vaccine for humans. Pasteur labelled all such protective procedures as vaccination in honor of Jenner’s work using cowpox (or *vaccinia*). A second approach to vaccination was also developed in the 19th century. This procedure involved killing (or inactivating) the pathogen. For example, formaldehyde was often used to inactivate pathogens. This procedure had to be done just right so that the killed pathogen still had proteins that were more or less intact so that antibodies and T cells that targeted them upon vaccination were protective against a real infection. Interestingly, Pasteur and others in the 19th century did not know about our immune system.

In the 20th century, the two approaches for vaccination developed in the 19th century were used by Jonas Salk and Albert Sabin to develop their polio vaccines. Salk’s method used an inactivated vaccine, and Sabin’s was a live attenuated vaccine. Wide-spread use of Sabin’s vaccine has resulted in the near eradication of polio in most of the world. Most vaccines developed in the 20th century roughly followed one of the two approaches noted above. For example, an inactivated vaccine is used to protect against influenza.

With the development of recombinant DNA technology, a third phase of vaccine development emerged. Scientists learned how to make large amounts of just the pathogen’s proteins using recombinant DNA technology. These proteins could be used as the vaccine material to elicit immune responses that could target the pathogen’s proteins upon infection. The efficacious vaccines that protect against Hepatitis B virus and Human Papillomavirus infections are examples of such so called “subunit vaccines”. Another innovation in the 20th century was that additives (called adjuvants) were added to vaccines to help stimulate innate immune responses that are critical for the development of potent adaptive immune responses. Adjuvants are a very important component of subunit vaccines. As mechanistic understanding of innate immunity has emerged, the design of adjuvants is becoming more systematic, rather than purely empirical. For example, adjuvants are designed to stimulate toll like receptors.

In the 21st century, the COVID-19 pandemic led to the successful deployment of another effective and safe vaccine modality. All past vaccines included either the pathogen against which protection was desired or its proteins. Information about the proteins that need to be targeted by the immune system to protect against infection can also be encoded in the corresponding DNA or RNA sequences. The mRNA vaccines that proved to be highly effective for protection against infection by SARS-CoV-2, the virus that causes the COVID-19 disease, encode information about the spike proteins of this virus as messenger RNA encapsulated in a lipid nanoparticle. The nanoparticles are designed to enter human cells. The RNA encoding the spike protein is translated to the protein using the host cell’s machinery, and it elicits an efficacious antibody response. The RNA is quickly degraded by various mechanisms. Another approach that encodes information about pathogenic proteins in the form of DNA was also developed. In this approach, a virus that typically causes mild diseases, like an Adenovirus, is first engineered such that it cannot replicate. Therefore, it cannot cause disease. The DNA encoding the pathogen’s proteins is then inserted into the DNA of the engineered Adenovirus. The Adenovirus can infect human cells, and the DNA encoding information about the pathogen’s proteins is transcribed and translated into the proteins by the host cell’s machinery. This elicits an immune response. The infected cells are soon cleared by the immune response.

In spite of the success of vaccines, we do not have effective against various pathogens. Prominent examples are Human Immunodeficiency Virus (HIV), Hepatitis C Virus, tuberculosis, dengue and malaria, many of which wreak havoc around the world. We do not have a broadly effective vaccine against seasonal variants of influenza either, and attempts to predict the right vaccine for the ensuing year are not always successful. Many of these pathogens share two features: 1] They present themselves in different guises, thus making them hard to target with specific immune memory responses. For example, new mutant strains of influenza evolve with time and become dominant in the population. 2] They often degrade, or hide from, the immune system. For example, the parasite that causes malaria infects red blood cells that do not express MHC proteins and so are not targeted by T cells.

HIV has characteristics which are extreme examples of both these features. It is a highly mutable virus with a rapid replication rate. Thus, it generates many mutant strains when it infects a person, and an enormous diversity of viral strains is circulating in the human population ((Fig. 4.1). The diversity of HIV strains dwarfs that of influenza. The high mutability allows HIV to evade natural or vaccine-induced immune responses [12]. For example, the infecting strain may not be the one for which vaccine-induced memory immune responses exist; even if the infecting strain is targeted by the memory response, if the virus is not eliminated rapidly, the infecting strain can mutate in the host to escape from this response. Furthermore, HIV principally infects and eventually kills human T helper cells, thus degrading the adaptive immune system. This is the reason why Acquired Immunodeficiency Syndrome (AIDS), the disease caused by HIV infection, results in a severe state of immunodeficiency. This results in patients being afflicted by many normally easy to control infections.

Successful vaccination against pathogens that have evolved sophisticated strategies to evade human immune responses will benefit from the development of firm scientific principles that can guide vaccine design [10]. At least three important issues must be studied in this regard: 1] What are the appropriate targets in the pathogen’s proteins on which to focus vaccine-induced immune responses such that the ability of the pathogen to evade such responses by mutation while simultaneously maintaining their viability/virulence is severely limited or eliminated? The answer to this question will determine the design of the active component of the vaccine, or the immunogen. 2] How can immune responses targeting the mutational vulnerabilities identified by answering question 1 be induced by vaccination in humans with diverse genotypes? 3] How should the immunogen be delivered? It is interesting that the mRNA and Adenovirus-vectored vaccine modalities that were first deployed during the COVID-19 pandemic were in development for potential use as delivery vectors for HIV vaccines.

A convergence of several factors is beginning to enable us to take the first steps toward addressing these questions. Biologists and clinicians can collect enormous amounts of data on sequences of mutant strains of pathogens, and it is also becoming possible to interrogate the immune system on an unprecedented scale. Both the immune system and pathogens function and interact via collective processes that involve myriad individual components, thus making mechanistic interpretation of this data complex. Computer scientists and statistical physicists have begun to play a role in translating this type of data to mechanistic knowledge that can help address the questions noted above. Engineers and clinician/scientists are beginning to design ways to more effectively deliver vaccines as well.

In this chapter, we will focus primarily on defining the mutational vulnerabilities of highly mutable pathogens (question 1 above). For concreteness, we will consider only one pathogen, HIV, but we will point out how related challenges for influenza are different. We will also note how immunogens could be designed to elicit T cell responses that may target the identified mutational vulnerabilities. In the next chapter, we will consider how vaccination strategies can be designed to elicit antibodies that target mutational vulnerabilities (question 2 above).

4.2: Brief description of the biology of HIV

HIV was transmitted to humans from non-human primates, and is estimated to have been circulating in small populations of humans for nearly a century [13]. The first well-documented cases were reported in 1981 in the United States. To date, HIV has infected over 75 million people, about 40 million people are estimated to have died from complications associated with AIDS, and 680,000 HIV-related deaths were recorded in 2020. In developed nations, HIV infections can be effectively controlled and managed because of the wide-spread availability of potent drugs, but it cannot be cured. In many parts of the world HIV continues to be a major public health problem, with sub-Saharan Africa being the epicenter of the disease. For example, although the mortality rates are beginning to stabilize, each day there are approximately 1000 new HIV infections in South Africa. A vaccine or cure is needed to eradicate HIV from the planet, but no successes have been reported after more than forty years of work and expenditure of large amounts of money.

As is obvious from its name, HIV is a virus. Viruses can be classified into two broad categories, DNA and RNA viruses. DNA viruses carry their genetic information in the form of DNA, while RNA viruses encode their genome as RNA. Smallpox is an example of a DNA virus and influenza and SARS-CoV-2 are examples of RNA viruses. Viruses infect host cells and hijack the host cell’s transcriptional machinery. Thus, the viral DNA or RNA is transcribed and translated to result in the synthesis of viral proteins and subsequent assembly of new virus particles that can infect other cells and people.

HIV belongs to a class of RNA viruses called retroviruses. HIV has an outer membrane through which protrudes its spike. The spike is made up of non-covalently bonded trimers of two “Envelope” proteins called gp 120 and gp 41 (Fig. 4.2). The outer membrane surrounds a nucleocapsid made up of structural proteins which encloses two copies of the virus’ RNA genome and other key proteins important for virus replication (Fig. 4.2). The life cycle of HIV is depicted schematically in Fig. 4.3. The trimeric spike binds to host cell surface proteins to initiate infection [14]. For example, the spike binds to the CD4 co-receptor expressed on the surface of T helper cells [15,16]. This binding event leads to a conformational change in gp 120 and the dissociation of gp 41, which then forms a six-helix bundle. The conformational change enables gp 120 to bind to a second receptor (called a co-receptor) on the surface of the host cell. These binding events and the free energy gained from gp 41 forming an ordered state result in fusion of the virus’ membrane with that of the host cell membrane, resulting in release of the nucleocapsid in to the cytoplasm. The nucleocapsid is then uncoated, thus releasing its contents. A viral protein, Reverse Transcriptase, then converts the RNA genome into the complementary DNA. This viral DNA, called a provirus, is transported into the nucleus of the host cell along with a viral protein called Integrase. Integrase inserts the viral DNA in to the genome of the host cell, thus infecting this cell for its lifetime. The HIV genes code for polyproteins (a number of concatenated proteins), and the transcriptional machinery of the host cell is hijacked to express these polyproteins. The polyproteins are chopped up by a HIV protein, called Protease, resulting in individual proteins that mediate viral function. In a series of steps, the virus’ proteins are properly assembled at the membrane of the host cell [17]. A part of the host cell membrane becomes the membrane for a new virus particle as it buds out and matures into a functional virus.

HIV has nine genes, with four of them, Gag, Pol, Nef, and ENV being the most important. Gag encodes structurally important proteins (such as the ones that make up the nucleocapsid), Pol codes for Reverse Transcriptase, Integrase, and Protease, ENV codes for gp 120 and gp 41, and Nef codes for proteins with many functions that include a role in downregulation of CD4 and HLA proteins. Downregulation of HLA proteins suppresses T cell responses and downregulation of CD4 helps the virus bud out of infected cells because that inhibits binding of the viral spike to the infected cell’s CD4 co-receptor.

