Maximum entropy models for antibody diversity

Thierry Mora[a,1], Aleksandra M. Walczak[a,b,1], William Bialek[a,b], and Curtis G. Callan, Jr.[a,b,2]

[a]Joseph Henry Laboratories of Physics and Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544; and [b]Princeton Center for Theoretical Science, Princeton University, Princeton, NJ 08544

Contributed by Curtis G. Callan, February 10, 2010 (sent for review January 2, 2010)

Recognition of pathogens relies on families of proteins showing great diversity. Here we construct maximum entropy models of the sequence repertoire, building on recent experiments that provide a nearly exhaustive sampling of the IgM sequences in zebrafish. These models are based solely on pairwise correlations between residue positions but correctly capture the higher order statistical properties of the repertoire. By exploiting the interpretation of these models as statistical physics problems, we make several predictions for the collective properties of the sequence ensemble: The distribution of sequences obeys Zipf's law, the repertoire decomposes into several clusters, and there is a massive restriction of diversity because of the correlations. These predictions are completely inconsistent with models in which amino acid substitutions are made independently at each site and are in good agreement with the data. Our results suggest that antibody diversity is not limited by the sequences encoded in the genome and may reflect rapid adaptation to antigenic challenges. This approach should be applicable to the study of the global properties of other protein families.

D regions | immune receptor proteins | statistical models

The number of possible amino acid sequences exceeds the number of individual protein molecules that have ever been synthesized. As a result, the limited set of sequences that we see today carries a signature of evolutionary history (1). But not all of the limitations are historical—randomly chosen sequences will not fold into stable, compact structures (2, 3), and carrying out specific functions places yet more requirements on the sequence. Regardless of the balance between historical and functional constraints, the stochastic nature of evolutionary change means that the sequences we observe should be thought of as being drawn out of a probability distribution. The goal of this paper is to construct an approximation to this distribution, by using a limited but biologically important example, the problem of antibody diversity.

The ensemble of *all* proteins is daunting, so most work focuses on particular families of proteins. The most tractable examples are those in which the relevant segments of the proteins are short, and experiments provide many independent samples of sequences from the family. For a family of small proteins that mediate protein–protein interactions, methods were developed to generate artificial sequences that are consistent with the patterns of single site substitutions and correlations between substitutions at pairs of sites; remarkably, most of these artificial sequences fold into functional structures (4, 5). Although this work did not lead to an explicit construction of the underlying probability distribution, the implicit model is equivalent to a maximum entropy model that captures pairwise correlations but ignores higher order interactions (6) and thus connects to other efforts to describe biological networks with simplified models (7–12). Maximum entropy methods have since been used to look at protein–protein interactions in bacterial signaling (13) and at the serine proteases (14).

A key feature of the maximum entropy approach is its intimate connection to statistical mechanics (15, 16). Maximum entropy models predict the underlying probabilities in the form of a Boltzmann distribution, thus assigning an effective energy to every amino acid sequence in our ensemble. Natural questions about this statistical mechanics problem have clear biological correlates: What is the entropy in sequence space or, equivalently, the allowed diversity of functional proteins? Does the energy landscape break up into multiple valleys, corresponding to clusters of closely related proteins? Are the barriers between these valleys large, so that different clusters are isolated, or are there paths that can smoothly mutate one class of sequences into another? Are the interactions among substitutions at different sites strong or weak? Is it possible that these interactions are tuned to some special values, perhaps analogous to critical points in statistical mechanics? Here we approach these problems in the context of antibody diversity.

For antibodies, sequence diversity has a direct biological function, setting the range of antigenic challenges to which the organism can respond. Classical work has emphasized the combinatorial diversity generated by piecing together different segments of the antibody molecule, each of which is encoded in the genome (17). Very recently, it has become possible to provide the sequences of essentially every single antibody molecule in individual organisms (18), and this explosion of data invites us to look more closely at the diversity within the combined segments, beyond that represented in the genome itself. As we will see, for the zebrafish studied in ref. 18, this nongenomic diversity is substantial and concentrates in short segments of the molecule, the D regions of these molecules. This combination of focus on short sequences and a nearly complete sampling of the relevant ensemble provides a unique opportunity to address the theoretical questions outlined above.

Defining the Problem

All jawed vertebrates are endowed with an adaptive immune system that responds to and "remembers" a wide range of challenges from the environment. One major component of the immune system are the B cells, each of which expresses multiple copies of a single antibody molecule on its surface. Binding to these molecules is the fundamental step by which the system recognizes an antigen, and hence the diversity of these molecules defines the range of pathogens to which the organism can respond effectively (19). During the development of B cells, the genome is modified by recombination to encode a single antibody sequence assembled from three pieces termed V, D, and J. In the zebrafish (20), there are 39 choices for the V region, 5 for D, and 5 for J, for a total of 975 possible VDJ combinations or classes. During recombination, nongenomic nucleotides are randomly added and others are removed at the VD and DJ junctions, generating what is called junctional diversity. Furthermore, during the lifetime of the organism, the antibody sequences encoded in proliferating B cells undergo somatic hypermutation. Finally, B cells that successfully bind pathogens proliferate, whereas B cells that

Author contributions: T.M., A.M.W., W.B., and C.G.C. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

[1]T.M. and A.M.W. contributed equally to this work

[2]To whom correspondence may be addressed. E-mail: ccallan@princeton.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/1001705107/DCSupplemental.

are not used are eliminated. As a result, the expressed repertoire of antibodies is a complex combination of VDJ class, phylogenic history, and pathogen environment.

