PROTEINS: Structure, Function, and Genetics 11:297–313 (1991)

Information-Theoretical Entropy as a Measure of Sequence Variability

Peter S. Shenkin, Batu Erman, and Lucy D. Mastrandrea
Department of Chemistry, Barnard College, New York, New York 10027

ABSTRACT **We propose the use of the information-theoretical entropy, $S = -\Sigma p_i \log_2 p_i$, as a measure of variability at a given position in a set of aligned sequences. p_i stands for the fraction of times the i-th type appears at a position. For protein sequences, the sum has up to 20 terms, for nucleotide sequences, up to 4 terms, and for codon sequences, up to 61 terms. We compare S and V_S, a related measure, in detail with V_K, the traditional measure of immunoglobulin sequence variability, both in the abstract and as applied to the immunoglobulins. We conclude that S has desirable mathematical properties that V_K lacks and has intuitive and statistical meanings that accord well with the notion of variability. We find that V_K and the S-based measures are highly correlated for the immunoglobulins. We show by analysis of sequence data and by means of a mathematical model that this correlation is due to a strong tendency for the frequency of occurrence of amino acid types at a given position to be log-linear. It is not known whether the immunoglobulins are typical or atypical of protein families in this regard, nor is the origin of the observed rank-frequency distribution obvious, although we discuss several possible etiologies.**

Key words: information theory, entropy, variability, sequence comparison, immunoglobulins, antibodies

INTRODUCTION

It seems appropriate to recall here how the subject of this paper, which seems distant from Cyrus Levinthal's main research interests, in fact emerged out of work that was begun in his laboratory. During the mid-1980s, Cyrus Levinthal, Richard Fine, David Yarmush, Huajun Wang and one of us (P.S.S.) collaborated in an effort to predict immunoglobulin loop conformations.[1,2] We had the benefit of an algorithm (random tweak) that allowed us to explore the conformational space of long loops, and this enabled us to target for modeling the full lengths of the complementarity determining regions (CDRs), as set forth by Kabat et al.[3] Around the same time, several other groups[4,5] published similar efforts that took advantage of the observation that the crystallographic structures then available for immunoglobulins exhibited structural variability over only parts of Kabat's CDRs. These workers were able to obtain good results by modeling only those parts, either by direct structural analogy to existing structures or by using search algorithms appropriate to shorter loops.

Although there is no a priori reason to assume that the regions that are most variable in sequence must also be the most variable in tertiary structure, we nevertheless found it interesting that this seems not to be the case, and the present authors decided to examine the primary-structure variability criterion which Wu and Kabat originally used to define the CDRs. In their classic 1970 paper,[6] these investigators defined the variability of a position in a set of aligned sequences as follows:

$$V_K = k/p_1. \quad (1)$$

We use the symbol V_K to represent Wu and Kabat's variability measure. In the definition, k is the number of different amino acid types that appear at the position in question, and p_1 is the fraction of times the most common amino acid type at this position appears there:

$$p_1 = n_1/N,$$

where n_1 is the number of times the most common amino acid type appears and N is the total number of sequences occupied at the site in question. In the following discussion, we use the symbols p_i and n_i to refer to the probability of appearance and the occupancy number, respectively, of the i-th-most-common amino acid type at a given position.

Received May 28, 1991; revision accepted July 9, 1991.

Address reprint requests to Dr. Peter S. Shenkin, Department of Chemistry, Barnard College, 3009 Broadway, New York, NY 10027.

Lucy D. Mastrandrea and Batu Erman were Pew Foundation Undergraduate Research Fellows at Barnard over the summers of 1989 and 1990, respectively. Ms. Mastrandrea was then a student at Manhattan College and is now at the State University of New York, Buffalo, New York. Mr. Erman was then a student at Hamilton College and is now in the Department of Biological Sciences, Brandeis University, Waltham, MA.

Abbreviations: CDR, complementarity determining region; ALC, all light chains; AHC, all heavy chains; HHC, human heavy chains.

V_K ignores all p_i except p_1. We thought that a better definition of variability might be one that takes all p_i into account, and there are reasons, discussed below, to propose the information-theoretical entropy for this use. This measure is defined by the equation

$$S = -\sum_{i=1}^{k} p_i \log_2 p_i. \quad (2)$$

This measure describes how "spread out" the distribution of the k types is and was proposed by Shannon et al.[7] as a measure of the average information per symbol contained in a message when the a priori probabilities of the symbols are given by p_i. A similar equation describes the ideal molar entropy of mixing in thermodynamics; within this context a highly variable position is viewed as one in which several types "mix" at a given position; the greater the number of types and the more uniform the composition, the greater the variability.

Garnier and co-workers[8] have employed algorithms based on the Shannon entropy for protein secondary-structure prediction. Stormo and co-workers[9,10] have used the Shannon entropy to analyze the expected and observed occurrence of protein binding sites in DNA sequences, and Berg and Von Hippel[11,12] have demonstrated a relationship between the statistical entropy of DNA binding sites and the thermodynamics of protein–DNA binding. Our use of this concept is related to that of Stormo and co-workers, but our definition as well as our purpose differs from theirs in several ways, as discussed below.

Finally, Jores et al.[13] have proposed an extension to Kabat's measure in which the occurrence of pairs, rather than singlets, of amino acids at a given position is used in an equation similar to Equation (1). These investigators claim that this measure allows clearer delineation of hypervariable regions when applied to T-cell antigen receptor sequences.

METHODS

A database of immunoglobulin sequences equivalent to that which appears in the compendium by Kabat et al.[3] was supplied in computer-readable form by Professor Elvin Kabat of Columbia University and by Harold Perry of Bolt, Beranek and Newman. The database was accessed and results computed using a C program called Var written by the authors and available to interested researchers on request.

To produce the data used in this paper, Var was run in a mode that considers residues GLU, GLN, ASP, and ASN to be distinct, and ignores sequence positions in which ambiguity is specified as GLX or ASX. Thus, variability plots (see Fig. 2) should be compared with the corresponding plots in the compilation of Kabat et al.[3]

TABLE I. Immunoglobulin Chain Numbering Schemes

Light chains		Heavy chains	
Kabat et al.[3]	Present study	Kabat et al.[3]	Present study
a. Sequence numbers			
0–27	0–27	0–35	0–35
27A–F	28–33	35A–B	36–37
28–95	34–101	36–52	38–54
95A–F	102–107	52A–C	55–57
96–106	108–118	53–82	58–87
106A	119	82A–C	88–90
107–109	120–122	83–100	91–108
		100A–K	109–119
		101–113	120–132
b. Complementarity determining regions			
CDR 1 24–34	24–40	31–35B	31–37
CDR 2 50–56	56–62	50–65	52–70
CDR 3 89–97	95–109	95–102	103–121

When Kabat et al.[3] make their variability plots, they ignore the positions that are lettered additions to their generic numbering scheme, such as positions 27A–F for light chains. We include such positions on our plots and number all positions consecutively for the purposes of labeling the figures. The conversions relating the two numbering systems are given in Table I.

Weighted least-square fits to linear relationships between log p_i and i were made as follows. Each n_i is a count of the number of times type i appears among the N sequences available at a given position. We may assume statistical counting errors,[14] so that $\sigma(n_i) \approx \sqrt{n_i}$. Since we are fitting log n_i/N vs. i to a straight line, we need to weight each (i, log p_i) pair by $w_i = 1/\sigma^2$ (log p_i). If $\sigma(n_i) << n_i$, we can use the differential relationship $d \log p_i = dp_i/p_i = dn_i/n_i$ to infer that $\sigma(\log p_i) \approx \sigma(n_i)/n_i \approx 1/\sqrt{n_i}$ which gives $w_i \approx n_i$. In fact, the assumption that $\sigma(n_i) \approx \sqrt{n_i}$ is not a good one for small n_i, particularly when the distribution is discrete, since under these conditions Poisson counting statistics are poorly approximated by a Gaussian distribution. The differential approximation is also poor for small n_i, since here the presumed $\sigma(n_i)$ has the same order of magnitude as n_i itself. Nevertheless, we proceeded with the described weighting scheme, for lack of a clearly superior alternative. This scheme does weight observations more strongly the greater the value of n_i, which is the chief feature we wish to preserve, and is a good approximation except for very small counts.

Calculations were performed on Silicon Graphics and Sun workstations. Plots were produced using the software packages Mathematica (Wolfram Research, Champaign, IL 61826) and Grtool (Paul Turner, Department of Environmental Science and

TABLE II. Comparison of V_K, S, and V_S*

Position	N	n_1	n_2	n_3	V_K	S	V_S
Example 1. Surprising grouping of positions by V_K							
a	1,000	500	499	1	6.0	1.01	12.09
b	1,000	500	500		4.0	1.00	12.00
b	1,000	500	250	250	6.0	1.50	16.79
Example 2. Unrobustness of V_K							
a	9,999	9,999	—	—	1.0	0.000	6.000
b	10,000	9,999	1	—	2.0	0.0015	6.006
c	10,001	9,999	1	1	3.0	0.0029	6.012

*Fictitious sequence data. See text for further details.