Upon successful infection, the virus replicates rapidly, and the viral load (that is, the number of viruses circulating in the host) increases (Fig. 4.4). Infected cells ultimately die as new virus particles bud out, and so the number of CD4 T cells in an infected patient declines rapidly. As immune responses develop, the viral load decreases and then stabilizes to a steady state value [21]. The number of CD4 T cells also increases and stabilizes. During this phase when the viral load and CD4 T cell counts are stable there is a dynamic “arms race” between the virus and the immune system. The immune system mounts a response directed at the prevailing viral strains which then mutate to evade these responses, then new responses develop, and the cycle continues. Without treatment the immune system ultimately loses the battle, viral load goes up, and CD4 T cells decrease to low numbers (Fig. 4.4). At this point, the individual’s immune system is severely compromised, and many opportunistic infections ensue leading to AIDS and ultimately to death.

The two main sources of mutations in HIV’s lifecycle arise from reverse transcription not being a very high-fidelity process and mistakes in generating HIV proteins from the viral DNA using the human transcriptional machinery. Mutations arise at an average rate of 3 × 10-5 per base pair per replication cycle [18]. HIV’s genome is about 104 base pairs in length, and so this implies that during every replication cycle the probability of evolving a mutant strain is 0.3. Moreover, Reverse Transcriptase can hop from one of HIV’s RNA molecules to another. When two different RNA genomes are available, this causes recombination of the genomes of two viral strains, leading to even more diversity. Fitting parameters in ordinary differential equations describing viral dynamics to data from patients treated with drugs revealed that HIV replicates very rapidly [19], producing 1010 to 1011 virus particles per day in infected humans [20]. Many of the mutant strains that are produced cannot form infectious virus particles, but several mutants are viable. The high mutation and replication rates of HIV and the chronic infection it causes are the main reasons underlying the extraordinary diversity of HIV strains circulating in the population and in individuals (Fig. 4.1). It is also the principal reason why we do not have an effective vaccine or cure for HIV.

4.3: Determining the mutational vulnerabilities of HIV

At first glance, determining the mutational vulnerabilities of a virus like HIV appears to be a simple problem given that today we can sequence large numbers of viruses derived from diverse patients. These data are the results of a large-scale natural experiment, and provide an opportunity to learn the mutational vulnerabilities from virus evolution in humans with diverse genotypes. Looking for sites or regions that exhibit a relatively low level of amino acid variation (e.g., measured by low entropy of amino acid variation) in the sequence data should help identify mutational vulnerabilities that could be the targets of potent immune responses. This is because to evade an immune response focused on such sites or regions the virus would have to evolve mutations from favored amino acids to others. Such mutations would likely hurt the virus’ ability to replicate and function, as there must be a reason why the sequence data exhibit a relatively low frequency of amino acid variation at these sites or regions. However, this strategy can be blunted because it has been observed in different cases that the virus can evolve other mutations, so-called compensatory mutations, which can partially restore the fitness cost incurred by making the primary immune-evading mutation [29,30]. Even though such compensatory interactions can be relatively rare, the high mutability and replication rate of HIV can allow them to be sampled. This is especially so when immune responses provide a selection force that promotes the evolution of one of the involved mutations; viz., the one that allows the virus to escape immune responses.

Therefore, if one wishes to determine the mutational vulnerabilities of a virus like HIV, one needs to define the collective mutational pathways that HIV uses to evade human immune responses in order to avoid targeting the involved sites or regions with vaccine-induced immune responses. On the flip side, one also needs to determine the combinations of mutations that the virus cannot make and remain viable, as these are targets for potent immune responses. This is because such immune responses would corner the virus between being killed by the immune responses or evolving unfit mutant strains that cannot function effectively. In short, one needs to determine the fitness landscape [31,32] of the virus. By that we mean the ability of the virus to replicate, function and propagate infection in humans as a function of its sequence, with explicit account for the coupled effects of mutations at different protein sites. Fig. 4.5 illustrates the concept of a fitness landscape using a 2-dimensional representation of sequence space; i.e., of a cartoon virus comprised of one protein with two sites. With a vaccine-induced immune response, one wishes to target sites or regions which are such that mutations therein correspond to a fitness valley. One also wishes to block the mountain passes corresponding to the collective compensatory mutational pathways that HIV uses to go from one fitness hill to another adjacent one when the first one is under immune attack.

Armed with the fitness landscape, one could design immunogens that could induce potent T cell responses. Such an immunogen would not be comprised of whole HIV proteomes as is traditional, but would contain only parts of it chosen according to three criteria: 1] Regions that are rife with compensatory pathways are minimized. 2] Regions wherein multiple mutations are especially deleterious are maximized. 3] Regions that can be presented by people with diverse HLA genes are maximized. Knowledge of the fitness landscape can also guide the choice of immunogens that could induce potent antibody responses, a topic that we will consider in the next chapter.

*4.3.1: Inferring the fitness landscapes of HIV polyproteins*

Inference of the fitness landscape from sequence data can only be done in a statistically meaningful way if a sufficient number of sequences are available. At the time of writing, the number of sequences of whole genomes of HIV strains circulating in the population is insufficient for this purpose. However, a sufficient number of sequences of the HIV polyproteins are available. Therefore, we will focus on the fitness landscapes of HIV polyproteins. The fitness landscape of the virus thus obtained ignores co-evolution of proteins encoded by different genes. From a biological standpoint, this limitation could be a problem. Fortunately, intergenic epistasis (coupled effects of mutations) seems to be rare for HIV, but it is not absent. From a methodological standpoint, this is a detail, since the same methods could be applied to whole genomes when a sufficient number become available (albeit with greater numerical challenges).

We wish to obtain the intrinsic fitness landscape of HIV polyproteins and then push and pull on this landscape with a vaccine-induced immune response so as to corner the virus as described above. The available data on HIV polyprotein sequences in circulation was obtained by sequencing viruses extracted from patients, and is thus a record of virus evolution in the human population. Inferring the intrinsic fitness landscape characterizing HIV polyproteins from this data requires accounting for the role of human immunity in shaping the prevalence of circulating strains represented by the sequence data. Thus, the inference of the fitness landscape must proceed in two steps: 1] First, infer a model for the prevalence of circulating strains based on the sequence data (the prevalence landscape); 2] second, deconvolute that effects of human immunity to infer the desired fitness landscape.

*4.3.1a: Inferring the prevalence landscape of HIV polyproteins*

The prevalence landscape of any polyprotein is a mathematical representation of the probability P(**z**) of observing a sequence, **z**, in the sequence data [46]. This data also has information on the probability of observing single mutations at every site of a polyprotein, double mutations at every pair of sites, triple mutations at every triplet of sites, etc. Any mathematical model for P(**z**) that can recapitulate these mutational correlations will accurately describe the prevalence landscape. One way to infer P (**z**) approximately is to ask: what is the “least biased” model for P(**z**) that constrains the one- and two-point mutational correlations to be the observed ones. Only one and two-point correlations are considered because typical finite sequence data sets do not contain statistically meaningful values for the higher order correlations.

As described in the Appendix to this chapter, exploiting the connection between statistical mechanics and information theory, one may interpret “least biased” to mean the probability distribution P(**z**) which has the maximum entropy subject to the constraints of preserving the observed mutational correlations [47]. A similar approach has been used to infer contacts in protein structures [48,49], correlations between the firing of neurons [50], etc. Related methods have also been employed to study structural properties of HIV protease [51] and inter-protein interactions [52]. Indeed, the development and use of such inference methods is an active area of research.

For most HIV polyproteins, a number of amino acids are observed at each site in the circulating viral strains. So, the sequences are represented by a “Potts” model, with a number of possible states of each site. For ease of illustration of the maximum entropy-based inference procedure, let us use an “Ising” model. Amino acids at a site are either the most frequently observed one (represented by a value = 0) or a mutant (represented by a value = 1). The generalization to Potts models is straightforward. With this simplified notation, the quantity we wish to maximize is

(1)

where the constraints are enforced through the Lagrange multipliers α, and the sets {hi} and {Jij}. Here pi and pij are the observed one- and two-point mutational correlations, respectively. Maximizing this functional with respect to P (z) yields:

(2)

The formula for P (**z**) is the familiar Boltzmann distribution in statistical mechanics with 1/kBT = 1 (kB is the Boltzmann constant). The analog of the partition function, Z, ensures that the probability distribution is properly normalized. In analogy with Ising models, we will refer to the {hi} as fields and the {Jij} as couplings, and they constrain the one and two-point mutational correlations to the observed values. The more negative the value of hi, the lower the frequency of observed mutations at site i. Positive and negative values of Jij correspond to compensatory and antagonistic effects of double mutations at the i-j pair of residues, respectively.

Once we choose the form of P (**z**) as in Eq. 2, the fields and couplings that recapitulate the observed mutational correlations can also be shown to maximize the likelihood of the observed sequence data. The likelihood of the observed data is given by the product of the probabilities predicted by the model for observing each sequence in the data. The log of this likelihood divided by the number of sequences B () is thus:

(3)

where **z**(k) represents the *k*th sequence. Maximizing with respect to the parameters (the fields and the couplings) yields:

(4)

Recognizing that the first terms in the two equations above are pi and pij, respectively, and the second terms are the model predictions for the same quantities, shows that the fields and couplings that maximize the likelihood of the data are also the ones that recapitulate the observed one- and two-point mutational correlations. Therefore, the fields and couplings we seek can be obtained by maximizing the likelihood of the sequence data. This inference problem is referred to as the “inverse Ising/Potts” problem.