The experiments of ref. 18 give us a snapshot of the complete antibody repertoire in each of fourteen zebrafish, labeled A–N. More precisely, these experiments extracted the mRNA for the complementarity determining region 3 (CDR3) of the heavy chain of IgM molecules and reverse transcribed, amplified, and then sequenced the resulting cDNA by using high throughput methods. It will be important in our analysis that the amplification step has biases, and so all averages over the distribution of sequences must be reweighted by a primer-dependent amplification, as discussed in ref. 18 (*SI Text*). Each fish yielded from 28,000 to 112,000 sequence reads of ∼200 nucleotides covering the last 90 nucleotides of V and all of D and J.

The V and J segments of all the sequences are easily recognized by aligning with the genome, discarding a small fraction of sequences with stop codons or frame mismatches. The situation for D regions is more subtle, and so we define the D region to be all the residues that lie between the identifiable parts of the V and J segments, as explained more fully in *SI Text*.

We find that the D region is much more diverse than expected from its genomic origin and concentrates most of the nongenomic diversity, as illustrated in Fig. S1. Most obviously, in the genome D regions range from 11 to 14 nucleotides, whereas in the sampled sequences the D regions range from 1 to 6 amino acids (3 to 18 nucleotides; Fig. S1*A*). If we try to match each sequence to one of the genomic sequences, the quality of these assignments typically is quite poor (Fig. S1*B*). By using mutual information between residue positions as a measure of variability within VDJ classes (see *SI Text*), we find that residues in the D region are both variable and correlated even within a given D class, whereas the V and J regions show very little diversity within their classes (Fig. S1*C*). Junctional diversity, somatic hypermutations, or other mechanisms may be the source of this nongenomic D variability and could explain the poor quality of the D assignments. Independent of the mechanism, these results suggest that, in trying to define the distribution of sequences represented in the system, we should focus our attention on the D region.

To be precise, we describe each observed D region sequence as $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_L)$, where L is the length of the sequence. At each site along the sequence, σ_i can take on 20 different values, corresponding to the 20 possible amino acids ($\sigma_i = \mathrm{Ala}, \mathrm{Arg}, \mathrm{Asn}, \ldots$). We would like to know the probability $P(\boldsymbol{\sigma})$ that any particular sequence will be found in the antibody repertoire of each individual. The difficulty is that there are $\sim(20)^{L_{\max}}$ possible sequences, where $L_{\max} = 8$ is the maximum length of the D region; in principle, each sequence can occur with a different probability, and hence the number of possible sequences is also the number of parameters required to specify the distribution. This number, $\sim 2.5 \times 10^{10}$, is much larger than the number of independent measurements that we can make and perhaps even larger than the number of B cells in the entire zebrafish at any one moment. How, then, can we make progress?

Maximum Entropy Models of the D Region

Whereas experiments cannot characterize the entire distribution $P(\boldsymbol{\sigma})$, it is possible to make reliable measurements of many averages over this distribution. For example, we can characterize the probability that any single amino acid appears in the sequence, $P_1(\sigma)$. Further, we can characterize the probability that two particular amino acids appear separated by a distance k along the sequence, $P_2(\sigma, \sigma'; k)$, and we can do this for nearest neighbors ($k = 1$), next-nearest neighbors ($k = 2$), and so on. Notice that these quantities do not refer to specific sites along the sequence but rather to pairs of sites separated by given distances; in this way, we can analyze sequences that have variable lengths and are difficult to align, as observed for the D regions. We could continue along this line, characterizing the probability of occurrence of triplets, quartets, etc., but at some point we will run out of data.

The central idea of maximum entropy models is to take some limited set of averages seriously as a characterization of the system and then build the least structured model for the distribution $P(\boldsymbol{\sigma})$ that is consistent with these data (15, 16). Formally, minimizing structure means maximizing the entropy

$$S[P] = -\sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma}) \log_2[P(\boldsymbol{\sigma})]. \quad [1]$$

Here we will find the maximum entropy distribution consistent with the single residue frequencies, $P_1(\sigma)$, with the pairwise distributions of amino acids along the sequence, $P_2(\sigma, \sigma'; k)$, and with the observed distribution of lengths of the D region, $P(L)$. Finding this model distribution, which we denote $P^{(m)}$, involves solving an optimization problem (maximize S) subject to constraints (the observed distributions). Because of the connection between maximum entropy distributions and statistical mechanics, the form of the solution is well known.