Engineering. Oregon Graduate Institute of Science and Technology, Beaverton, OR 97006).

RESULTS

V_K, S, and V_S as Measures of Variability

We first show by example that S and a related measure, which we call V_S, correlate better than V_K with intuitive notions of variability. Consider a position in a set of aligned sequences occupied by three types with populations (500, 499, 1). It seems clear that this position is slightly more variable than one exhibiting occupancies (500, 500) and a good deal less variable than one exhibiting occupancies (500, 250, 250). As shown in Example 1, Table II, however, V_K assigns equal variabilities to the first and third cases and a significantly lower variability to the second. By contrast, S orders the variabilities as our intuition dictates. This ordering is also exhibited by another measure, which we call V_S, and which is also exhibited in Table II. We define V_S by

$$V_S = 6 \times 2^S. \tag{3}$$

We shall see later that V_S is linearly related to V_K. Since V_S is a monotonic function of S, these two measures will exhibit similar trends.

V_K exhibits the undesirable mathematical property of being discontinuous over its domain. V_K takes on a value of one when all sequences are occupied by the same amino acid at the position in question. When two amino acids appear, k in Equation (1) is equal to 2 and, since by definition p_1 must be less than unity, V_K cannot take on values between one and two. This discontinuity is not fatal to the purpose for which the measure was designed; nevertheless, it comes as a surprise, and has no intrinsic meaning within the context of sequence variability.

Finally, V_K is unrobust in the extreme. Suppose one is comparing a large number of sequences and that at some position only one type has been observed. If a new sequence is then determined, and this sequence exhibits a new type at the position in question, V_K will jump from one to something more than two, regardless of the value of N; thus, a 100% increase in V_K can be caused by an infinitesimal change in the data. This is not a consequence of the discontinuity of V_K between one and two; the same phenomenon can occur when a third amino acid type is found at a position at which only two were previously observed. Here, for large N, the jump in V_K can be as great as 50%. By contrast, as N increases, S and V_S exhibit less and less variation when k changes incrementally. Example 2, Table II provides a numerical example. We would thus expect S and V_S to converge on self-consistent values more quickly than V_K as a database of sequences grows.

It is useful to summarize some additional properties of S, V_S, and V_K. S and V_S are continuous over their respective ranges. When a single type occupies all sites, S is equal to its minimum value, S_{min}, which is zero; here the term (1 ln 1) in Equation (2) is clearly equal to zero, and terms of the form (0 ln 0) go to zero by virtue of l'Hopital's rule. S reaches its maximum value, S_{max}, when all types exhibit equal occupancy. Here, if k_{max} is the maximum number of types, the occupancy of each type is $1/k_{max}$, and we have $S_{max} = \log_2 k_{max}$. For proteins, $k_{max} = 20$, giving $S_{max} \approx 4.32$. Application of Equation (3) to S_{min} and S_{max} gives $V_{S,min} = 6$, always, and $V_{S,max} = 120$, for proteins.

V_K takes on its minimum value, $V_{K,min} = 1$, under the same conditions that minimize S. The maximum values of V_K, S, and V_S also occur under the same conditions: when k_{max} types appear, each with $p_i = 1/k_{max}$, Equation (1) gives $V_{K,max} = k^2_{max}$, which is 400 for proteins. Except at their minimum and maximum values, S and V_K do not exhibit a functional relationship; for any value of either of them, a range of values is possible for the other.

Intrinsic Meaning of S

Unlike V_K, S has an intrinsic meaning that enables us to say just what we mean when we claim that one position is more variable than another. Simple combinatorics[15] tells us that if a group of N objects falls into k types, n_1 objects of the first type, n_2 of the second, etc., up to n_k, then the number of ways of ordering these objects into distinguishable arrangements is

$$W = \frac{N!}{\prod_{i=1}^{k} n_i!}. \tag{4}$$

In the limit of large N, Stirling's approximation gives

$$\log_2 W = -N \sum_{i=1}^{k} p_i \log_2 p_i \tag{5}$$

with $p_i = n_i/N$ as before. Comparison with Equation (2) shows that NS can be interpreted as the logarithm to the base two of the number of ways the

objects can be reordered in a distinguishable fashion.

This is the information-theoretical equivalent of Boltzmann's famous hypothesis, $S = k_B \ln W$, which is the historical origin of the statistical understanding of the thermodynamic entropy.[15] In statistical mechanics, one conventionally uses the natural logarithm, rather than the base-two logarithm, and Boltzmann's constant, k_B, appears to render the results compatible with the system of units used in thermodynamics. In addition, there is a small notational difference. In thermodynamics, the symbol S is generally used for the extensive quantity that in our terminology is NS; for the intensive quantity which we denote by S, thermodynamicists generally use $\bar{S}$. Our usage corresponds to that of information theory[7] and, in the language of that discipline, if a set of objects can be classified into k types, and if type i has p_i as its a priori probability of appearance, then S as calculated from Equation (2) is the expected number of bits of information conveyed when we are told to which type a previously unidentified object belongs. For example, if a particular position in a set of aligned protein sequences has Shannon entropy S, then we gain on average S bits of information by being told which amino acid actually appears at that position in a particular sequence. We expect to gain NS bits of information when we are told which amino acid type appears at this position in each of a list of N sequences, and this list of known types then specifies one distinguishable ordering of the N objects.

Finally, we define

$$k^* = 2^S. \tag{6}$$

k^* is the number of types that would give the observed value of S if all the types occurred with equal frequency $p^* = 1/k^*$. k^* is equal to unity when only one type appears and to k_{max} (20, for proteins) when S is equal to S_{max}. k^* need not be an integer, and $k^* \leq k$, the equality arising only when the k types actually do occur with equal frequency. By Equation (3), $V_S = 6k^*$.

Although the number of sequences available for a calculation of S at a given position may not be large enough for Stirling's approximation to hold to high precision, the sequences available can be thought of as a sampling from the much larger ensemble that occurs or could occur in nature. The calculated value of S then provides an estimate of the entropy-per-position of the larger ensemble. The current sequence database is, however, not a random sample of the entire ensemble, since the order in which sequences are accumulated is biased by the convenience and interests of experimentalists. Therefore, S values based on an existing sequence database are likely to exhibit distortions from the true ensemble values; this is true for V_K as well.

TABLE III. Overall Characteristics of Data Sets Used

Dataset	ALC*	AHC†	HHC‡
Sequences	1,071	767	158
Positions	123	133	132
Occupied sites††	59,974	50,842	7,698

*All light chains.
†All heavy chains.
‡Human heavy chains.
††(Occupied sites < Sequences × Positions), since not every sequence is occupied at every position. This can be due either to insertions and deletions in the sequences or to incomplete experimental data.

Other information-based measures than the one we employ are possible. In their study of polynucleotide binding sites, Schneider et al.[10] use an information measure that reflects the degree of nonrandomness of appearance of the four bases at a position in a set of aligned sequences. This involves weighting the observed frequencies of appearance of the types at a site by the overall frequencies observed in the genome. In our application, this would ascribe different information values to two sites, one of which was occupied exclusively by serine and the other of which was occupied exclusively by histidine, since these two types have greatly differing overall frequencies of appearance. The measure used by Schneider et al. could be useful in the analysis of protein sequences; it might be interpreted as a measure of the discrimination or "strength of selection" of a site for the residues which occupy it. Our purpose here, however, is to measure variability per se, and this is accomplished by the definition given in Equation (2). As an example, this definition assigns a variability of zero to any site occupied by only a single amino acid type.

In terms of information theory, Equation (2) represents the average amount of information conveyed when the type present at some position in a single sequence is revealed, given a priori knowledge of the probabilities of appearance of all the types at that position. The measure by Schneider et al. represents the average amount of information conveyed per sequence when the probabilities of appearance of the types at a given position are revealed, given a priori knowledge of their probabilities of appearance in the database as a whole—in their case, the genome.

Observed Values of V_K, S, and V_S for Immunoglobulin Families

We will examine results for three groupings of immunoglobulins: all light chains (ALC), all heavy chains (AHC), and, for certain data, human heavy chains (HHC, a subset of AHC) as well. The CDRs of HHC are especially clearly delineated in variability plots.[3] Table III gives overall statistics for the three

groups, and Table IV gives sample output from applying the Var program to dataset ALC.

Figure 1a is a scatter plot of S vs. V_K for all positions in ALC and AHC datasets. Figure 1b is the corresponding plot of V_S vs. V_K. The lower and upper boundaries represent the minimum and maximum possible values, respectively, of S or V_S, given a value of V_K, in the limit of infinite N. Methods for calculating these boundaries are given in a Supplement available on request from the author.

The relationship between S and V_K shown in Figure 1a appears logarithmic; therefore we guessed that a plot of k* vs. V_K would appear roughly linear, and Figure 1b bears out this supposition. We find the average value of V_K to be six times the average of k*; thus, the definition in Equation (3) was adopted so as to define an S-based measure, V_S, which could be inspected on the same scale as V_K. These plots are similar in appearance to Figure 2 in the paper by Jores et al.[13] Their pairwise variability measure correlates well with V_K^2, as might be expected from its definition; however, they do not pursue the implications of this correlation.