The simplest way to solve this problem and infer the fields and coupling employs a Monte-Carlo scheme. You start with an initial guess for {hi} and {Jij}, and generate a large ensemble of sequences according to P (**z**) ~ e-H (**z**) using Monte-Carlo sampling. Briefly, this is done as follows. Attempt to mutate sites in a given sequence with a certain probability (e.g., the mutation rate of HIV). Accept the mutation (the new sequence thus generated) with a probability equal to , where ) is the change in H upon making the mutation. Repeating this procedure many times and harvesting sequences after some “equilibration time” generates a set of sequences according to the prevalence probabilities predicted by Eq. 2. Compute {pi} and {pij} from this ensemble of sequences and compare the values to the empirically observed ones in the sequence data. Use a standard gradient descent method to generate a new estimate for {hi} and {Jij}, and repeat until a convergence criterion is met. Such a method is computationally prohibitive for all but the smallest systems because of the burden of carrying out the Monte-Carlo sampling a large number of times. For example, with just an Ising representation for the HIV protein, p24 (which is 231 amino acids long), such an algorithm required of the order of 108 steps in each Monte-Carlo sampling step to obtain converged averages for {pi} and {pij}, and 20,000 iterations for convergence of a gradient descent method.

The reason that direct and simple methods like the one described above become intractable is that the likelihood of the data depends upon the partition function (Eq. 3). In principle, as there are 20 possible amino acids (or states) for each site, the partition function for a Potts model for a protein of length, N, involves summing 20N terms. The computational cost thus grows exponentially with the length of the protein. This is reflected in the computational method above as the Monte-Carlo sampling is tantamount to calculating the partition function. A number of methods have been developed to address this challenge. The reader is encouraged to peruse descriptions of algorithms like iterative scaling, pseudo-likelihood, perturbation expansions, mean-field (or Gaussian) models, minimum probability flow, etc. Here, we will discuss only one algorithm that has proven to be useful for most HIV proteins.

*The Adaptive Cluster Expansion (ACE) Method*

This method takes advantage of the fact that, for most proteins, the matrix of couplings, {Jij}, is sparse. Notice that Eq. 3 can be expressed in the following ways:

(5)

The first term on the right-hand side can be interpreted as an analog of the Helmholtz Free energy, and the second as the negative of the energy. So, can be viewed as an entropy, S. We can define an intensive entropy, S\*, as follows:

(6)

S\*, is referred to as the intensive cross-entropy between the data and the inferred model. Maximizing the likelihood of the data is tantamount to minimizing the cross entropy. Therefore, the fields and couplings that minimize the cross entropy are those that constrain the one and two-point mutational correlations predicted by P (**z**) to be the observed values. The ACE algorithm minimizes the cross entropy while taking advantage of the sparsity of the matrix of couplings.

Before we describe the ACE algorithm, let us note that the sequence data on which any such algorithm operates is under-sampled because only a finite number of sequences are available. A procedure that finds the fields and couplings by minimizing the cross entropy must account for this under-sampling. For sufficiently small values of B, Bpi or Bpij could be much less than unity. If so, the corresponding single or double mutant would likely not be observed in the data. This would result in inferring that the magnitude of the corresponding field or coupling is infinite (so, that the weight of a sequence with a mutation at i, or the double mutant at the i,j pair, is zero). This inference would be incorrect, and purely a manifestation of under sampling of sequences. For sequences of viral proteins, such as the cases of interest in this chapter, this problem is of concern mostly for estimates of the couplings because the values of B are sufficiently high. To illustrate this point, consider uncorrelated mutations wherein pij = pip. It is easy to see from this example that the value of Bpij can be much less than 1 for values of B for which Bpi > 1.

A principled way to account for finite sampling of sequences is obtained by assuming some prior knowledge of the distribution of the couplings. Using Bayes theorem in statistics, we can write:

(7a)

or equivalently,

(7b)

In Eq. 7, the set of fields and couplings ({hi} and {Jij}) are denoted by and , respectively, is the likelihood of the sequence data, given the fields and couplings, is a prior estimate of how the fields and couplings are likely to be distributed, is the likelihood of the fields and couplings having specific values given the data, and is the probability of observing a particular set of sequences in the data out of all possible Potts models (appears only as a normalization for ).

As we noted earlier, for the sequence data on HIV polyproteins, the finite size of the data principally impacts the estimated two-point mutation correlations. The matrix is expected to be sparsebecause mutations at most pairs of residues are unlikely to have coupled effects. In the absence of more precise information, it is therefore reasonable to assume that the couplings, Jij, are distributed as a Gaussian with zero mean:

(8)

As before, N is the number of sites in the polyprotein. Now noting that, by definition, the log likelihood of the data, given the fields and couplings, is , the numerator on the right-hand side of Eq. 7b can be written as:

(9)

with . Given that is just a normalization factor in Eq. 7b, Eq. 9 is proportional to . We would like to find the fields and couplings that maximize , which is equivalent to minimizing . That is, to obtain the best estimate for the fields and couplings while accounting for under sampling of the sequence data, we should minimize (referred to as the regularized cross entropy). It should be evident from Eqs. 6, 7b and 9 that minimizing the regularized cross entropy is the same as maximizing the likelihood of the data with a Gaussian prior for our expectation about the matrix **.** The regularization term penalizes large values of Jij during the optimization procedure, thus partially mitigating the effects of under sampling in the sequence data. Although some systematic ways have been proposed to choose the value of , the choice is really somewhat arbitrary. It is sometimes chosen such that the probabilities of observing higher order mutations in the sequence data are also predicted reasonably by the inferred quadratic model for P (**z**).

Since we want to obtain the fields and couplings by minimizing the regularized cross entropy, it is appropriate to ask whether has a unique minimum with respect to the fields and couplings; i.e., is its Hessian with respect to these parameters positive definite? The terms in the Hessian matrix of S\* can be obtained from Eq 6. They have the following form:

(10a)

(10b)

(10c)

That is, the Hessian of the cross entropy is the covariance matrix of mutational correlations. A covariance matrix is positive definite, which can be most readily illustrated by considering a zero centered two-point covariance matrix, **C**, with individual terms, <zizj>. For the matrix, **C**, to be positive definite, we require that the scalar, , for any vector, , be positive. This scalar quantity can be re-written as follows to show that it is positive:

(11)

The term added to S\* in Eq. 9 in order to regularize it has a Hessian that equals , where ***I*** is the identity matrix. So, the regularized S\* has a positive definite Hessian, and so has a unique minimum with respect to the fields and couplings.

The ACE algorithm infers the fields and couplings that minimize the regularized cross entropy using a recursive algorithm that considers finite size clusters of sites on a polyprotein that are connected by couplings. Starting with independent sites, the size of the connected clusters grows as per the following algorithm:

1] Start with single independent sites. In this case, we can directly minimize the cross entropy to obtain estimates for the fields as the regularization term affects only the values of the couplings. Minimizing the cross entropy (Eq. 6) leads to the following expression:

(12)

Therefore,

(13)

Calculate the entropy, Si, for each site as .

2] Now construct clusters of all pairs of sites, and minimize the cross entropy including the regularization term () to obtain new estimates for the fields and couplings. The minimization can be carried out using various methods, such as gradient descent. The starting guess for the fields is the one obtained in step 1, and an independent pair approximation can be made to estimate the initial values of the couplings. Upon convergence of the minimization algorithm, compute the entropy, - , of the two-site clusters by sampling configurations of each cluster using the inferred values of the fields and couplings. Then, compute the change in entropy between each two-site cluster and that obtained by adding up the entropy for the corresponding single sites (obtained in step 1). If the entropy reduction is larger than a threshold value (T), then the particular coupling constant (Jij) and two-residue cluster is kept. Otherwise, the connection is not included (corresponding Jij = 0).

3] Now construct all three-site clusters starting from the connected pairs obtained in step 2, and minimize the cross entropy including the regularization term () to obtain new estimates for the fields and couplings. The starting guesses for the fields and couplings are the ones obtained in step 2, and additional ones for the new couplings. Upon convergence of the minimization algorithm, compute the entropy (-of the three-site clusters by using the newly inferred values of the fields and couplings to sample configurations of each cluster. Compute the change in entropy upon considering three-site clusters compared to the sum of the entropies of all the corresponding connected pairs obtained in step 2. A particular three-site cluster and corresponding couplings are kept only if the entropy reduction is larger than the threshold value (T).

4] Keep growing cluster sizes in this way. For a given value of T, the procedure will end with a disjoint set of clusters that cannot be grown further without an entropy change less than T. Using the fields and couplings thus inferred sample configurations using Monte-Carlo simulations to obtain the one and two-point mutational probabilities. If these probabilities compare well with the observed values estimated from the data, then the procedure has converged. If not, repeat steps 1 – 4 with a lower value of T.

The Los Alamos National Laboratory maintains a database of sequences of various HIV polyproteins and proteins. Multiple sequence alignments (MSA) of these proteins can be downloaded, and the observed one and two-point mutational correlations can be calculated. The protein p24 is part of the Gag polyprotein of HIV, and it forms the viral nucleocapsid. Fig. 4.6 shows a comparison of the mutational correlation functions calculated by sampling sequence configurations according to Eq. 2 using the converged fields and couplings inferred by the ACE algorithm and that observed in the MSA for the p24 protein. Similar levels of agreement can be obtained for other HIV polyproteins. As noted earlier, since there is substantial amino acid variation at sites of HIV polyproteins, a Potts representation is required, rather than the simple Ising representation we used for illustrative purposes.