We can write $P^{(m)}$ in the form of the Boltzmann distribution, as if the sequences represented the state of a physical system in thermal equilibrium:

$$P^{(m)} = \frac{1}{Z} \exp[-E(\boldsymbol{\sigma})], \quad [2]$$

where the effective energy of each sequence is

$$E(\boldsymbol{\sigma}) = -\mu(L) - \sum_{i=1}^{L} h(\sigma_i) - \sum_{k=1}^{K} \sum_{\substack{i,j \\ i-j=k}} J_k(\sigma_i, \sigma_j). \quad [3]$$

To complete the analogy to thermodynamics, we should think of the temperature as being such that $k_B T = 1$. Then $\mu(L)$ acts like a chemical potential for adding residues, $h(\sigma)$ is a uniform biasing field that prefers some amino acids over others, and the couplings J_k describe the interactions between amino acids at different sites, reaching across a range K, as schematized in Fig. 1*A*. The hs, Js, and μs must be chosen such that $P^{(m)}(L)$, $P_1^{(m)}$, and $P_2^{(m)}$ agree with the data.

Calculating $P^{(m)}(L)$, $P_1^{(m)}$, and $P_2^{(m)}$ from the full distribution $P^{(m)}(\boldsymbol{\sigma})$ is hard in general, and the inverse problem of inferring the model parameters from these observables is clearly not easier. We solve the inverse problem by combining Monte Carlo simulations with gradient descent (see *SI Text*). The number of parameters can be fairly large, $399K + 19 + L_{\max} \sim 10^3$, although vastly smaller than the number of possible parameters $(20)^{L_{\max}}$. To test the validity of our method and control for overfitting, we learned the maximum entropy distribution from only half of the sequences (training set). Then the model predictions were compared to the second half of the data (testing set). We solved the inverse problem and tested our solution for all 14 fish and for different interaction ranges $K = 1, 2, 3, 4$. Our results showed excellent agreement with the data, as illustrated in Fig. 1*B* for the pairwise frequencies in fish A.

Testing and Exploring the Model

The maximum entropy model is the least structured model consistent with the observed pairwise correlations among amino acids, but of course there is no guarantee that nature is described by this minimal model. To test the model, we look systematically at its predictions for measurable quantities that are not already used in determining the model parameters. If we can convince ourselves that these predictions are at least approximately correct, we can take the model more seriously and ask what it tells us about the nature of antibody diversity.

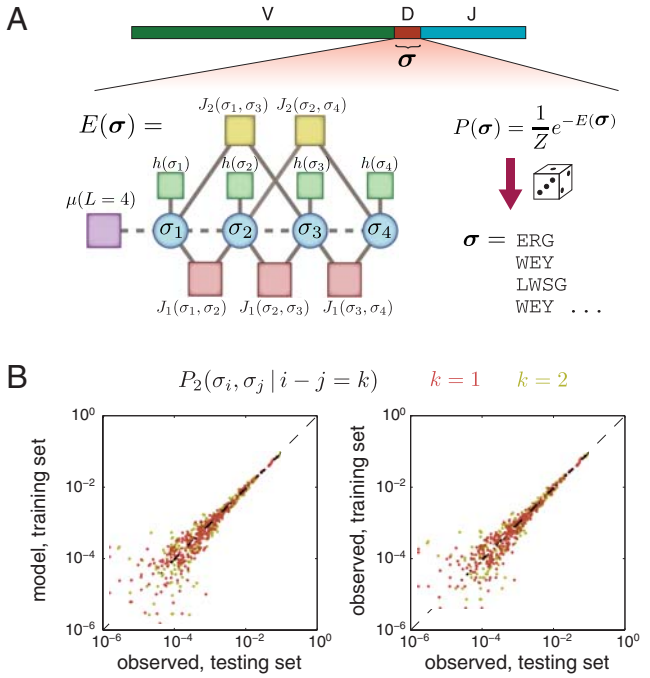

Fig. 1. Maximum entropy model. (*A*) The model of the D region is viewed as a system of interacting residues $(\sigma_1, \ldots, \sigma_L)$ in thermal equilibrium, schematized here by its interaction network for $K = 2$. To each sequence $\boldsymbol{\sigma}$ is associated an energy $E(\boldsymbol{\sigma})$ (Eq. **3**). Then the sequences of the repertoires are drawn at random from the Boltzmann distribution (Eq. **2**). (*B*) Fit quality and control for overfitting. Pairwise frequencies of nearest- ($k = 1$, red) and second-nearest neighbor ($k = 2$, yellow) residues. (*Left*) Comparison between the model prediction, where the model was fitted with the training data, and the testing data. (In this figure the maximum interaction range is $K = 2$, but $K = 1$, 3, and 4 gave similar results.) (*Right*) Direct comparison between the training data and the testing data. The scatter is of the same magnitude, showing that the model is as precise as the data allow.