The unpopulated portions of the allowed regions in Figure 1a,b correspond to patterns of site occupancy—sets of p_i—never observed in the immunoglobulins. We observe that although most of the allowed ranges of S and V_S are populated, only the lowest third of the V_K range is sampled. More importantly, for any exhibited value of V_K, less than one-half the corresponding feasible range of S or V_S is inhabited; similarly, less than one-half the feasible V_K range is populated for any observed value of S or V_S. The net effect is that V_K and the S-based measures are more highly correlated than they would be if the "occupancy space" implied by the boundaries in the figure were randomly sampled.

The correlation between V_K and the S-based measures persists when variability plots are examined, as in Figure 2a,b,c for datasets ALC, AHC, and HHC, respectively. The correlation exhibits position-to-position regularity; there are very few adjacent positions where V_K and V_S move in opposite directions. The net effect is that the same residues appear highly variable using both V_K and V_S, and both measures identify the same CDRs with about the same degree of ambiguity. Apparently, occupancy patterns of the sort discussed in connection with Table II, which lead to great disparity between V_K and V_S, are rare among the immunoglobulins.

Statistical Models

We wish to further elucidate the nature of the site occupancy patterns that the immunoglobulins do exhibit. Simple models that use summary statistics from the database to infer the relationship between V_K and the S-based measures fail badly. In the simplest possible model, amino acid type distributions at sites are obtained by random selection from a pool of amino acids reflecting the overall composition of the proteins in the database, which is given at the head of the output of the Var program (Table IV). The values of V_K, S, and V_S derived from this model, which we call Model 0, are displayed in Figure 1. Since Model 0 treats sites as samples from the same pool, it predicts no change in variability from position to position beyond that caused by statistical sampling error; thus, the great range of variabilities exhibited by the immunoglobulins is not accounted for. Furthermore, the single variability predicted by this model is much greater than that observed in any immunoglobulin position.

Of course, as biologists or chemists, we are not surprised. We expect that every position in a set of homologous protein sequences will have structural or functional requirements. The finding that no position in the database is neutral with respect to amino acid preference is therefore as expected, although it is imaginable that some position in some protein might exhibit such neutrality. If such a site were found, however, we would view it as an anomaly of potentially great interest.

Consider now not the total amino-acid composition of the protein, but the average rank-frequency distribution of the positions. Disregarding amino acid identities, we sum n_1 for all the positions, n_2 for all the positions, and so on, and use the resulting "grand occupancies" to calculate values of the p_i, V_K, S, and V_S. The corresponding "grand rank-frequency distribution" can be used to calculate variabilities, and we call this procedure Model 1. Like Model 0, Model 1 predicts a single set of variabilities for all positions, sampling error aside. These results are also shown in Figure 1a,b. Unlike the results from Model 0, these points lie in densely populated regions; however, the model still does not mimic the broadness of the observed variability distribution, which is far too great to attribute to sampling error, given that we are sampling approximately 1,000 sequences. Furthermore, this model does not illuminate the origin of the grand rank-frequency distribution that is its basis.

Of course, the failure of model 1 also comes as no surprise. We know from Wu and Kabat's original hypothesis,[6] later borne out by structural studies,[16] that some positions must be variable in order to accommodate a diversity of antigen binding functionalities, whereas others must be invariant, or nearly so, for structural reasons. Any successful model must allow a broad range of variabilities to be exhibited.

Let us now consider a model in which V_K is considered known, and the most likely values of S and V_S are then calculated. Details of this calculation, which we call Model 2, are described in the Supplement. If V_K is specified, only certain (k,p_1) pairs are possible. In Model 2, once we have constrained p_1 and k, we assume that the other (k–1) types have

TABLE IV. Sample Var Output, All Light Chains

```
All Light Chains
Mode 1: ignoring explicit GLX and ASX, treating GLN, GLU, ASN, ASP separately.
Total of 1071 sequences, 59974 sites, compared over 123 positions.
  20 different amino acids appear.
Overall amino acid distribution (number and fraction):
      S       T       G       L       V       A       Q       P       I       Y
   8961    5646    5106    4547    3960    3938    3833    3297    2991    2708
 0.1494  0.0941  0.0851  0.0758  0.0660  0.0657  0.0639  0.0550  0.0499  0.0452
      D       K       R       E       F       N       C       M       W       H
   2320    2301    1947    1745    1660    1360    1271     973     892     518
 0.0387  0.0384  0.0325  0.0291  0.0277  0.0227  0.0212  0.0162  0.0149  0.0086
Entry numbers: 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35
ENTRY          HUMAN KAPPA LIGHT CHAINS SUBGROUP I
ENTRY          HUMAN KAPPA LIGHT CHAINS SUBGROUP II
ENTRY          HUMAN KAPPA LIGHT CHAINS SUBGROUP III
ENTRY          HUMAN KAPPA LIGHT CHAINS SUBGROUP IV
ENTRY          HUMAN LAMBDA LIGHT CHAINS SUBGROUP I
ENTRY          HUMAN LAMBDA LIGHT CHAINS SUBGROUP II
ENTRY          HUMAN LAMBDA LIGHT CHAINS SUBGROUP III
ENTRY          HUMAN LAMBDA LIGHT CHAINS SUBGROUP IV
ENTRY          HUMAN LAMBDA LIGHT CHAINS SUBGROUP V
ENTRY          HUMAN LAMBDA LIGHT CHAINS SUBGROUP VI
ENTRY          MOUSE KAPPA LIGHT CHAINS I
ENTRY          MOUSE KAPPA LIGHT CHAINS II
ENTRY          MOUSE KAPPA LIGHT CHAINS III
ENTRY          MOUSE KAPPA LIGHT CHAINS IV
ENTRY          MOUSE KAPPA LIGHT CHAINS V
ENTRY          MOUSE KAPPA LIGHT CHAINS VI
ENTRY          MOUSE KAPPA LIGHT CHAINS VII
ENTRY          MOUSE KAPPA LIGHT CHAINS MISCELLANEOUS
ENTRY          MOUSE LAMBDA LIGHT CHAINS
ENTRY          RAT KAPPA LIGHT CHAINS
ENTRY          RABBIT KAPPA LIGHT CHAINS
ENTRY          RABBIT LAMBDA LIGHT CHAINS
ENTRY          OTHER KAPPA LIGHT CHAINS
ENTRY          OTHER LAMBDA LIGHT CHAINS
ENTRY          MISCELLANEOUS LIGHT CHAINS
    pos.
ipos name   N  k  n1     S      Vs     Vk  AA's and frequencies
Begin data.
  0     0  25  1  25 0.0000   6.00   1.00 A 25
  1     1 825 11 561 1.6172  18.41  16.18 DEAQNSVKYFH 561 118 52 46 23 14 5 2 2 1 1
  2     2 954 15 623 1.8912  22.26  22.97 IVSAYFLNTPDQEGM 623 121 53 46 38 22 20 9 8 7 2 2 1 1 1
  3     3 961 16 628 1.7381  20.02  24.48 VQALEMDKTIGHSPRW 628 185 39 29 22 16 14 9 7 3 2 2 2 1 1 1
  4     4 971  6 510 1.3948  15.78  11.42 MLVIPT 510 385 44 30 1 1
  5     5 965 11 928 0.3241   7.51  11.44 TSIAKLMNPQV 928 15 12 3 1 1 1 1 1 1 1
  6     6 902  3 890 0.1075   6.46   3.04 QEV 890 11 1
  7     7 931 11 522 1.7544  20.24  19.62 STPDEAQFINR 522 243 97 24 19 16 6 1 1 1 1
  8     8 940 12 773 1.1447  13.27  14.59 PASTEHQRVGIM 773 45 39 30 25 14 5 4 2 1 1 1
  9     9 919 13 380 2.2144  27.85  31.44 SALGTKPDFINRV 380 285 118 46 33 18 18 9 6 2 2 1 1
 10    10 790 15 497 1.8534  21.68  23.84 STIFPLVYEAGKMNW 497 113 67 56 22 11 8 8 2 1 1 1 1 1 1
 11    11 924 11 554 1.8167  21.14  18.35 LVMNTAKEFIQ 554 187 84 30 28 19 15 2 2 2 1
 12    12 912 15 537 1.8764  22.03  25.47 SAPTEYQVLFGDMNR 537 153 122 45 28 8 4 4 3 2 2 1 1 1 1
 13    13 906 12 375 2.0093  24.15  28.99 VALTGEMIDFQS 375 342 55 51 45 26 6 2 1 1 1 1
 14    14 896 11 613 1.5134  17.13  16.08 STAPVFNGIQY 613 122 101 37 7 4 4 3 2 2 1
 15    15 902 10 283 2.1398  26.44  31.87 PVLASIMFTK 283 267 242 61 26 15 3 2 2 1
 16    16 894  5 860 0.2771   7.27   5.20 GSEVR 860 26 3 3 2
 17    17 807 10 239 2.2409  28.36  33.77 EDQGKTNSHL 239 229 168 124 34 7 2 2 1 1
 18    18 859 12 319 2.3781  31.19  32.31 RTKSQPIEGLVM 319 221 134 82 52 30 7 5 4 2 2 1
 19    19 866  7 590 1.1078  12.93  10.27 VAIFGLS 590 241 31 1 1 1 1
 20    20 857 11 645 1.1132  12.98  14.62 TSIRAKVEMNY 645 169 16 10 5 5 3 1 1 1 1
 21    21 833  7 593 1.2756  14.53   9.83 ILMVFRS 593 128 93 9 8 1 1
 22    22 795 13 419 1.5620  17.72  24.67 STNKFAGIDLQRY 419 294 38 30 3 2 2 2 1 1 1 1 1
 23    23 781  1 781 0.0000   6.00   1.00 C 781
 24    24 705 13 351 2.1824  27.23  26.11 RQSKTAIEGDFLV 351 100 94 83 54 8 4 3 3 2 1 1 1
 25    25 713  9 451 1.4335  16.21  14.23 ASGTRFMVI 451 181 60 10 4 2 2 2 1
 26    26 688 10 616 0.7545  10.12  11.17 STDNGARLHI 616 21 17 13 10 4 3 2 1 1
 27    27 652 15 336 2.2104  27.77  29.11 QSEKTGANDHRFPVY 336 93 84 62 47 8 6 4 3 3 2 1 1 1 1
 28   27A 310 10 279 0.7174   9.87  11.11 SGNRTADHKV 279 11 9 4 2 1 1 1 1 1
 29   27B 267  8 133 1.6450  18.77  16.06 LVISFANT 133 96 26 6 3 1 1 1
 30   27C 226 12  84 2.4652  33.13  32.29 LVDYSNEKQART 84 50 34 28 17 4 2 2 2 1 1 1
 31   27D 283 16  70 2.9064  44.98  64.69 SHYGNTDAWQCEFIKR 70 56 46 39 30 19 9 3 3 2 1 1 1 1 1 1
 32   27E 220 10 122 2.0134  24.22  18.03 SAKDNRTGVY 122 37 23 16 15 2 2 1 1 1
 33   27F 116  8  52 2.1576  26.77  17.85 VKGSIRNY 52 28 17 10 3 3 2 1
 34    28 575 15 103 3.2527  57.19  83.74 DNSYTGVHILAFKEM 103 98 97 55 52 39 38 30 28 23 4 4 2 1 1
 35    29 643 16 178 2.8276  42.59  57.80 GISTVKQDANRPELYF 178 172 112 48 33 32 20 12 8 8 6 4 3 3 3 1
 36    30 625 17 158 3.0303  49.02  67.25 SNVKYGIDRHATEFLCQ 158 142 87 64 53 40 20 19 15 7 5 5 4 2 2 1 1
 37    31 590 17 193 2.6165  36.79  51.97 SNTHKDIGYQARELPVW 193 167 108 31 24 14 12 10 8 6 5 4 3 2 1 1 1
 38    32 629 17 428 1.9203  22.71  24.98 YFNSWRDAGHLTECMPQ 428 51 39 27 21 15 12 9 6 5 5 5 2 1 1 1 1
 39    33 618 10 378 1.7356  19.98  16.35 LMVAIPYFGS 378 109 62 44 16 3 3 1 1 1
 40    34 558 16 166 2.7417  40.13  53.78 ANHSYEQGTIDCVKFW 166 133 113 53 26 15 12 11 8 6 5 3 3 2 1 1
 41    35 600  3 598 0.0356   6.15   3.01 WLY 598 1 1
 42    36 543  7 425 1.0872  12.75   8.94 YFVLIHQ 425 71 32 9 3 2 1
 43    37 495  9 410 0.8251  10.63  10.87 QLREHDKTV 410 73 4 2 2 1 1 1 1
 44    38 488  9 447 0.6000   9.09   9.83 QEKHLGPVY 447 18 10 7 2 1 1 1 1
 45    39 508 14 435 1.0040  12.03  16.35 KRHLTYFNVDAEGS 435 28 13 9 4 4 3 3 3 2 1 1 1 1
 46    40 512  9 423 0.9817  11.85  10.89 PSQALTGRF 423 58 9 6 6 5 2 2 1
 47    41 476 10 384 1.0121  12.10  12.40 GDEHKNQRSV 384 65 13 4 2 2 2 2 1 1
 48    42 456 13 226 2.3309  30.19  26.23 QKTHGSEARIFLN 226 67 62 35 28 15 8 5 5 2 1 1 1
 49    43 444  8 159 2.1534  26.69  22.34 SAPLTRCV 159 118 98 36 28 3 1 1
```