*Constructing Potts models*

The analyses described above are easily generalized to Potts models, but the numerical cost grows quickly with the number of amino acids that are explicitly considered at each site of the protein. Some simplifications are possible that balance the conflicting demands of accurate representation of the fitness landscape and computational costs. The data for the frequency of amino acids that appear rarely at a given site is noisy because of under sampling. This suggests that it may be appropriate to explicitly represent the more frequently occurring amino acids at a site explicitly while the rarer ones are grouped into a single pseudo amino acid. The choice of which amino acids are grouped together can be done in a systematic way.

Based on the probability of observing amino acid, a, at site, i, (pi (a)) we can rank order the amino acids according to their probability of occurrence, with the most frequent one being ranked 1. Grouping the rare amino acids together, we can write the probabilities of occurrence of amino acids in this representation as:

(14)

Here the amino acids labeled 1 through ki are the ki most frequently observed amino acids, and the others are grouped into one pseudo amino acid which appears with a frequency equal to

(15)

where qi is the number of amino acids observed at site, i, in the sequence data.

The entropy corresponding to amino acid variation at site, i, is: . The entropy after grouping the least frequent amino acids as per Eqs. 14 and 15 is:

(16)

where is the fraction of entropy that is captured by the representation described by Eqs. 14 and 15. The value of ki, and hence, , can be chosen such that the error introduced by Eqs. 14 and 15 is comparable to the error in the sequence data due to under sampling of rare amino acids at a site of the protein.

If there are B sequences, amino acid, a, will be observed n times at site, i, as per the binomial distribution:

(18)

The variance of this binomial distribution, is , and so the error in the data for pi (a) is:

(19)

We wish to choose ki such that the error introduced by grouping the less frequent amino acids in to a single pseudo amino acid is comparable to the variance in the observed sequence data (Eq. 19). The ratio of the error to the variance, given a choice of ki (or ), is denoted as in the formula below:

(20)

We can choose ki such that is roughly unity. Alternatively, we can define a value, , that is averaged over all sites of the protein as follows:

(21)

The values of ki for the different sites can be chosen such that is roughly unity. The values of ki thus chosen will be different for each site. Empirically, it is found that, for example, for the large HIV polyprotein (ENV) setting the fraction of the entropy recovered in the coarse-grained representation in Eqs. 14 and 15 to be roughly 0.95 makes ~ 1.

*4.3.1b: Connection between prevalence and fitness of HIV mutants*

In order to relate the prevalence landscape to the intrinsic fitness landscape we have to consider how the evolutionary history of HIV circulating in the human population while being subjected to human immunity has shaped the prevalence of different virus strains. Eigen considered the evolution of a swarm of species (e.g., virus strains) in the limit of an infinite total virus population. The resulting deterministic equation for the temporal variation of the frequency of each strain is given by [71]:

(22)

Here, is the frequency of strain a, is the rate of mutation from strain a to b, and is the replication rate of strain g (which corresponds to its fitness in our treatment above). The last non-linear term ensures that the strain frequencies, , are normalized. From a biological standpoint, this term embodies competition between strains and so the growth rate of strain a () relative to the average replication rate of the prevailing strains, , is pertinent.

Leuthäusser showed that the non-equilibrium process described by Eq. 22 is isomorphic with the equilibrium statistical mechanics of a two-dimensional Ising magnet [72]. To see how this isomorphism emerges, define the following variable:

Using Eq. 22, the derivative of is given by:

This linear equation can be written in discrete form as follows:

Note that the entries in the vector, re not normalized. We may view to be proportional to the number of progeny of type b in generation 1, given that the number of strains of each type in generation, 0, is given by . The strain frequencies can be obtained as follows:

Each element of the matrix, W, is proportional to a transition probability for going from one strain to another in a single generation, as the following argument shows. Suppose that our starting swarm of viruses in generation 0 is composed of a single strain, a. This is often the case for HIV infection since clinical evidence suggests that usually a single strain establishes infection. If the number of copies of the initial strain is N0, then the numbers of copies of another strain, b, in generation, 1, is given by:

and the frequency or probability of observing strain b in generation 1 is:

So, Eq. 25 describes a Markovian process with the matrix, W, being proportional to the transition probability matrix that does not depend on past history. The Markovian property of the process implies that the probability of observing the vector after n evolutionary generations, given the vector, , , is:

(29)

Each term in the summand in Eq. 29 is proportional to the matrix, W. In other words, repeated operation (multiplication) with the matrix W on yields the probability of occurrence of a particular evolutionary trajectory, with the strain vectors in each generation being, , , , …….. To obtain , we then need to sum over all such evolutionary trajectories (intermediate strain vectors). In other words, we have a path integral formulation.

Fig. 4.7a is a visual depiction of an evolutionary trajectory starting from one strain with a particular sequence (first row) and evolving one other strain in each generation. The sequence depicted in a particular row is the progeny of the sequence above it. The identity of the amino acid at a particular site in a sequence is shown using the Ising representation. A particular configuration of this 2-dimensional “Ising magnet” represents an evolutionary trajectory. The weight of an evolutionary trajectory is the summand in Eq. 29. Since each term in the summand is proportional to W, the probability of a particular trajectory, P (T) is:

(30)

where is the element of W describing a transition from the strain in generation, q, to the one in q+1.

How can we make use of this path integral formulation of evolution of a viral swarm to address how the prevalence landscape of HIV polyproteins is related to the intrinsic fitness landscape that we seek. The evolutionary history of HIV can be described as a very long evolutionary trajectory starting from a founder strain that infected the first patient. The virus evolved for a few generations in a particular patient, and then in another who was infected by this person, and so on. During evolution in each person, the selection forces at play are the virus’ intrinsic fitness and the immune responses of this patient. This information would be contained in W. The prevalence landscape was obtained by sampling sequences from different patients. If we obtained using W for evolutionary trajectories that were very long and had passed through a number of individuals, we could estimate the probabilities of occurrence of different strains in the sequence data if all sequences had been collected from patients after evolutionary trajectories of length n. In reality, samples were collected from diverse patients and the value of n was different in each case. So, averaging over many values of n would yield the prevalence landscape. We could relate how the prevalence landscape thus obtained is influenced by the intrinsic fitness of viral strains and human immunity because W contains this information.

Each element of W describes a transition from one sequence to another, with the diagonal elements representing transitions to the same sequence. Two strains, q and q+1, differ from each other because the amino acids at a number of sites are different. A simple metric of this difference between two sequences is the Hamming distance (dq+1, q), which counts the number of sites that have different amino acids. Mutations can only occur if the virus replicates, and if m is the mutation probability per replication cycle, can be written as follows:

where, as earlier, N is the length of the protein and l is the number of possible amino acids at each site. In the Ising representation that we are using for illustrative purposes, l = 2.

In the Ising representation of sequences in Eq. 2 and all that followed, zi was either 0 or 1. For algebraic simplicity in what follows, we will use an Ising variable si equal to to define . We will make the analyses consistent with Eq. 2 later using . With si = to ,

Using the fact that ax = exp [x ln a], and Eqs. 31 and 32, allows us to express as:

where h and M are defined as follows:

Notice that M is positive for m < 1/2, which is true for any realistic virus.

Using Eq. 30, given a particular founder strain, the probability of observing a particular evolutionary trajectory, T (Fig. 4.7a), is proportional to the product:

The normalized probability of observing a particular trajectory, T, is:

where H (T) is the “Hamiltonian” corresponding to a particular trajectory, and Z is the “partition function” obtained by summing over trajectories. As advertised earlier, in this path integral formulation, the non-equilibrium dynamical problem of virus evolution is transformed in to studying the equilibrium statistical mechanics of a 2-dimensional Ising magnet. Each configuration of the Ising magnet represents an evolutionary trajectory with the weight of an evolutionary trajectory given by Eq. 36 (Fig. 4.7b). Note that the term proportional to M (which is greater than 0) acts like a ferromagnetic nearest-neighbor coupling between “spins” of the Ising magnet across rows. This term promotes the occurrence of the same amino acids at the same site in successive generations, which is a consequence of m < 1/2 (and so M > 0).

In order to relate the prevalence to intrinsic fitness following the approach outlined earlier, we need to write an expression for , which is the effective fitness in generation, q. That is, includes the effects of intrinsic fitness and the immune pressure to which the virus is subjected at that time. The immune pressure acts only on some sites of the virus as antibodies and T cells elicited in a person target only a limited number of sites at any given time. The immune pressure reduces the effective fitness of a strain with a sequence that is the target of the immune response since strains with that sequence are neutralized or the corresponding infected cells are killed. The efficacy of neutralization and killing is determined by the strength of the immune response. Concomitantly, the immune pressure also acts to promote the evolution of mutations at the targeted sites as such mutations would enable immune escape and an increase in the virus’ effective fitness. The extent to which the immune pressure promotes the evolution of mutations depends upon the strength of the response and the effects of “escape” mutations on the intrinsic fitness of the mutated strain. If we take the form of the intrinsic fitness to be as in Eq. 2 and incorporate the effects of immunity as described above, we can write Eq. 36 as follows:

In eq. 37, the strength of the immune response at site i in generation q is denoted by ; the prime indicates that the immune response acts only at a few sites. The can be viewed as external fields that act selectively on some sites (or spins of the Ising magnet). In writing the effects of the immune response as in Eq. 37, we have assumed that depends only on q, and not on effective memory immune responses that were elicited in the past. As we will discuss later, this is not true for influenza and is approximately true for HIV. In writing Eq. 43, we also transformed the spin variables to the {0,1} Ising spin representation.