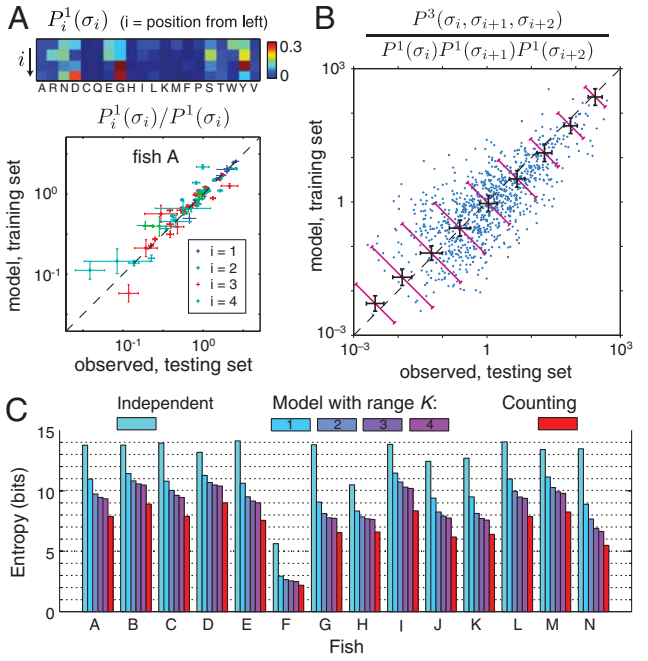

Fig. 2. Local observables and the entropy are well captured by the model. (*A*) Position-dependent amino acid frequency. (*Top*) Frequency as a function of position $i = 1, \ldots, 4$ from the left end of the sequence. (*Bottom*) Comparison between model and data of position-dependent frequencies, normalized by the prediction of the independent model. Error bars are obtained as the standard deviation over many choices of partition between training and testing sets. (*B*) Comparison of triplet frequencies of contiguous amino acids, normalized by the prediction of the independent model. The small crosses illustrate one choice of the training/testing partition. The black error bars represent the average measurement error made on a triplet frequency at that frequency value, obtained as the standard deviation over many choices of the training/testing partition. The diagonal error bars show the average error between model and data. (*C*) Entropy of all fish: from frequency counting, from the independent model, and from the maximum entropy model with range $K = 1, \ldots, 4$.

Local Biases. The model we have constructed does not incorporate any site specificity—interactions between amino acids depend on the distance between them but not on their absolute location along the sequence (Eq. **3**). But, because amino acids at the start or end of the sequence have only half the number of neighbors that are available to sites in the middle of the sequence, the model predicts "end effects" that will be manifest as position-specific biases in amino acid composition. As shown in Fig. 2*A*, these predicted biases can be large, so that the probability of finding particular amino acids at specific sites, $P_i^1(\sigma)$, can vary by more than two orders of magnitude. These predictions are in very good agreement with the data. We emphasize once again that these predictions of site-specific substitution patterns are obtained from a model that has no explicit site-specific information (both h and J are position-independent) and that is learned from an ensemble of sequences that have not been aligned. In a similar spirit, we find good agreement between the predicted and observed probabilities of contiguous amino acid triplets (Fig. 2*B* and *SI Text*), even though the model has no explicit three-site interactions.

Zipf's Law. The space of possible sequences is so large that we cannot test the predictions for the distribution $P(\boldsymbol{\sigma})$ directly. Still, we can get a global view of the distribution through a Zipf plot, in which we put the observed sequences in order on the basis of their frequency of occurrence, and plot probability P vs. rank r, as in Fig. 3. We see that both the data and the predictions of the model are very close to obeying Zipf's law, $P \propto 1/r$ (21, 22), and the data and model agree very well with one another. The same pattern is observed in all fish, although the ranking of particular sequences varies. The dynamic range over which we can observe Zipf's law is limited by the number of independent sequences that are read in the experiments, but the model predicts that this behavior should continue even if this number were extended by an order of magnitude.

Zipf's law first attracted attention in the context of language (22), and many models have been proposed for the origin of this behavior. Even before Zipf's work, it was known that some growth processes with mutations can generate Zipfian distributions (21, 23). Because we have built a model out of measured pairwise correlations, with strong analogies to statistical mechanics, we emphasize that Zipf's law reflects the proximity of a critical point in the strength of the underlying interactions. The rank of a state $\boldsymbol{\sigma}$ is determined by the number of states with higher probability or lower energy in Eq. **2**. But counting the number of states is equivalent to measuring the (microcanonical) entropy, and then Zipf's law is the statement that the entropy grows linearly with the energy, with slope one (see *SI Text*). This locally linear relation between energy and entropy is characteristic of thermodynamic systems at a critical point (24) and could not emerge from a system of noninteracting units or even from an interacting system with slightly weaker or stronger correlations. Thus, the strength of correlations that we see in the real sequences corresponds to interactions with a critical strength, restricting the set of allowed sequences substantially but not forcing the system to "freeze" into a small set of possibilities.