(continued)

TABLE IV. Sample Var Output, All Light Chains (continued)

```
 50    44 446  8 378 0.8711  10.97   9.44 PFVINAEL 378 36 23 4 2 1 1 1
 51    45 446 11 318 1.5258  17.28  15.43 KRTQVELAFIN 318 43 42 22 9 4 3 2 1 1 1
 52    46 446 13 302 1.7839  20.66  19.20 LGRPVTSIAFMEH 302 40 36 26 15 12 4 3 2 2 2 1 1
 53    47 437  6 351 1.0203  12.17   7.47 LWVIMT 351 53 19 7 6 1
 54    48 435  6 415 0.3571   7.69   6.29 IVLMPS 415 9 5 4 1 1
 55    49 432 11 360 1.0178  12.15  13.20 YGFSKHNRDEQ 360 39 11 7 6 2 2 2 1 1 1
 56    50 418 15  79 3.3922  63.00  79.37 GDKRAEYLSNWTQHV 79 63 48 42 41 34 33 31 16 7 7 6 5 3 3
 57    51 427 12 183 2.3168  29.89  28.00 ATVDIMNGSFLR 183 113 58 18 18 10 10 8 6 1 1 1
 58    52 410  9 336 1.0507  12.43  10.98 SNTDKARYE 336 41 12 7 7 2 2 2 1
 59    53 401 14 158 2.5597  35.38  35.53 NKTSRQDEYGILAF 158 77 60 45 22 17 4 4 4 3 3 2 1 1
 60    54 424  8 229 1.3035  14.81  14.81 LRQSKEVW 229 176 7 6 3 1 1 1
 61    55 420 16 175 2.8071  41.99  38.40 AEPFQYHGVDKISMTW 175 67 48 37 18 14 13 10 10 8 7 5 5 1 1 1
 62    56 417 13 299 1.5403  17.45  18.13 SPTDAIENGLRVY 299 42 40 13 7 5 3 3 1 1 1 1 1
 63    57 427  8 417 0.2240   7.01   8.19 GDESNTVW 417 2 2 2 1 1 1 1
 64    58 411  5 338 0.7555  10.13   6.08 VITFY 338 68 3 1 1
 65    59 413  5 384 0.4119   7.98   5.38 PSQTV 384 26 1 1 1
 66    60 411 14 136 2.1659  26.92  42.31 ASDVENTHKLPGQY 136 119 117 11 8 6 3 2 2 2 2 1 1 1
 67    61 418  4 415 0.0728   6.31   4.03 RKNS 415 1 1 1
 68    62 449  5 444 0.1096   6.47   5.06 FILQV 444 2 1 1 1
 69    63 446  8 376 0.8567  10.87   9.49 SKTRAGIL 376 42 22 2 1 1 1 1
 70    64 447  5 438 0.1775   6.79   5.10 GASDV 438 4 3 1 1
 71    65 436  9 425 0.2417   7.09   9.23 SGRALQTVY 425 3 2 1 1 1 1 1 1
 72    66 441 12 321 1.4767  16.70  16.49 GKLRNSTAEIMV 321 43 40 18 4 4 4 3 1 1 1 1
 73    67 413  9 357 0.7921  10.39  10.41 SIYAFLDPT 357 39 7 3 2 2 1 1 1
 74    68 423  8 388 0.5780   8.96   8.72 GRASDEKV 388 20 4 4 3 2 1 1
 75    69 400 13 325 1.1530  13.34  16.00 TNDASQKGHIRVY 325 32 20 7 5 3 2 1 1 1 1 1 1
 76    70 423 11 206 2.2624  28.79  22.59 DSKQTEAHGLN 206 79 43 34 26 24 3 3 2 2 1
 77    71 427  8 247 1.5494  17.56  13.83 FAYVLPCI 247 88 83 3 2 2 1 1
 78    72 431  6 250 1.3843  15.66  10.34 TSAIRY 250 135 42 2 1 1
 79    73 430  4 417 0.2189   6.98   4.12 LFPV 417 11 1 1
 80    74 417 10 290 1.6590  18.95  14.38 TKNAIRSEPG 290 52 28 12 8 8 6 5 5 3
 81    75 426  5 417 0.1827   6.81   5.11 IVLST 417 5 2 1 1
 82    76 417 10 297 1.4757  16.69  14.04 STNHDGQFRY 297 50 32 24 6 3 2 1 1 1
 83    77 420 12 118 2.4954  33.83  42.71 SGRPDNTCAEIV 118 115 84 45 27 20 5 2 1 1 1 1
 84    78 421  7 166 1.9065  22.49  17.75 VLMATIQ 166 148 58 42 4 2 1
 85    79 392  7 229 1.2819  14.59  11.98 EQKRLHT 229 144 10 5 2 1 1
 86    80 402 14 154 2.7127  39.33  36.55 APTCSEQVDYGRFN 154 70 48 43 26 25 15 8 4 3 2 2 1 1
 87    81 395  8 302 1.1999  13.78  10.46 EDAGMNVQ 302 50 28 7 3 2 2 1
 88    82 398  3 393 0.1095   6.47   3.04 DNV 393 3 2
 89    83 391 11 120 2.8169  42.28  35.84 AEFLVIDMTSG 120 65 64 43 39 24 11 11 8 5 1
 90    84 418  5 331 0.8896  11.12   6.31 AGTSV 331 76 6 3 2
 91    85 407 12 173 2.3166  29.89  28.23 TVIDMSHENAPY 173 87 55 51 21 7 6 2 2 1 1 1
 92    86 425  3 422 0.0671   6.29   3.02 YFK 422 2 1
 93    87 421  4 313 0.9440  11.54   5.38 YFLH 313 99 7 2
 94    88 428  1 428 0.0000   6.00   1.00 C 428
 95    89 409 16 241 2.1669  26.94  27.15 QASLFMGCHWEKRNPV 241 57 29 23 21 11 8 4 3 3 2 2 2 1 1 1
 96    90 397  9 260 1.8250  21.26  13.74 QLSHGTAVN 260 41 31 20 19 11 9 4 2
 97    91 411 14 112 2.9589  46.65  51.38 WYSGAFHNDLTRQI 112 91 65 51 18 16 14 11 8 7 7 6 4 1
 98    92 405 18  83 3.3306  60.36  87.83 NYSDTGKRALWVEIFHQC 83 73 62 49 42 18 13 12 10 10 8 7 5 4 3 3 2 1
 99    93 401 17 145 3.0323  49.09  47.01 SEHYTNGDRIQALVKMF 145 57 40 38 30 22 19 17 9 6 4 3 3 3 2 2 1
100    94 396 16  66 3.4792  66.91  96.00 SNLVYDTPFIWAGRCM 66 63 48 48 38 28 22 20 17 12 10 8 8 6 1 1
101    95 406 16 257 2.0735  25.25  25.28 PHLSNGTAQYRDEFIV 257 39 31 25 13 9 8 7 4 4 3 2 1 1 1 1
102   95A  59 14  13 3.2193  55.88  63.54 SDNTPGHKQAEFLY 13 10 9 9 4 3 2 2 2 1 1 1 1 1
103   95B  29 11  10 2.9708  47.04  31.90 GVAHEWLMRTY 10 4 3 3 2 2 1 1 1 1 1
104   95C  10  6   3 2.4464  32.70  20.00 DGYFTV 3 2 2 1 1 1
105   95D   5  4   2 1.9219  22.74  10.00 SDNY 2 1 1 1
106   95E   2  2   1 1.0000  12.00   4.00 NV 1 1
107   95F   2  2   1 1.0000  12.00   4.00 EY 1 1
108    96 356 18  67 3.3698  62.02  95.64 WLYRIFVPTGHSANQKCM 67 63 63 34 29 22 19 15 10 7 7 5 4 4 3 2 1 1
109    97 359 11 240 1.5419  17.47  16.45 TVIASGLMPYN 240 78 13 7 7 3 3 3 2 2 1
110    98 366  5 362 0.1087   6.47   5.06 FEILR 362 1 1 1 1
111    99 370  3 367 0.0754   6.32   3.02 GAR 367 2 1
112   100 367  8 221 1.7008  19.50  13.29 GAQSTPRV 221 73 44 13 10 4 1 1
113   101 363  3 361 0.0548   6.23   3.02 GDP 361 1 1
114   102 361  3 357 0.0969   6.42   3.03 TSQ 357 3 1
115   103 357 11 293 1.0916  12.79  13.40 KERTNQMDGHY 293 32 12 6 4 4 2 1 1 1 1
116   104 343  3 246 0.8827  11.06   4.18 LVG 246 96 1
117   105 323  9 196 1.5555  17.64  14.83 ETVDSILNQ 196 82 29 10 2 1 1 1 1
118   106 333  8 136 1.7312  19.92  19.59 IVLMKNFS 136 118 70 3 2 2 1 1
119  106A  83  4  80 0.2816   7.29   4.15 LQTV 80 1 1 1
120   107 325  8 233 1.2470  14.24  11.16 KGRSTELN 233 68 11 6 3 2 1 1
121   108 251  6 168 1.3504  15.30   8.96 RQGCVW 168 49 29 2 2 1
122   109 147  4  63 1.8760  22.02   9.33 PTDA 63 35 27 22
End data.
```