Eq. 37 allows us to study the relationship between the prevalence landscape and the intrinsic fitness landscape as follows. We could assume that the intrinsic fitness landscape is the one inferred from the prevalence data and then use Eq. 37 to numerically sample trajectories of varying length, n (to mimic that the virus evolved for different times before it was sampled from a patient). In these trajectories, would have to switch periodically to mimic evolution in patients that impose different immune pressures on the virus. The probability of obtaining different strains in generation, n, for different values of n could thus be obtained. By averaging over different values of n, and realizations of we could obtain the prevalence landscape. Comparing this prevalence landscape to the assumed intrinsic fitness landscape could then provide insights into how human immune responses, the finite mutation probabilities (M) and the intrinsic virus fitness landscape shape the prevalence landscape. Such insights can also be obtained more readily from an approximate analysis of Eq. 37, which we now describe.

We can analyze Eq. 37 variationally using the Feynman-Bogoliubov bound. The partition function, Z, can be rewritten as:

(38)

where Hv (T) is a different Hamiltonian, indicates the average of the quantity in angular brackets using the Hamiltonian, Hv, and Zv is the partition function corresponding to Hv (T). Jensen’s inequality now tells us that the average of the exponential of a quantity (a convex function) is greater than or equal to the exponential of the average of the quantity, which when applied to Eq. 38 obtains:

(39)

If Hv is chosen appropriately, the right-hand side of Eq. 39 can be determined exactly. The resulting expression can then be maximized with respect to parameters in Hv to obtain an approximate Hamiltonian that variationally bounds the real Hamiltonian. Analysis of the properties of this variationally optimized Hamiltonian can lead to the insights we seek.

The Hamiltonian corresponding to the prevalence landscape is of quadratic form (Eq. 2), and so we choose Hv to be quadratic form:

where and are in-row couplings and fields, respectively, the optimal values of which are to be obtained using the variational approximation; i.e.:

Solving these equations while taking care to separate the surface layer (n) from the others (since only the probabilities of occurrence of strains in generation n are of interest and interpretable) obtains the following formulas:

In obtaining the first line in Eq. 42, we have assumed that . Since is the average of whether a mutant or wild-type amino acid is present at residue i in generation, n, this approximation should be reasonably good for m < 1 and sufficiently large, n. The details of the algebra in going from Eq 41 to Eq. 42 is left as an exercise for the reader, and is available in one of the suggested readings for this chapter.

Eq. 42 describes how a set of effective fields and couplings, and , depend on the intrinsic fitness landscape (described by {hi} and {Jij}), human immunity (), and mutational fidelity (M) using a variational approximation. In this approximation, only the fields are affected by these effects. However, is not explicitly determined by Eq. 42 because its right-hand side depends on Hv through . Hv cannot be determined without knowing . So, Eq. 42 must be solved self-consistently to obtain , as is typical for such variational, or mean-field, calculations.

However, several key insights in to the relationship between the prevalence and intrinsic fitness landscapes can be obtained without carrying out such a calculation. The parameters, and , describe the prevalence landscape because they can be used to compute the probability of observing different strains sampled from humans. As described earlier, however, we must average the term representing the immune pressure in the expression for in Eq. 42 over the immune fields corresponding to people with diverse HLA types and number of generations, n. Let us denote the value of averaged over different values of n and human HLA types (or types of immune responses) as , and the value of averaged over various generations, n, as . After these steps of averaging, we obtain the following averaged expression for the fields () that determine the prevalence landscape:

Therefore, the Variationally optimized Hamiltonian that incorporates the effects of immune pressure, mutational fidelity, and intrinsic fitness (i.e, the approximate Hamiltonian that can be used to obtain the prevalence landscape) can be written as:

Eq. 44 provides insights in to the relationship between the prevalence and fitness landscapes as well as the underlying biological factors that determine this relationship. Because hi is negative and is positive, Eq. 44 implies that the effect of the immune pressure at a particular protein site is to make the virus more mutable than that due to the intrinsic fitness of the virus. This is a reflection of the fact that the immune pressure imposed at a site promotes mutations because the resulting mutant viral strain is no longer subject to immune attack, thus making the mutant effectively more prevalent. Thus, human immune responses promote the exploration of sequence space by the virus population. How large is this effect? Because of the great diversity of immune (especially T cell) responses in the human population, most regions of the viral proteome are targeted by a small fraction of people; so, acts only for a small number of evolutionary generations corresponding to the fraction of individuals who target residue, i, with their immune responses. Thus, we are led to the conclusion that has a relatively small value when averaged across the human population. That is, human immune responses do not significantly affect the relationship between the prevalence and fitness landscapes.

The conclusion above depends on our earlier assumption that the immune fields acting in a particular generation did not depend on immune fields in a previous generation. Why can persons that were originally infected not be infected again later along the evolutionary trajectory, and mount effective memory responses elicited earlier. This would correlate immune fields effective at controlling a particular viral infection across time, thus significantly complicating our analysis of the evolutionary dynamics. The simplification we made originates in the nature of human immune responses against HIV. Although a small number of people, called elite controllers who usually possessing certain HLAs, target epitopes that enable them to control viral load to low levels [24], most such people do not clear the infection. Furthermore, most people do not control the viral load to low levels with any kind of effective immune response. So, the global population of HIV has largely not been persistently subjected to a few effective classes of natural or vaccine-induced immune responses. Therefore, the global HIV population has not evolved in narrowly directed ways to avoid past effective herd memory responses in the human population. Contrast this situation with that for influenza, which has been subjected to very effective vaccine-induced or natural antibody responses that clear the virus. The resulting effective memory immune responses in the human population continuously drive the evolution of the global influenza population in specific directions to evade such responses. The simplification wherein effective immune fields developed in the past in the human population are assumed to largely not drive evolution of the global HIV population cannot be made for influenza.

How does mutational fidelity (term proportional to M in Eq. 44) affect the relationship between ai and hi? The term, , is expected to be less than one half. This is because the mutation probability is less than half, and also because HIV is a chronic infection that is transmitted from one host to another. If the virus is forced to make an immune evading mutation in a particular individual and this mutation has an intrinsic fitness cost, such deleterious escape mutations can revert over time when the virus is transmitted to a new host (whose immune response likely will not target the same site) [73]. Therefore, in Eq. 44, the term, , corresponding to mutational fidelity adds a negative value to , making it more difficult to observe a mutation at site, i, in the prevalence landscape than that predicted by the intrinsic fitness landscape alone. How large is this effect? If we compare two strains, this term is added to the Hamiltonian for every site for which the amino acids in the two strains are different. The prevalence of two strains that are equally intrinsically fit (same {hi} and {Jij}) could thus be different. The strain that has many more mutations compared to some reference strain will be less prevalent. However, this confounding effect of mutational fidelity is likely to be significant only if we compare strains that differ from each other by many mutations. For strains that differ by just a few mutations from each other, as those that evolve in a single patient over time, this effect is likely to be small. Note also that this effect is further mitigated by recombination of different viral strains during infection and replication.

A way to summarize our conclusions from the analysis exploring the connection between fitness and prevalence landscapes is that, within some range of mutational distances, the HIV population is at steady state, and the rank order of the prevalence of different strains should be statistically in accord with their relative intrinsic fitness. This is decidedly not the case for influenza as it is continuously driven far from equilibrium for reasons that were noted earlier. Of course, in individual hosts HIV evolves to evade host immunity, forcing HIV to adapt and explore the sequence space. If a mutation that evades host immunity comes at a substantial fitness cost to the virus, compensatory mutations often arise to restore lost fitness, and so mutations at these combinations of residues are observed more frequently than by chance in the circulating virus population. Similarly, some combinations of mutations that are especially deleterious for the virus are observed less frequently than by chance. These correlations, which reflect intrinsic viral fitness effects observed because host-pathogen riposte forces the virus to sample sequence space, are reflected in our inferred landscape. Thus, this landscape is expected to describe the collective mutational pathways that HIV can use to evade host immunity and those that it cannot.

4.4: Tests against in vitro measurements and clinical data

The inference procedure for the prevalence landscape is approximate and the sequence data is incomplete. Furthermore, the approximate calculations and biological arguments noted above seem reasonable but could be inaccurate. Therefore, the veracity of our conclusion that the inferred prevalence landscape for HIV proteins is an adequate statistical proxy for the intrinsic fitness can only be established by testing predictions against *in vitro* experiments and clinical data. Let us consider examples of how such tests can be used to test such predictions, beginning with *in vitro* experiments.

*4.4.1: Tests against in vitro measurements of fitness*

The inferred fitness landscape can be used to calculate the value of the Hamiltonian (or “energy”) corresponding to particular mutant sequences relative to a reference sequence. The model would predict that the replicative fitness of the mutant strain relative to that of the reference sequence should correlate negatively with the energy difference between the mutant strain and the reference sequence (Eq. 2). The mutant sequences for which predictions are made can be generated through site-directed mutagenesis, and then their relative fitness can be measured by assaying their growth rates when placed in culture with human cells that HIV can infect. Such experiments are carried out in the absence of human immunity, and so reflect the intrinsic fitness of mutant strains. Fig. 4.8a shows such a comparison between experiments and model predictions for strains of HIV with mutations in the Gag polyprotein, which contains HIV’s structural proteins [46,79]. As is evident, while not perfect, the comparison is reasonably good. The ENV polyproteins comprise the spike on the surface of HIV particles. The fitness landscape of ENV is the hardest to infer because of its long length, much higher variability and occurrence of insertions and deletions compared to other HIV polyproteins. Fig. 4.8b shows a comparison between predictions based on the inferred landscape and roughly 100 *in vitro* measurements of infectivity. Again, the comparison is reasonable.