Entropy. The fundamental quantity in a maximum entropy construction is the entropy S itself. Entropy measures the diversity in sequence space and hence is also a fundamental quantity from a biological point of view. If we imagine that sequences are constructed by choosing amino acids at random, then the entropy could be as large as $\log_2(20)$ bits per residue, or a total of ~15 bits for the average length D region. For almost all fish (F is an exception and is excluded from further analyses), the observed biases in the use of the different amino acids do not reduce this very much; that is, if we choose amino acids independently at every site but with the observed frequencies,

$$P_{\rm ind}(\boldsymbol{\sigma}) \equiv P(L)\prod_{i=1}^{L} P_1(\sigma_i), \qquad [4]$$

then the entropy $S[P_{\rm ind}]$ of this independent model is nearly $\log_2(20)$ bits per residue. We can think of the maximum entropy model as part of a hierarchy, in which the entropy is reduced every time we take account of additional correlations (25). As shown in Fig. 2*C*, the entropy is reduced significantly as we take account of correlations between neighboring amino acids, corresponding to $K = 1$ in Eq. **3**. It is reduced further when we include next-nearest neighbors ($K = 2$), and the reduction seems to plateau as we include more distant neighbors ($K = 3, 4$). Including all of these pairwise correlations pushes the total entropy well below 10 bits for all fish, so that out of tens of thousands of possible sequences, most of the distribution is concentrated in only a few hundred ($\sim 2^S$) sequences, and this is consistent with what we observe in the Zipf plots (Fig. 3). This restriction of sequence space is even more dramatic when we realize that, given the maximum length of the D regions, there really are tens of millions of possible sequences.

The difference between the entropy of the independent model and the true entropy, $I = S[P_{\rm ind}] - S[P]$, measures the overall strength of correlations in the system and is called the multiinformation. The maximum entropy model predicts a value for $I^{(m)} = S[P_{\rm ind}] - S[P^{(m)}]$ that must be smaller than I, and the ratio $I^{(m)}/I$ measures the fraction of the correlated structure that we capture in our model. The difficulty is that, because sequence space is large, estimating the entropy $S[P]$ is difficult. Methods are available, however, that allow us to estimate $S[P]$ even when we do not have enough samples to accurately estimate $P(\boldsymbol{\sigma})$ itself, as explained in *SI Text* and ref. 26. By using these methods, we find, as shown in Fig. 2*C*, $I^{(m)}/I$ in the range from 0.67 to 0.91 across the different fish. Thus our maximum entropy model, on the basis of only pairwise correlations, captures between two-thirds and 90% of all the correlated structure in the distribution of sequences.

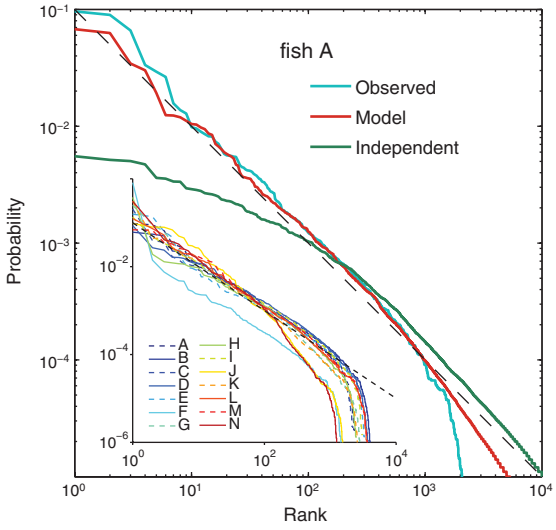

Fig. 3. The distribution of D regions obeys Zipf's law. Probability of D region sequences as a function of their rank in fish A, as observed from frequency counting (*Blue Line*), and as predicted by the independent (*Green Line*) and the maximum entropy model with $K = 2$ (*Red Line*). The dashed line has slope −1. (*Inset*) The same for all fish, from frequency counting.

Comparison Between Fish. The analysis of entropies shows that the repertoires of individual fish span only a tiny fraction of the possible sequence space. Do the repertoires of different fish overlap with each other, or are they distinct? To answer this question, we first computed a similarity factor $\mathrm{Sim}[P_\alpha, P_\beta]$ between repertoire distributions (see *SI Text*). This factor takes values between 0 and 1 and measures the difficulty of guessing to which of the two repertoires (α or β) a given sequence belongs. Fig. S2 shows the similarity factor for all pairs of fish, as calculated by the maximum entropy model (see *SI Text*). Whereas the choices of V, D, and J segments are correlated with the family relations among the fish (18), this measure of similarity among D regions is not.

To study repertoire specificity beyond two fish, we looked at the average information that the sequence $\boldsymbol{\sigma}$ of a single antibody molecule carries about the identity α of the fish from which it is drawn,

$$I(\alpha;\boldsymbol{\sigma}) \equiv \sum_{\boldsymbol{\sigma},\alpha} P(\boldsymbol{\sigma},\alpha)\log_2\left[\frac{P(\boldsymbol{\sigma},\alpha)}{P(\boldsymbol{\sigma})P(\alpha)}\right], \qquad [5]$$

where $P(\boldsymbol{\sigma},\alpha)$ is the probability that a sequence picked at random in the dataset be $\boldsymbol{\sigma}$ and come from fish α. Fig. 4 represents this mutual information as a function of the fish entropy $S_\alpha = -\sum_\alpha P(\alpha)\log_2[P(\alpha)]$ for many subgroups of fish of various sizes. The fish entropy is an upper bound to the mutual information and is reached only when sequences give perfect information about which fish they came from, i.e., when each sequence belongs to one fish uniquely. Although the mutual information remains far from this upper bound, it keeps growing linearly with the entropy as the size of the group is increased, each fish adding its own unique diversity. Importantly, this individuality of the sequence ensembles depends dominantly on correlations, because in the independent model, $P_{\rm ind}(\boldsymbol{\sigma})$ from Eq. **4**, the mutual information between identity and sequence is roughly a factor of four smaller (Fig. 4, *Lower Inset*). All 13 fish do not suffice to cover the potential diversity of D regions, as evidenced by the absence of saturation.