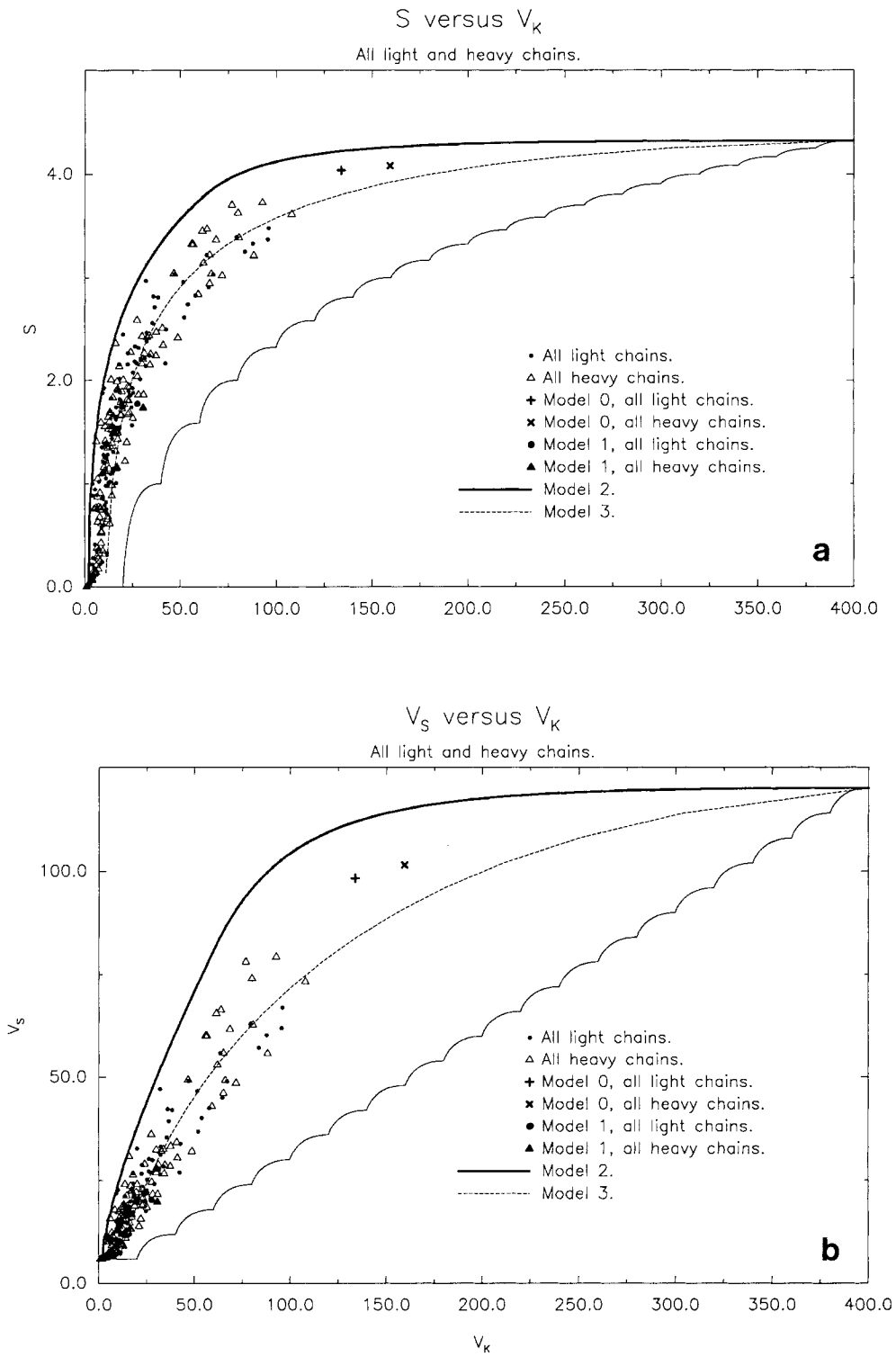


Fig. 1. Scatter plots. **a.** S vs. V_K. **b.** V_S vs. V_K. The points for Model 1 lie near (30, 1.8) in Figure 1a and near (30, 20) in Figure 1b. See text for details.

equal probabilities of occurring. This model allows V_K to vary over its feasible range. When Model 2 is worked out in detail, it predicts that S or V_S should lie on the upper boundary exhibited in Figure 1. This makes sense, because S is maximized whenever the p_i are equal, and the model makes all the p_i

beyond p_1 equal; however, this result does not mimic the data. Model 2 makes a range of V_K appear, but it predicts that S and V_S will be considerably greater than the values actually observed, given any V_K value. This model leaves unexplained the origin of the assumed range of variabilities, just as Model 1 left unexplained the origin of the grand rank-frequency distribution.