We can ask whether the accuracy of the predictions would be worse if we inferred a model that does not include the effects of coupling between mutations (i.e, with the fields only). For Gag, the comparison is worse. But, the study with Gag mutants shown in Fig. 4.8a was done as a part of a collaboration between immunologists and the physical scientists who inferred the landscape. They were interested in testing whether the predictions made for epistatic interactions were correct. So, some of the mutant strains that were tested included strains with mutations that were predicted by the inferred landscape to be strongly coupled (as per the magnitude of the corresponding Jij values). The fitness measurements for the ENV mutant strains shown in Fig. 4.8b were conducted independently, and before, the predictions using the fitness landscape. As has been noted before, the matrix, **J**, is sparse because mutations at many residues in a protein are not coupled. As a result, if you take a random set of mutant strains (as for ENV) and compare fitness predictions using landscapes inferred with or without the couplings, the statistical differences are not substantial as many strains do not have mutations at residues that are coupled. But, as we have noted earlier, because of the high mutation and replication rates of HIV *in vivo*, rare sets of mutations in coupled residues could be sampled to affect the virus’ ability to evade human immunity. We will shortly discuss comparisons with *in vivo* clinical data that will make this point vivid. The importance of these effects can also be observed in other types of data as we now discuss.

We anticipate that mutations in sites that are in contact, connected, or in close spatial proximity in the functional structure of a protein or protein complex are more likely to be coupled. A large number of studies have been conducted to study bacterial protein families using methods similar to maximum entropy models. Methods have been developed to determine whether two residues are directly coupled (or in contact) in protein structures based on the **J** matrix. One such method is the Direct Coupling Analysis (DCA), which we do not discuss here. Applying the DCA method to the inferred fitness landscape of the ENV protein can predict residues that are in contact in the viral spike of HIV. This is a prediction that strictly relies on the importance of couplings in the **J** matrix. We can then compare these predictions of the inferred model to crystal structures.

Since the **J** matrix was inferred from virus sequences from patients, we expect that it reflects important contacts in the functionally relevant trimeric spike of HIV, not monomers of the constituent ENV proteins. The native HIV spike is very unstable and so has proven to be difficult to crystallize. A mimic (called SOSIP) has been prepared by stabilizing the trimer with a few disulfide bonds, and its structure is available. So, it may be appropriate to test predicted contacts against those observed in SOSIP’s structure. Fig. 4.9 shows a comparison of the predicted contacts (function of **J**) to those observed in the crystal structure. With high probability, the top twenty pairs of sites as measured by the size of the corresponding inferred values of Jij are correctly predicted to be contacts in the structure. A large fraction of the false positives in this top group of pairs are in particular loops of the trimeric spike or in CD4 contact residues. As noted earlier, upon binding to CD4 on host cells, the ENV proteins that make up the HIV spike undergo conformational changes. Indeed, upon HIV’s spike binding to CD4, it is well-documented that conformational changes occur in one of the loops for which we observe false positives based on structural considerations alone. So, it is possible that the predictions from our inferred landscape reflect this *in vivo* effect that is not captured in the SOSIP crystal structure. These results show the importance of the couplings inferred in the **J** matrix for virus function. Related comparisons have also been described for protein assemblies important for virus function that involve the p24 protein.

As noted earlier, cocktails of antiretroviral drugs that simultaneously target multiple HIV proteins that are important for replication have been very effective in making HIV infection a controllable condition. HIV protease is one of the targets of antiretroviral drug therapy through the action of a class of drugs known as protease inhibitors. The virus is able to evolve mutations that increase its resistance to protease inhibitors alone, but these may be associated with substantial fitness costs to the virus. If so, the drug resistance mutations that are most likely to be observed should be the ones whose fitness costs can be compensated by other mutations. One may then ask whether the inferred fitness landscape can identify such potential drug resistance mutations based on the **J** matrix. When coupling constants corresponding to compensatory interactions exceed a certain value, the strain bearing both the mutation that confers drug resistance and the additional compensatory mutation becomes sufficiently fit. It has been shown that model predictions compare increasingly well with observed drug-resistance mutations as the cut-off value for the coupling constants is increased. These data also suggest that the inferred landscape can capture effects of coupling between mutations on the virus’ intrinsic fitness.

*4.4.2: Tests against clinical data*

Predictions based on the fitness landscape can be tested against clinical data in different ways. Some patients, called elite controllers, can control HIV infections without any therapy by maintaining very low viral loads. These individuals do not progress to AIDS many years after infection, and are less likely to infect other individuals. A number of factors have been posited to explain how elite controllers maintain low viral loads. In genome-wide studies, the strongest correlation is observed with their HLA genes. Some HLAs, such as HLA B57 and HLA B27, are strongly overrepresented in elite controllers. As we have discussed in Chapter 3, a part of the reason is likely the statistically more cross-reactive TCR repertoire of individuals with these genes. Another reason that has been noted by comparing cohorts of controllers and progressors who have the same HLAs is that the activated T cells in controllers exhibit more effective polyfunctional responses. Some studies have also implicated the antibody response in these patients. However, the factor that is implicated most strongly is the kind of peptides targeted by T cells in elite controllers. Mutations in these peptides incur a large fitness cost, and are more likely to be presented by the HLA molecules that are associated with elite controllers.

One can use the inferred fitness landscape to predict the fitness cost incurred upon evolving mutations at a T cell-targeted peptide (epitope). To estimate this quantity in a way that is relevant for diverse patients, it is important to account for the fact that the viral strains in different individuals have different protein sequences. This can be done by calculating the fitness cost incurred by mutations at any site averaged over different sequence backgrounds. Averaging over the sequence backgrounds takes in to account the effects of coupling between mutations. The average fitness cost can be calculated as follows:

(45a)

(45b)

Here is the average difference in energy between the T cell-targeted peptide, , and the same peptide, , with a single non-synonymous mutation, and is the rest of the protein sequence containing this peptide. The average over all possible sequence backgrounds, , in Eq. 45a can be calculated using standard Monte-Carlo simulations with the fitness landscape (Eq. 2). The fitness cost thus obtained accounts for the effect of a mutation by itself as well as the effects of coupling between mutations. Eq. 45b computes an average cost of evolving a mutation in the epitope, averaged over all possible mutations therein. This way of calculating the latter average emphasizes the contributions of mutations with the lowest fitness costs, as they are the most likely to evolve. Using this metric of the fitness cost, it was found that elite controllers target peptides associated with high fitness costs for evolving mutations in all possible sequence backgrounds. Thus, such escape mutations do not emerge rapidly and the individual’s T cell response can control the virus to low levels. With a vaccine-induced T cell response, one would like to elicit such responses that can corner the virus between being killed by T cells or evolving mutations that are likely to make the virus unviable.

One can also test predictions of the fitness landscape against data from patients in whom the virus evades their immune responses via mutation, and so they do not succeed in controlling the virus. Let us discuss one such example. In a cohort of patients from the US, Malawi, and South Africa, the epitopes targeted by the killer T cells (CTLs) of individual patients during a time frame spanning a range from the first detection of virus to shortly after the viral load peaked were identified (as described below). In these patients, the virus evaded the predominant CTL responses directed toward these epitopes via escape mutations. The time required for these mutations to evolve and their locations were measured or estimated from measurements (see below).

The peptide epitopes targeted by CTL responses are usually detected using an assay called ELISPOT. Peptides comprised of overlapping sites derived from a viral protein are presented on human APCs and displayed in different wells of the device. Blood samples from patients contain T cells. These T cells secrete cytokines, in particular IFNg, when they interact with their cognate peptide. One counts the number of “spots” of these cytokines that are secreted in individual wells. The number of spots thus detected corresponds to the extent to which a particular epitope is being targeted in the patient. In this manner, one can determine which epitopes are being targeted by a patient’s T cells, as well as the relative immunodominance of the epitopes being targeted. Epitopes targeted more dominantly are under stronger selection pressure to evolve mutations.

The time taken for a particular escape mutation to take over the population of viruses (escape time) in a patient can be estimated if blood samples are collected at different times. Often, the samples are not collected at regularly spaced intervals, and the escape time could correspond to a time point at which a sample was not collected. One way to estimate the escape time from such data is as follows. Let us define

(46)

where ni is the number of viral sequences observed at time point, i, and ki is the number of these sequences with an escape mutation at a targeted epitope. One can then use a logistic form for the fraction of sequences, f (t), with an escape mutation at time, t:

Here f0 and e are parameters that reflect the fraction of sequences with an escape mutation at time zero and the rate at which the logistic equation saturates to unity, respectively. One may interpret f (t) to be a probability of observing an escape mutation at time, t. Therefore, one can write the following expression for the likelihood (L) of the observed data (Eq. 46):

Maximizing L with respect to the parameters, f0 and e, provides an estimate for them based on the data. These values of f0 and e can then be used in Eq. 47 to determine the time point at which half the viral population is comprised of a sequence with an escape mutation in the targeted epitope (f (tescape) = 0.5). Data sets like these, which provide information on the epitopes targeted by T cells in individual humans and estimates of the locations of escape mutations and the escape times offer several opportunities to further test inferred fitness landscapes.

For example, the data noted above provide examples of how the time to evolve a particular escape mutation is influenced by the rest of the sequence of the corresponding protein. Different patients are infected with viruses bearing different mutations. Comparing the time to evolve the same escape mutation in pairs of patients that target the same peptide may thus provide insights into the in vivo importance of the interactions embodied in the **J** matrix. Fig. 4.10 (A and B) illustrates one such example. Two patients, labeled CH185 and CH159, target the same epitope in Gag. In CH185, the escape mutation evolved in 122 days after the T cell response was first detected, while in CH159 the escape mutation had not evolved in over 1100 days. Fig. 13 shows that this is because of differences in the sequences of the viruses that infected these individuals. The circles in the figure depict the rest of the protein that contains the targeted epitope. The marked residues are those where mutations existed in the virus that infected the patient. Blue curves indicate that the fitness landscape predicts that the element of the **J** matrix corresponding to the preexisting mutation and the ultimate escape mutation is positive (compensatory). Red curves indicate negative values of the corresponding element of the **J** matrix (antagonistic interactions). The thicknesses of the lines reflect the relative magnitudes of the predicted J-couplings. Patient CH159 was infected with a viral strain that, compared to the infecting strain in CH185, contained more mutations that coupled negatively (and strongly) to the putative escape mutation. Therefore, it was much more difficult for the escape mutation to evolve and grow out in patient CH159. Fig. 4.10 C shows other examples of such a situation in this cohort of patients, further highlighting the importance of the effects of couplings between mutations for host-pathogen riposte in individual humans. The values of reported in Fig. 4.10 C were computed using Eq. 45.