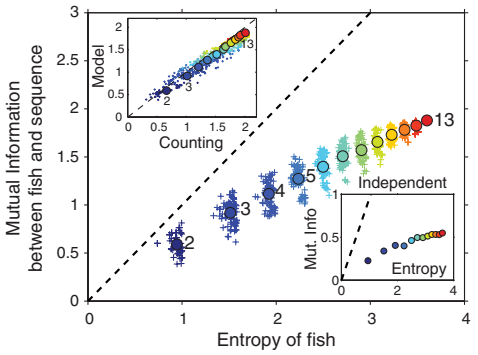

Fig. 4. Fish repertoires overlap yet are specific. Mutual information between fish and sequence vs. the entropy of fish. Each point is a subgroup of all 13 fish (excluding fish F), color-coded by its size (from dark blue to red). Filled circles are averages over groups of each size. (*Upper Inset*) Comparison between mutual information estimated from counting observed sequences and that predicted by the maximum entropy model. (*Lower Inset*) Mutual information vs. fish entropy, as predicted by the independent model.

Multivalley Landscape. The energy function in Eq. **3** includes competing interactions—the couplings J can be positive or negative, favoring both correlated and anticorrelated amino acid substitutions at different sites. From the statistical mechanics of disordered systems (27) we know that such competition can lead to "frustration" and many metastable states. A metastable state is defined as a local minimum of the energy landscape or, in probabilistic language, a local maximum of the probability distribution. Does this happen in the case of antibody diversity?

Our model assigns an energy to every sequence, but to find local minima in this landscape we need to define "local." Because mutations occur at the level of nucleotides, we work in the space of nucleotide sequences; to assign a (free) energy to nucleotide sequences, we translate to amino acids, compute $E(\boldsymbol{\sigma})$ from Eq. **3**, and add a correction term for the entropy of codon usage. Then we say that two sequences are adjacent if (*i*) they differ by one nucleotide, (*ii*) they differ by one nucleotide insertion and one deletion, or (*iii*) they differ by three insertions or three deletions; the last criterion is necessary because, by construction, the lengths of D regions is a multiple of 3. With this conservative definition, we find ~10 local minima per fish; examples are shown in Fig. 5. Some of these states correspond to the D regions encoded in the genome, as shown in Fig. 5*A*, but many do not. The structure of the energy landscape, and hence the probability with which sequences appear in the organism's antibody repertoire, thus has elements that are not simply a record of genomic history but presumably reflect rapid adaptation to the antigenic environment.

Each metastable state defines a basin of attraction or valley in the energy landscape, and we can assign each sequence to its corresponding valley by moving "downhill": Starting from a given sequence, go to the lowest energy neighbor, and continue doing so until the energy stops decreasing and a metastable state has been reached. Fig. 5*B* represents the energy of all sequences in a basin of attraction as a function of their distance (in number of steps) to the metastable states; although there are differences of detail, the different basins have very similar structures. As we explore away from the minimum energy in each basin, at some point we reach the "pass" that connects neighboring valleys; the trajectories over these passes are analogous to the trajectories from reactants to products in a chemical reaction, with the pass identifiable as the transition state (28). Because the sequences are not too long, we can find these paths by a conventional Monte Carlo procedure (see *SI Text*), and in most cases we found continuous paths through adjacent observed sequences between metastable states. When the two metastable states had the same length, we found paths where each step was a single nucleotide mutation. Fig. 5*C* summarizes the connections among the seven most populated metastable states in the repertoire of fish A. Taken together, these results on the energy landscape imply that the repertoire explores much of the sequence space and is not slaved to the genomic templates or to any specific sequence arising in the adaptation process.

Summary and Discussion

The formation of the antibody repertoire is an example of an accelerated evolutionary process under selective pressure. Antibodies in a given organism are correlated both through their genomic origin and as a result of the adaptation history. In this study we have analyzed the repertoire of B cell antibodies by building compact models of the hypervariable region of their heavy chain, on the basis of the principle of maximum entropy.

The reduction of parameters achieved by the model is enormous. Even though we are looking at the relatively short hypervariable D regions, there are tens of millions of possible sequences, and in principle each sequence occurs with a different probability in the repertoire. In contrast, the number of parameters of our model is of order $400K$, where K is the interaction range. Importantly, this number scales reasonably with sequence size, making our approach tractable for systems in which the relevant sequence is much longer, including the hypervariable regions in other species. The compactness of the model allows for generalization, so we can predict quantities that are not deducible simply by counting sequences in the observed sample: the overall size of the repertoire, the overlaps between repertoires of different individuals, and the probability of finding new, as yet unobserved, sequences in larger samples from the same individual.