A Correlative Model

It is possible to imagine more elaborate models similar in spirit to those just described, but these would likely share similar flaws. Therefore, we now turn to the sequence data in order to determine which sorts of site occupancy patterns do, in fact, occur among the immunoglobulins, and which do not. We were led to the examination of rank-frequency distributions (p_i as a function of i) from the following consideration. Suppose the rank-frequency distributions at all sites exhibited a similar, predictable form. Then, if we knew p_1, we could predict the other p_i. The minimal occupancy information required for the calculation of V_K could then be used to predict the additional occupancy information used in calculating the S-based measures. In fact, a sufficient condition for this predictability is that sites with similar variabilities exhibit similar rank-frequency type distributions.

Figure 3a and c illustrates p_i vs. i for data sets ALC and AHC, respectively. It is clear that the distributions are similar for similar values of S and change systematically as S changes. These trends are more clearly visible in Figure 3b and d, in which the p_i axis is logarithmic.

Figure 4a,b,c illustrates the same data exhibited in Figure 3b, except that weighted linear least-squares lines fitted to the data points are also shown. The S range has been divided into three segments for ease of visualization. The nearly horizontal sets of data points that appear at high i in many positions consist of several types with counts (n_i) of unity. Since the regression weights of these points are therefore low, they may lie far from the corresponding lines. Even so, the correlation coefficients of these lines are high. There are 38 regression lines shown in Figure 4c, only three of which exhibit correlation coefficients[14] with absolute values less than 0.9. The correlation probability,[14] which represents the probability that the observed value of $|r|$ would be exceeded in a sample of k points drawn from an uncorrelated population, is <0.01 for all the lines shown, and is <0.001 for all but four of them; typical values are of order 10^{-5}. The AHC data set exhibits similar statistics, and the less variable positions in both sets continue to be this well correlated down to S values of about unity, which corresponds to V_S and V_K values of about 12. Below this, the correlation becomes somewhat weaker, exhibiting typical correlation probabilities of .01.

We now show that the simple assumption of log-linearity of the rank-frequency distributions at immunoglobulin sites provides a semiquantitative explanation for the observed relationship between V_K and the S-based measures.† We start with the three relationships

$$p_i = p_1 e^{\alpha(i-1)}. \tag{7}$$

$$\sum_{i=1}^{k} p_i = 1. \tag{8}$$

$$S_e = -\sum_{i=1}^{k} p_i \ln p_i. \tag{9}$$

Equation (7) is the assumption of log-linearity of rank-frequency diagrams, and Equation (9) defines an entropy measure based on the natural logarithm; it is easiest to use natural logarithms in what follows and later revert to our previous definitions.

From (7) and (8) we have

$$p_1^{-1} = \sum_{i=1}^{k} e^{\alpha(i-1)} = \frac{1-e^{\alpha k}}{1-e^{\alpha}} \tag{10}$$

where the second equality comes from the identity

$$\sum_{i=1}^{k} x^{i-1} = \frac{1-x^k}{1-x}.$$

It is also convenient to use the analog of (6):

$$k^* = e^{S_e}. \tag{11}$$

Note that k^* will have the same value whether calculated from Equation (6) or from Equation (11).

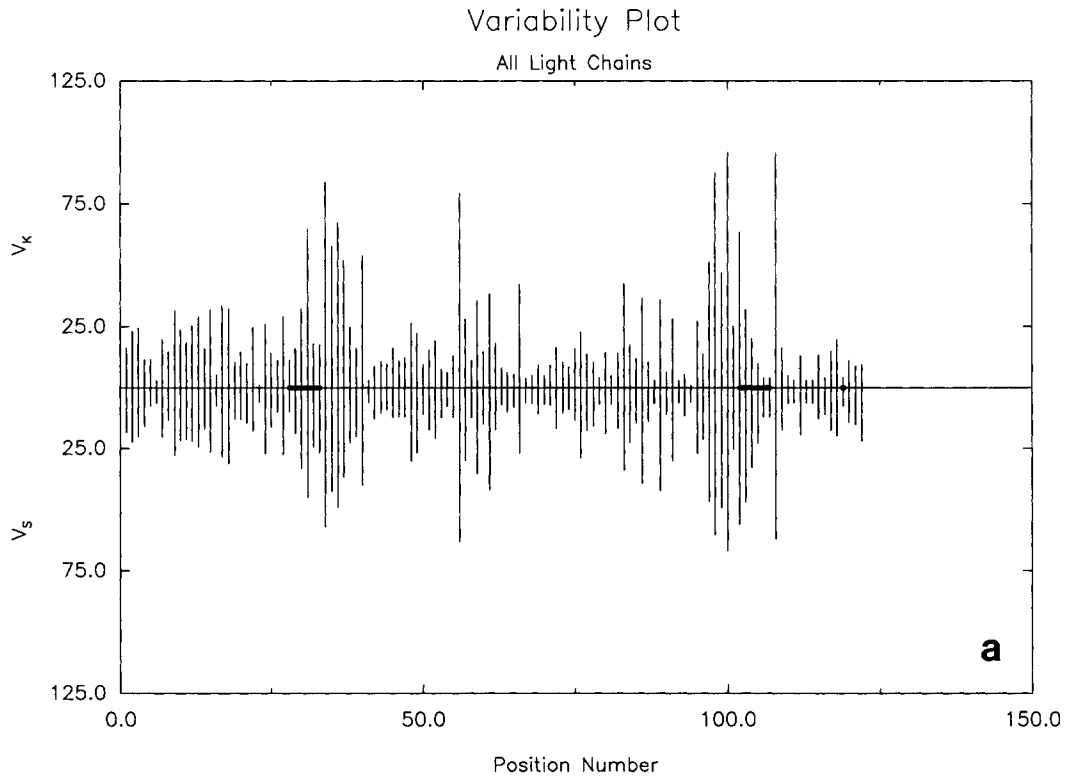
Substitution of (7) into (9) gives

$$\begin{aligned} S_e = \ln k^* &= -\ln p_1 - \alpha p_1 \sum_{i=1}^{k} (i-1) e^{\alpha(i-1)} \\ &= \ln p_1^{-1} - \alpha p_1 \frac{dp_1^{-1}}{d\alpha} \\ &= \ln \frac{1-e^{\alpha k}}{1-e^{\alpha}} - \frac{\alpha[(k-1)e^{\alpha(k+1)} - ke^{\alpha k} + e^{\alpha}]}{(1-e^{\alpha k})(1-e^{\alpha})} \end{aligned} \tag{12}$$

where the last two equalities come from the two parts of (10).

For proteins, k^* lies in the range [1, 20]. For any value of k^*, the possible values of k lie in the range [k^*, 20,] and for each value of k, feasible values of p_1 lie in the range [/k, 1). Now, if the p_i are related to each other according to Equation (7), a site with a given value of k^* (hence of S_e) may take on any value of k in the feasible range, but for each such value of k, p_1 and α will be uniquely defined. For a given k^*, the value $k = k^*$ is associated with equal occupancies of the types, so that $p_1 = 1/k$ for all i and $\alpha = 0$. For $k > k^*$, we will have $p_1 > 1/k$ and α

†We are indebted to an anonymous referee for suggesting this line of analysis.



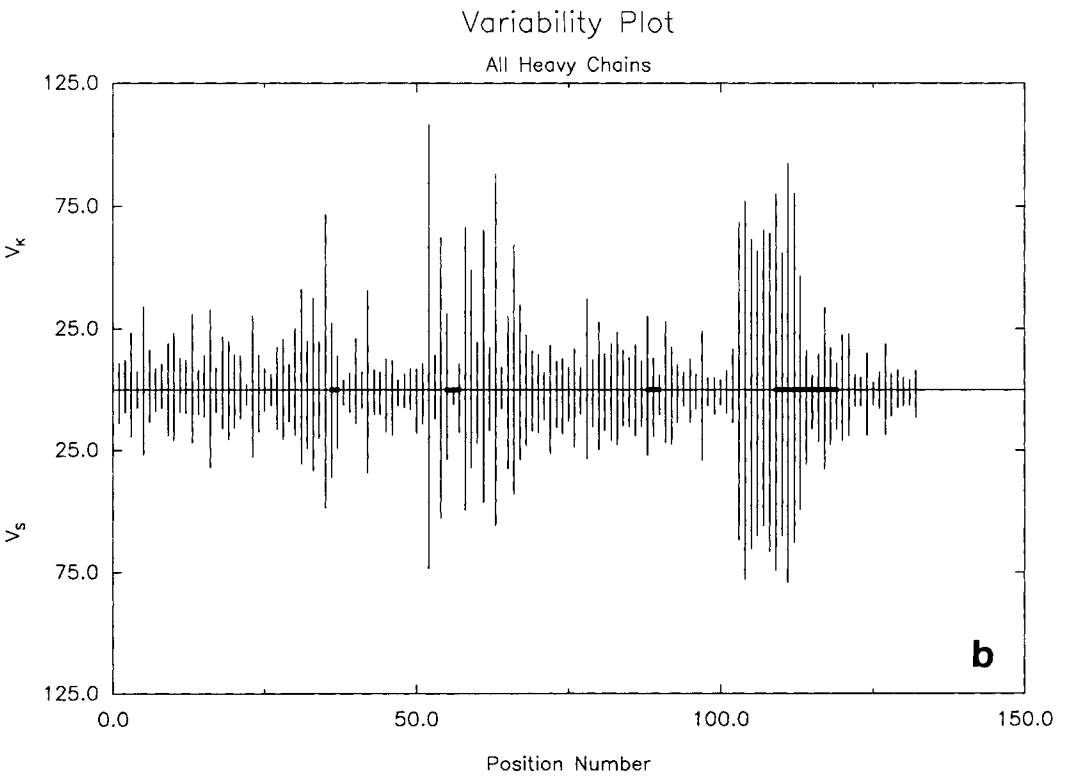


Fig. 2a,b. Legend appears on page 307.