Eq. 45 reflects the shortest path to evolving an escape mutation and does not account for the dynamics of the evolution of the virus via complex trajectories in response to T cell pressure. As one example of phenomena not captured by Eq. 45, if a larger number of feasible trajectories are available for an escape mutation to emerge at a particular site or epitope, escape is facilitated because the “entropy” of escape trajectories is higher. The importance of such effects and the veracity of fitness landscapes can be tested by attempting to predict the dynamics of virus evolution in individual patients in the cohort described above. Specifically, one can ask if it is possible to predict the sites at which escape mutations emerged, and the relative times required for this to happen in these patients by combining computer simulations of the evolutionary dynamics with the inferred fitness landscapes and knowledge of the targeted epitopes (and their relative immunodominance) [82].

The evolutionary dynamics within each person is carried out using a method similar in spirit to Wright-Fisher dynamics in population genetics. In these dynamics, the virus population evolves in discrete time steps. At each time step, the virus can mutate and replicate in a fitness-dependent way. Mutations are carried out at the nucleotide level. This is important because mutation rates are known for nucleotides, not amino acids, and because this is the only way that allows proper treatment of mutational paths (transitions between certain amino acids are simply not possible in one step). The average rate of mutation for HIV is 3 \* 10-5 per base pair per replication step.

In addition to mutation, recombination of viral genomes can also occur if two different RNA strands (or viral genomes) are simultaneously present in an infected cell. Recombination can be simulated during the evolutionary dynamics in a simple way. Let r be the rate of recombination events per base pair per replication cycle (a quantity that has been estimated for HIV). During replication of each sequence, we pick the number of recombination sites (n) according to the binomial probability distribution, , where N is the length of the protein or genome under consideration, and is the number of ways of choosing n locations out of N. If a finite number of recombination sites results, the specific locations of the recombination sites are chosen from a random uniform distribution. A partner sequence is then chosen from those in the evolving swarm from a uniform random distribution. As an example of how a recombined sequence is created next, consider the case where there are two recombination sites, i and j. The recombined sequence has the original sequence between residues 1 to i, the sequence of the randomly picked other sequence from i+1 to j, and the original sequence from j+1 to L. This process mimics Reverse Transcriptase falling off of one RNA strand, hopping to another, and then returning. This is possible because RT is not a very processive enzyme.

After attempting mutations at each site of every sequence that is present at a particular time step and possible recombination events, the resulting sequences undergo fitness-dependent replication. The fitness of each sequence is determined by the corresponding value of energy, . As in Eq. 37, the energy depends upon the sequence and the immune pressure directed toward an epitope contained in the sequence. We assume that any nonsynonymous mutation within a targeted epitope is sufficient to allow the virus to avoid detection by epitope-specific T cells. Though this assumption is not always correct, there is experimental evidence that most mutations within an epitope tend to substantially impair T cell recognition as we have discussed in Chapter 3 [83]. A simple way to model the abrogation of immune pressure due to a mutation is to increase E (**z**) by a fixed value, b. As noted earlier, the available data contains information on the relative immunodominance of the epitopes in each patient. This information can be incorporated in to simulations by writing the following expression for b:

where %I is the relative immunodominance (varies between 0 and 1), bmin and bmax are the minimum and maximum values of the penalty in fitness due to immune pressure, respectively. As long as the value of bmin is chosen to be greater than the largest escape cost for possible escape mutations, the qualitative results are the same. That is, evolving mutations in epitopes targeted by T cells must confer a selective advantage. For the results we will show, bmin = 9 satisfies this condition, and bmax was be chosen to be twice the value of bmin. Modest variations in bmax do not change the qualitative results.

To model fitness-dependent selection, the survival probability of each sequence present after the mutation and recombination step is calculated as follows:

(50)

where is the average energy of the sequences prevailing at the current time step and E (**z**) includes the effects of intrinsic fitness and immune pressure. The value of E (**z**) used in Eq. 50 must reflect the intrinsic fitness of the virus strain, **z**, as immune pressure is explicitly accounted for in the simulations (fixed value of b calculated according to Eq. 49 is added to E (**z**)). So, we have introduced a quantity analogous to the “inverse temperature” in statistical physics ( = 1/kBT, where kB is Boltzmann’s constant and T is temperature). This parameter can be obtained by graphing the logarithm of *in vitro* replicative capacity measurements (Fig. 4.8) against values of E (**z**) obtained from the inference procedure (Eq. 2). The value of thus obtained for Gag mutants, for example, equals 0.07 [79].

It is worth reflecting on why < 1. The intrinsic fitness of a strain, without immune pressure or time to evolve mutations, is measured in the *in vitro* experiments. Let us denote the energies corresponding to the intrinsic fitness of a strain and its prevalence by E’ (**z**) and E (**z**), respectively. So, the prevalence, P (**z**), and the intrinsic fitness, f (**z**), are given by:

(51)

E (**z**) contains the effects of immune pressure, mutational fidelity, reversion of the mutations induced by immune pressure in one host, and other effects discussed in section 4.3.1b. As noted there, averaged across the population of hosts, the immune fields acting at HIV protein sites are expected to be small because of the diversity of human immune responses. Reversion of mutations induced by a host’s immune response when infection propagates to hosts that do not target the same epitopes, the contributions due to mutational fidelity, and recombination are all forces that drive the circulating HIV population to remain closer to the fittest strains than expectations based on intrinsic fitness considerations alone. Thus, one might posit that P (**z**) samples the distribution of E (**z**) at a lower temperature (it samples lower fitness strains less) compared to that expected based on intrinsic fitness considerations alone. In other words, to obtain f (**z**) we must sample the distribution of E (**z**) at a higher temperature than kBT = 1, which was used to infer P (**z**) in Eq. 2. That is equivalent to saying that E’ (**z**) = E (**z**) with < 1.

In the spirit of Wright-Fisher dynamics, let us assume that the size of the population of viruses in an individual is fixed. This is not a very good approximation for some of the cases in the data being considered because the virus population size is still rising if escape occurs before peak viremia (Fig. 4.4). However, let us examine the results that emerge from such approximate calculations. The population size that was used for the results we describe [82] is 104 virus particles, which corresponds to the effective population size for HIV populations estimated from data in individual patients. The effective population size can be thought of as the population size of a Wright-Fisher model that results in the same extent of genetic drift as the real system (whose population is different). To implement a fixed population size in the simulations, after surviving sequences are obtained using Eq. 50 the population size is restored by picking sequences from this set with replacement.

The inputs to the Wright-Fisher simulations are the fraction of each sequence in the virus population when the T cell response was first detected. If sequence data was not available at the time point when the T cell response was first detected, the most recent recorded sequences are used. Given this input, the simulations are carried out, and escape is signaled by sequences with a particular escape mutation exceeding half the prevailing population of viral strains.

In the evolutionary dynamics, the steps of replication, mutation/recombination, and selection occur in a single step. However, biologically these steps are separate and involve different time scales. Mutations occur rapidly during reverse transcription of the viral RNA into DNA and during replication. Selection effectively operates at the level of infected cells, which may or may not successfully produce new viruses and which may be killed by cytotoxic T cells. The typical lifetime of infected cells is around 2 days when productively infected [20]. Also, the simulations are carried out in the strong selection regime as the value of b is set to be sufficiently large that escape is favored even for epitopes where fitness costs are large. For these reasons, it is difficult to precisely connect generations of evolution in the simulations to real time. The predicted generations of evolution required for escape mutations to emerge reflect relative rates at which escape occurs at different epitopes. Let us see how the predictions of the evolutionary simulations which invoke many approximations compare with the clinical data on the sites where escape mutations evolve and the escape time.

Model predictions for the most likely and second most likely locations for escape mutations matched the clinical data for escape mutations that actually emerged in roughly 86% of the cases. Figure 4.11 shows that the clinically measured escape times correlate reasonably well with the predicted evolutionary generations (Spearman correlation of 0.73). The latter quantities were calculated by averaging over many simulations for the same epitope. There is considerable scatter, especially for cases where escape is neither very fast or slow. However, the simulation results accurately separate the epitopes wherein escape is rapid and those where escape mutations take long to evolve. The latter types of epitopes should be the targets of vaccine-induced T cell responses.

Two other points are noteworthy. The first is that if we only use Shannon entropy of the epitopes to predict relative escape times, then the Spearman correlation with the clinical data is only -0.22 (epitopes associated with higher entropy escape faster). This shows that the effects of coupling between mutations is important. That is, the nature of the epitope targeted, as characterized by the fitness costs of evolving mutations therein in diverse sequence backgrounds and the number of possible escape paths with low fitness cost that are available for an epitope, are important determinants of escape time. The other important determinant is the selection pressure imposed by the T cell response to evolve mutations. The higher the immunodominance (value of %I in Eq. 49), the higher the selection pressure. This is made clear by considering simulations where the relative immunodominance (values of %I) are not taken into account. Then the predicted times for escape correlate with the clinical data with a lower Spearman correlation of 0.53. In fact, if only the values of %I are used to predict escape times, a Spearman correlation of -0.53 is observed. The results of evolutionary dynamics taking both the fitness characteristics of the targeted epitopes and immunodominance into account lead to the highest correlation between computational and clinical results (Spearman correlation of 0.73).