The maximum entropy construction accounts for correlations between amino acid substitutions at different residue positions through an effective interaction structure. These interactions are strong enough to generate a dramatically different ensemble of sequences than would be expected if substitutions at each site were independent. The diversity of the repertoire is substantially

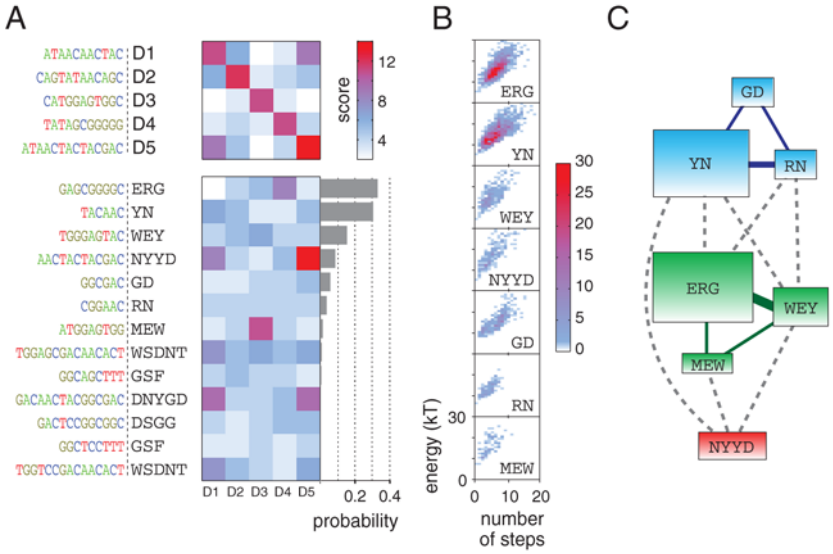

Fig. 5. Metastable states (data from fish A). (*A*) Lower: Scores of pairwise alignments between the genomic segments D1–D5 and the metastable states. The bar plot represents the total weight of the basins of attraction of each metastable state. Upper: Scores of alignments of the genomic segments with themselves and with each other are shown for comparison. (*B*) Basins of attractions of the 7 most populated states. A density plot represents the energy of the sequences vs. the number of steps separating them from their metastable state by steepest descent. (*C*) Connectivity of the sequence space. Lines indicate the existence of paths of adjacent sequences between two metastable states. When the link is a solid line, there exists a path made only of single-nucleotide mutations.

reduced (from an entropy of ∼14 bits to ∼8 bits), the distribution of sequences obeys Zipf's law, and the distribution has a complex structure of "metastable states," clusters of sequences with high probability.

We have addressed the question of individuality, by using our model and tools from information theory. At one extreme, the fish could be completely different from each other, with each fish bringing its own set of unique sequences. At the other extreme, fish could have more or less identical repertoires, sharing the same antibodies in the same proportions. We found an intermediate situation, where about 50% of the repertoire diversity was unique to each fish (Fig. 4), and the rest shared among all fish. As one concatenates the individual repertoires, including more and more fish, the size of the resulting metarepertoire must saturate, because the number of possible antibody sequences is finite. But this saturation is not reached even for 13 fish, meaning that each fish is still unique compared to all other 12 taken together and not only compared to each of them separately.

The details of the adaptation process undergone by the repertoire are largely unknown, and our model provides only a first step to aid in its study. What is the mutation mechanism? How do recognition and selection work? Our observation of Zipf's law provides an important constraint on these mechanisms. As we have emphasized, this behavior arises only if the interactions between substitutions at different sites have a critical strength. But these interactions are just a summary of the mutation and selection dynamics. There are simple growth processes with mutation that can generate Zipfian distributions (23), but much work remains to find a realistic model that generates the full structure of $P(\boldsymbol{\sigma})$.

The structure of the energy landscape underlying our model shows that the repertoire decomposes into several components. Each component is centered on a metastable state, a peak in the probability distribution of sequences. Some metastable states are closely related to the genomic templates, although rarely identical, whereas others are not attributable to any genomic template. We can think of these metastable states as markers of adaptation. For example, an infection could have caused the proliferation of antibodies particularly efficient for recognizing a specific antigen, thus creating a peak in the probability landscape, which suggests the possibility of using metastable states and their basins of attraction for probing infectious history, perhaps in experiments that follow the dynamics of the sequence ensemble over time.

The clusters associated with the metastable states are not completely disconnected from one other: We found continuous paths of observed sequences between most metastable states, which means that, far from being slaved to their genome, the D sequences have the freedom to explore sequence space extensively during the adaptation process, forming a large cloud of possibilities between the highly concentrated regions of the sequence space, i.e., the metastable states, whether they be genomic or not. The method we have used for finding these paths—a Metropolis walk in energy space—further illustrates the power of the maximum entropy model: Because it naturally favors low energy barriers, this algorithm is more likely to find paths where all sequences are present in the data. More generally, it could be used as a tool for retracing mutation paths between any two sequences and could lend us insight into the repertoire's evolutionary history.