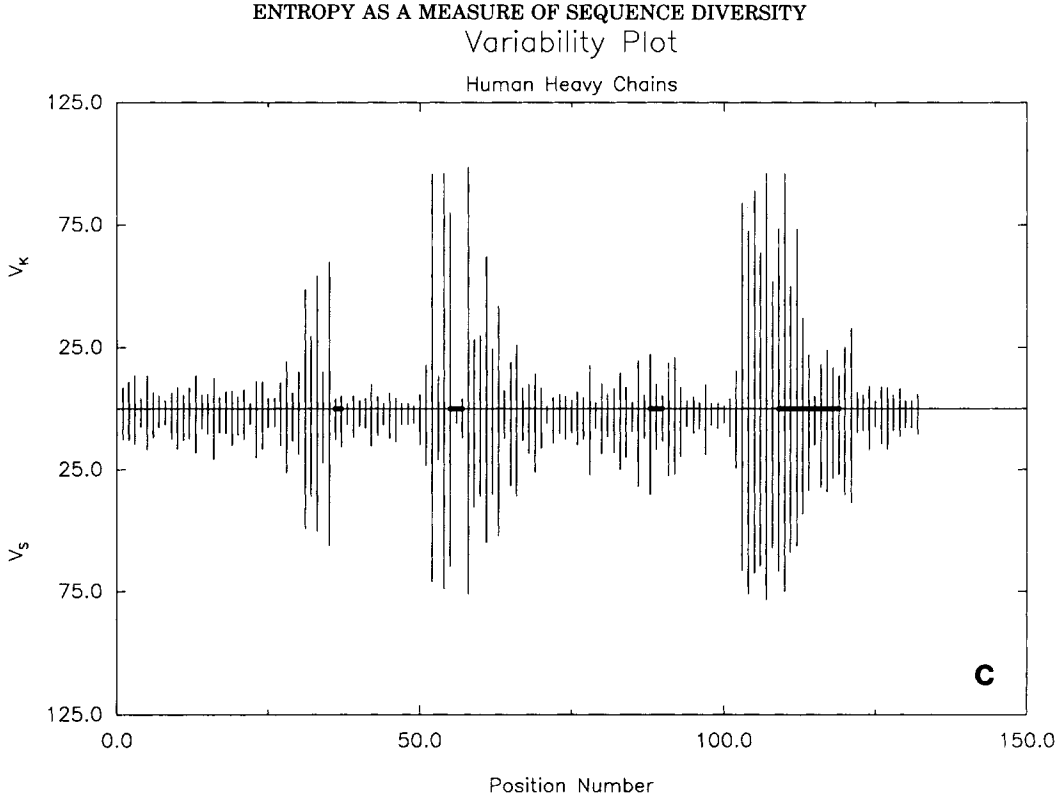


Fig. 2. Variability plots, V_K and V_S vs. sequence. **a.** All light chains. **b.** All heavy chains. **c.** Human heavy chains. The heavy lines along the x-axis indicate the locations of "lettered" positions (insertions) in the compilation of Kabat et al.[3]

< 0. Once we obtain k and p_1 we can calculate V_K, and thus investigate what constraints the assumption of Equation (7) places upon the populated region of Figure 1. This is Model 3.

The simplest way to envision the calculation is as follows. If k* is specified, Equation (12) defines an implicit function between k and α; thus, if k* and k are specified, α can be found numerically. Equation (10) can then be used to calculate p_1. In practice, we used a somewhat altered procedure that took advantage of our initial knowledge of the range of p_1. Given k* and k, which were sampled systematically, we first searched for two values of p_1 whose corresponding values of α bracketed ln k* when inserted into the right-hand side of Equation (12).

This search was conducted as follows. For increasing values of p_1 in [1/k, 1), $\alpha(p_1, k)$ was found by solving Equation (10), in the form $x = 1 - p_1(1 - x^k)$, by successive substitution. Upon convergence, α was taken as ln x. α was then used to evaluate Obj(α, k), an objective function obtained by subtracting ln k* from the right hand side of Equation (12). The bracketing values of p_1 were obtained by noting the change in sign of Obj(α, k) as p_1 approached unity. These bracketing values were used to initiate a binary search for the root, using an algorithm described by Press et al.,[16] section 9.1. Within the binary search routine we continued to evaluate the objective function via the intermediate determination of α values.

Note that k* determines S, through the relationship $S = \log_2 k^*$, and V_S, through the relationship $V_S = 6k^*$. For each value of k*, the V_K values obtained for the various values of k were averaged. The resulting functions, S and V_S vs. $V_{K,avg}$, are plotted in Figures 1a,b. The Model 3 curves fit the data well. Furthermore, the average value of V_K/k^* for V_K values between 1 and 135, which encompasses the observed range, is predicted, based on the Model 3 calculations, to be 7.3. This is reasonably close to the value of six found empirically and embodied in the definition of V_S. There are no adjustable parameters in this model: the shape of the Model 3 curves and the predicted V_K/k^* ratio are consequences solely of the assumption that the rank-frequency distribution of amino acid types at any position is log-linear, as observed in Figure 4 and expressed in Equation (7).

Fig. 3. Rank-frequency distributions. **a.** p_i vs. i, all light chains. **b.** $\log_{10} p_i$ vs. i, all light chains. **c.** p_i vs. i, all heavy chains. **d.** $\log_{10} p_i$ vs. i, all heavy chains. Each line represents a single position; the positions are sorted by S.

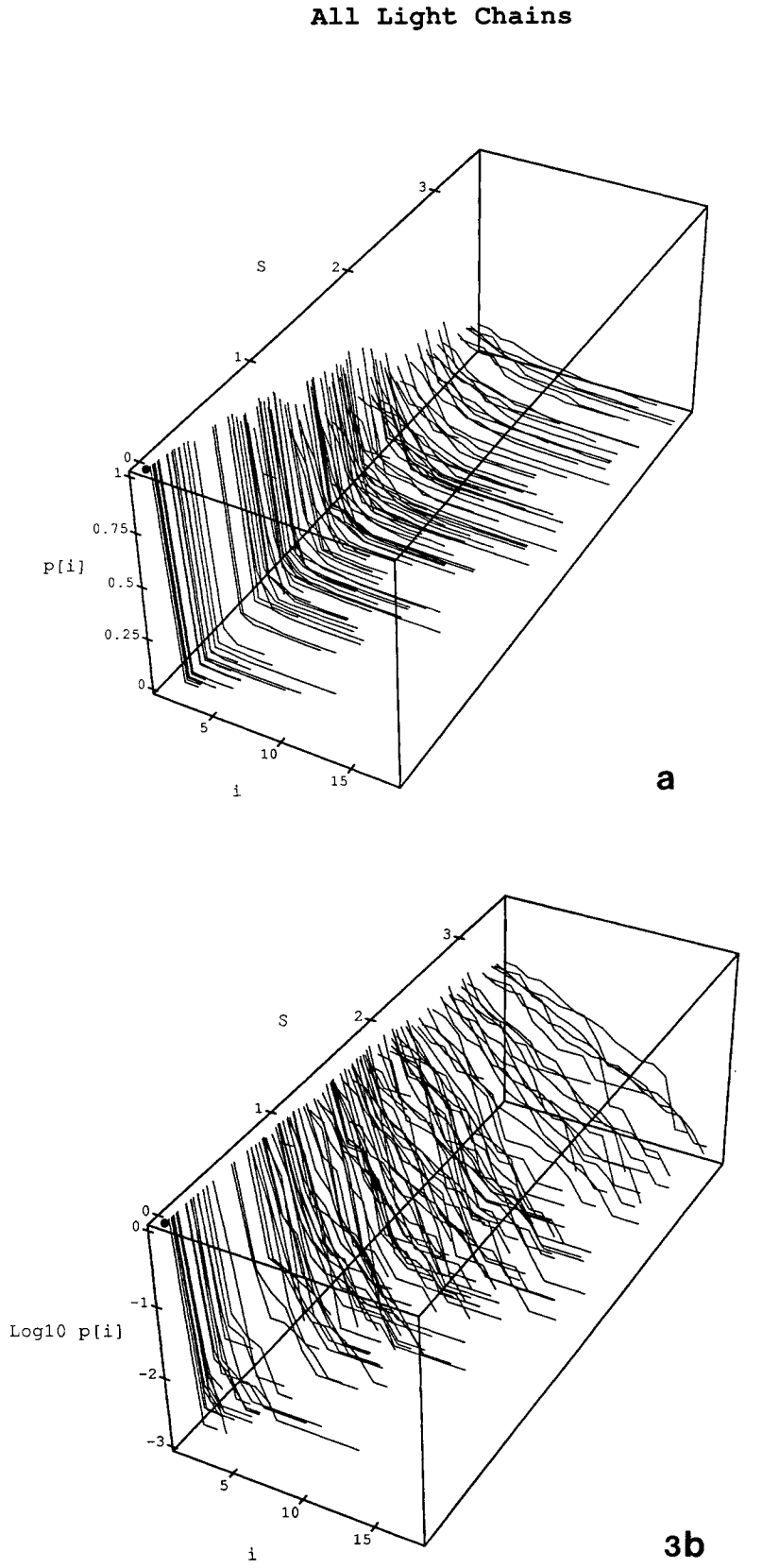


Fig. 3a,b. Legend appears on page 307.