A question for future exploration is to determine the extent to which the discrepancies between clinical data and predictions can be ascribed to different approximations made in the simulations of evolutionary dynamics or errors in statistical inference of the fitness landscape. For example, is the comparison with data better if we do not make the approximation of constant population size, or do not encapsulate mutation/replication and selection in one effective time step (by treating infection of new cells as a separate step)? Also, in the studies described, the evolutionary dynamics of escape was considered only one epitope at a time. In reality, multiple epitopes are simultaneously targeted in a single individual.

4.5 Concluding remarks

With a validated fitness landscape, one can work toward designing immunogens that may be able to elicit potent T cell responses that target mutational vulnerabilities of highly variable pathogens. Toward this end, Monte-Carlo simulations can be used along with Eq. 45 to determine regions of the viral proteome wherein mutations incur a large fitness penalty in all possible sequence backgrounds. Only these parts of the proteome would be then included in the immunogen. A major engineering challenge here is devising carriers that can efficiently deliver long peptide immunogens in a way that leads to a significant T cell response, an issue that also confronts the development of cancer vaccines. This topic is beyond the scope of this book, but we note that the Adenovirus vectors and mRNA platforms that proved to be successful for vaccines that were developed to protect against SARS-CoV-2 offer promise in this regard.

We mentioned that inferring the fitness landscape of influenza proteins from sequence data is far more challenging because it is continually evolving away from effective population-wide immune responses. Overcoming this challenge will be important for other such current and future pathogens. Analyses of protein structures can also yield interesting information on mutational vulnerabilities. But, structures of single proteins may not reveal various functional constraints that go beyond structure, and also not reflect constraints that arise because of the importance of forming multi-protein functional structures. Perhaps a combination of sequence and structural analyses, with explicit account for the temporal variation of human immunity targeting a pathogen could be developed as a general method to determine fitness landscapes of diverse mutable pathogens. Such a method could become an invaluable aid to vaccine design.

**Appendix**

Connection between statistical mechanics and information theory

In this appendix, we briefly describe the connection between statistical mechanics and information theory that led us to define the least biased model for the prevalence of HIV strains as in Eq. 2.

*A.1: Entropy in statistical mechanics*

We begin with some brief reminders of elementary statistical mechanics. Consider a macroscopic system with a fixed number of particles, volume, and energy (denoted by N, V, and E, respectively). The use of N for the number of particles is standard in most books on statistical mechanics, and is not to be confused with the protein sequence length for which the symbol N was used in the main text of this chapter. There is an ensemble of microscopic states consistent with the macroscopic state with fixed N, V, and E, and this ensemble is referred to as the microcanonical ensemble. All microscopic states in this ensemble are equally likely, and so, the probability, P, of being in any one of the microstates is:

(A1)

where is the total number of microstates in the ensemble. A basic tenet of statistical mechanics proposed by Boltzmann is that the entropy, which is a measure of disorder or uncertainty in the system, equals in the microcanonical ensemble, where kB is a constant named after Boltzmann.

It is more convenient to control and measure temperature (T) instead of the energy of an experimental system. Note that temperature is an intensive variable, which is not a function of system size, while energy grows with system size. The ensemble of microscopic states consistent with the macroscopic constraints of N, V, and T is referred to as the canonical ensemble. There is no reason to believe that the microstates in the canonical ensemble are equally likely as in the microcanonical ensemble, and indeed their probabilities are different as shown below.

A convenient way to keep a system at constant temperature is to surround it with a conducting wall and immerse it in a large bath. The system exchanges energy with the bath, and so its energy fluctuates, but this keeps the temperature of the system constant and equal to that of the bath. As the large bath is insulated from its surroundings, together the bath and the system are characterized by constant values of N, V, and E. Denoting the bath by B, the system with , and the system and bath together as , the following relationship holds:

(A2)

where is the energy of the system, is the energy of the bath, and is a constant. Now, suppose the system is in a particular microstate with energy, , then the energy of the bath is given by . There is only one way for the system to be in a particular microstate, and so the total number of microstates for the bath and system equals 1 \* , where is the number of microstates available to the bath when the system is in the microstate with energy . Since the system and bath together is characterized by constant N, V, and E (equal to ), the microstates of the system and bath taken together are equally likely. Therefore, the probability of the system being in a particular microstate with energy is given by dividing the number of microstates available to the system in this situation divided by the total number of possible microstates of ; i.e.,

(A3)

The denominator on the right-hand side of Eq. A3 is a constant, and so ~ . To keep the temperature of the system constant, it must be much smaller than the bath (or the bath and system taken together). So, we can carry out a Taylor expansion in powers of , and rewrite as follows:

(A4)

where in the last equality in A4, we have used Boltzmann’s definition of entropy in the microcanonical ensemble. Now standard classical thermodynamics tells us that , where is the temperature of the bath, which is equal to the temperature of the system, T. Therefore, we conclude that

(A5)

where Z is the partition function that normalizes the probability distribution.

What is the entropy in the canonical ensemble? Classical thermodynamics tells us that the equilibrium state of a system with fixed, N,V and T (canonical ensemble) is the one for which the Helmholtz free energy (A) is minimal. The Helmholtz free energy is given by . Note that the energy of a system in the canonical ensemble is the value of the energy averaged over microstates,. Using the definitions of E and A, and Eq. A5, it is easy to show that

(A6)

where is , an inverse temperature. Simple algebra using Eq. A6 and the definition of A yields that . Now substituting formulas obtained above for A and E into the relationship, , obtains

in “entropy units” (A7)

Notice that we obtained expressions for the pertinent quantities in the canonical ensemble by considering a system that overall (bath plus system) looked like a system in the microcanonical ensemble.

*A.2 Information and entropy*

In statistical physics, entropy is a measure of the uncertainty in the knowledge regarding the microstates in the system. The larger the number of possible microstates consistent with macroscopic constraints, the greater the uncertainty and the higher is the entropy. Now consider a question that has several possible answers. When we learn the answer, we gain information as there is no longer any uncertainty. The larger the number of possible answers, the higher the uncertainty. So, the information gained is also higher if there are more possible answers to the question. Intuitively, these statements suggest a connection between entropy and information gained. The precise connection between the concepts of information and entropy was first described by Shannon, and many fine expositions are available. A clear description in the context of biophysics can be found in the book by Bialek noted in the suggested readings, and the much shorter description below is inspired by it.

As noted above, the information gained should increase monotonically with the number of possible answers, N, to a question. Consider the following question: what is your name, and where do you live? This question has two independent parts as a person’s name does not depend upon where they live. The information gained by answering a question which has independent parts must be the sum of the information gained by knowing the answer to each independent part. If we associate a probability, pi, with each possible answer, i, being correct, Shannon showed that entropy is the only function of the set, {pi}, that is consistent with the two properties of information gained noted above. A rough proof is sketched below.

First, let us consider a situation analogous to the microcanonical ensemble in statistical mechanics – viz., one where there are N equally likely possible answers to a question. Furthermore, consider the general situation where m independent questions need to be answered, each with k possible answers. Therefore, N = km. Since the information gained by answering each independent part must add, the information gained, I (N), must obey the following relationship:

(A8)

Now consider a situation where n independent questions need to be answered, each with l possible answers, with the values of l and n such that the following condition holds:

(A9)

Since the information gained must be a monotonically increasing function of the total number of possible answers,

(A10)

If the information gained by knowing the answer grew with N as I (N) ~ ln (N), both Eqs A8 and A10 would be satisfied. This is because the logarithm is a monotonically increasing function, and the ln (ax) is x ln a. This suggests that, if there are N equally likely answers to a question, the information gained by knowing the answer is ~ log N. Notice that the entropy of a system in the microcanonical ensemble is also proportional to the logarithm of the number of possible microstates.

Now consider the more general situation, where there are still N possible answers, but they are not equally likely. The probability that the ith answer is correct is pi. The quantity we seek is the information, I ({pi}), gained by knowing the answer when the possible answers are characterized by a set, {pi}. To determine the properties of a system characterized by the canonical ensemble of microstates, which have unequal probabilities, we considered the system to be immersed in a large bath and together the system and bath comprised a microcanonical ensemble of states. We could then use the known properties of the microcanonical ensemble to help us determine the properties of the canonical ensemble. In a similar spirit, let us consider a system where the answers are equally likely to be correct. However, the possible answers are divided into groups, with each group having a different number of these equally likely answers. Because all answers are equally likely, pi is given by

(A12)

where the ith group contains ki of the equally likely answers, and the total number of equally likely answers is N = , where the sum is over the m groups.

In the way that we have formulated the problem, I (N) must be the information gained by knowing which group the answer belongs to, and which of the equally likely answers in that group is the answer. The answers to these two questions are independent. I (N) must therefore be equal to the sum of I ({pi}), and the information gained by knowing which of the ki equally likely answers in that group was obtained. Since the ki answers are equally likely, the latter quantity ~ ln ki. Furthermore, we must weight ln ki by the probability, pi, as each group must contribute as per its probability. Thus, we conclude that:

(A13)

Therefore,

(A14)

The expression for the information gained, given that the possible answers are characterized by the set, {pi}, is exactly the same as the entropy in the canonical ensemble where different microstates occur with different probabilities (compare Eqs. A7 and A14).

In the text of this chapter, we sought the least biased model for the probability distribution characterizing HIV protein strains in circulation (analog of {pi}). The least biased model is one where the answer is most uncertain, and so the information gained by knowing the answer should be maximal. Eq. A14 says that this is tantamount to seeking the probability distribution characterized by the maximal entropy.