Finally, the success of maximum entropy models in accounting for the higher order statistical structure of the sequence ensemble encourages us to think that this approach is more widely applicable. The maximum entropy formalism shows how, as in many statistical physics problems, the observable correlations between amino acid substitutions at any two sites provide the signatures of collective behavior in the system as a whole. The idea that crucial aspects of life should be viewed as emergent, collective phenomena has been discussed for decades. The challenge has been to move beyond metaphor by developing precise mathematical tools for extracting quantitative models of this collective behavior from experiment. We believe that we have taken useful steps in this direction in the work reported here.

ACKNOWLEDGMENTS. We thank S.R. Quake, J.A. Weinstein, and their colleagues for sharing their data and for several helpful discussions. This work was supported in part by National Science Foundation Grant PHY–0650617 and by National Institutes of Health Grant P50 GM071598; T.M. was supported by the Human Frontiers Science Program.

1. Pal C, Papp B, Lercher M (2006) An integrated view of protein evolution. *Nat Rev Genet* 7:337–348.
2. Branden C, Tooze J (1991) *Introduction to Protein Structure* (Garland Science, New York).
3. Cordes MH, Davidson AR, Sauer RT (1996) Sequence space, folding and protein design. *Curr Opin Struct Biol* 6:3–10.
4. Socolich M, et al. (2005) Evolutionary information for specifying a protein fold. *Nature* 437:512–518.
5. Russ WP, Lowery DM, Mishra P, Yaffe MB, Ranganathan R (2005) Natural-like function in artificial ww domains. *Nature* 437:579–583.
6. Bialek W, Ranganathan R (2007) Rediscovering the power of pairwise interactions. arXiv:0712.4397.
7. Schneidman E, Berry MJ, Segev R, Bialek W (2006) Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* 440:1007–1012.
8. Tkačik G, Schneidman E, Berry MJ, Bialek W (2006) Ising models for networks of real neurons. arXiv:q-bio/0611072.
9. Seno F, Trovato A, Banavar JR, Maritan A (2008) Maximum entropy approach for deducing amino acid interactions in proteins. *Phys Rev Lett* 100:078102.
10. Tang A, et al. (2008) A maximum entropy model applied to spatial and temporal correlations from cortical networks in vitro. *J Neurosci* 28:505–518.
11. Volkov I, Banavar JR, Hubbell SP, Maritan A (2009) Inferring species interactions in tropical forests. *Proc Natl Acad Sci USA* 106:13854–13859.
12. Dhadialla PS, Ohiorhenuan IE, Cohen A, Strickland S (2009) Maximum-entropy network analysis reveals a role for tumor necrosis factor in peripheral nerve development and function. *Proc Natl Acad Sci USA* 106:12494–12499.
13. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T (2009) Identification of direct residue contacts in protein-protein interaction by message passing. *Proc Natl Acad Sci USA* 106:67–72.
14. Halabi N, Rivoire O, Leibler S, Ranganathan R (2009) Protein sectors: Evolutionary units of three-dimensional structure. *Cell* 138:774–786.
15. Jaynes ET (1957) Information theory and statistical mechanics. *Phys Rev* 106:620–630.
16. Jaynes ET (1957) Information theory and statistical mechanics. II. *Phys Rev* 108:171–190.
17. Hozumi N, Tonegawa S (1976) Evidence for somatic rearrangement of immunoglobulin genes coding for variable and constant regions. *Proc Natl Acad Sci USA* 73:3628–3632.
18. Weinstein JA, Jiang N, White RA, Fisher DS, Quake SR (2009) High-throughput sequencing of the zebrafish antibody repertoire. *Science* 324:807–810.
19. Murphy KP, Travers P, Janeway C, Walport M (2008) *Janeway's Immunobiology* (Garland, New York).
20. Lieschke GJ, Trede NS (2009) Fish immunology. *Curr Biol* 19:R678–R682.
21. Newman MEJ (2005) Power laws, Pareto distributions and Zipf's law. *Contemp Phys* 46:323–351.
22. Zipf GK (1932) *Selected Studies of the Principles of Relative Frequency in Language* (Harvard Univ Press, Cambridge, MA).
23. Yule G (1925) A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, FRS. *Phil Trans R Soc B* 213:21–87.
24. Huang K (2008) *Statistical Mechanics* (Wiley, New York), 2nd Ed.
25. Schneidman E, Still S, Berry MJ, Bialek W (2003) Network information and connected correlations. *Phys Rev Lett* 91:238701.
26. Strong SP, Koberle R, de Ruyter van Steveninck RR, Bialek W (1998) Entropy and information in neural spike trains. *Phys Rev Lett* 80:197–200.
27. Mézard M, Parisi G, Virasoro MÁ (1987) *Spin Glass Theory and Beyond* (World Scientific, Singapore).
28. Hänggi P, Talkner P, Borkovec M (1990) Reaction-rate theory: Fifty years after Kramers. *Rev Mod Phys* 62:251–341.