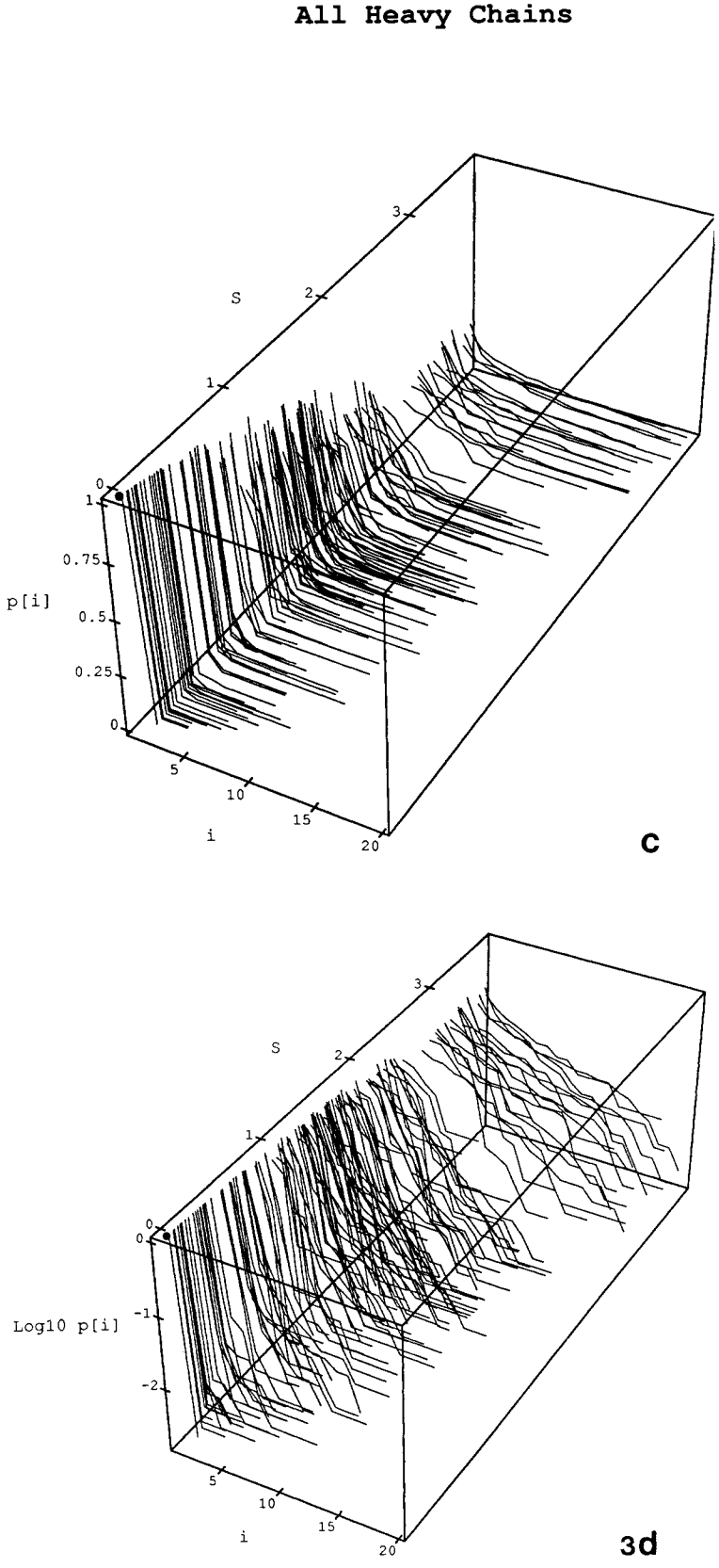
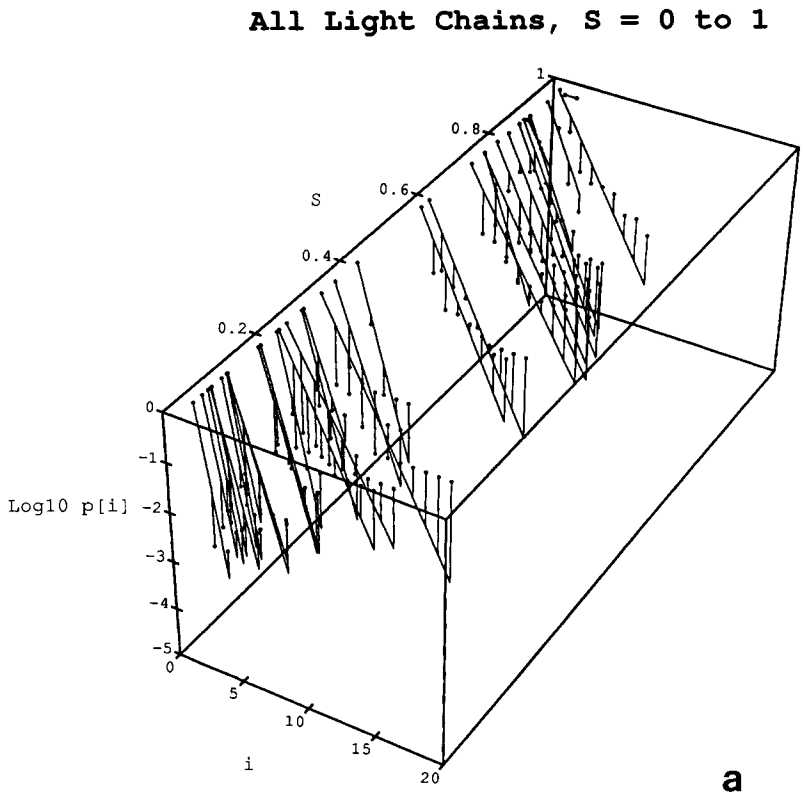


Fig. 3c,d. Legend appears on page 307.



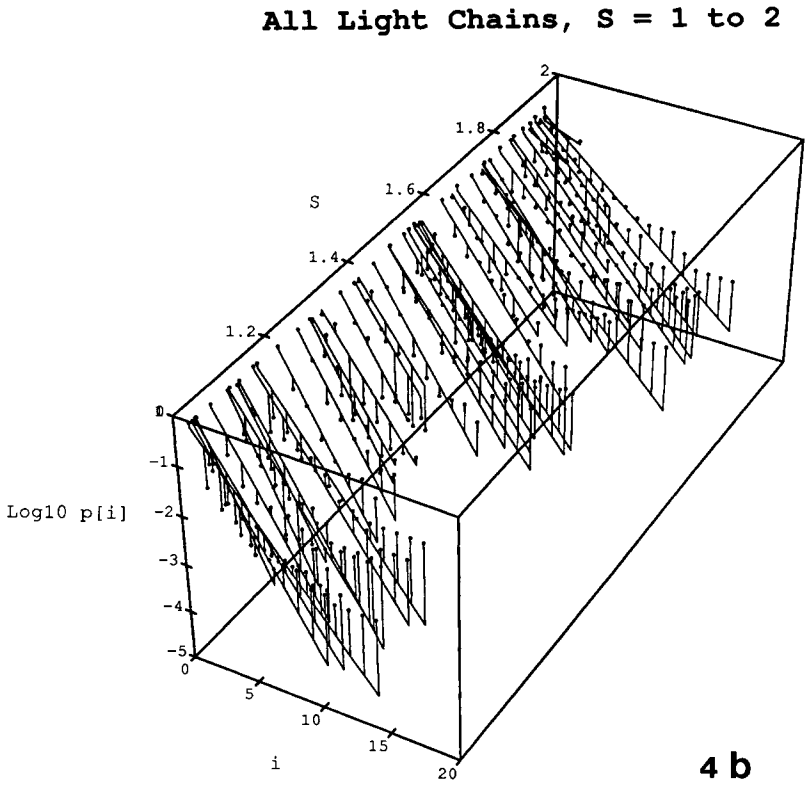


Fig. 4a,b. Legend appears on page 311.

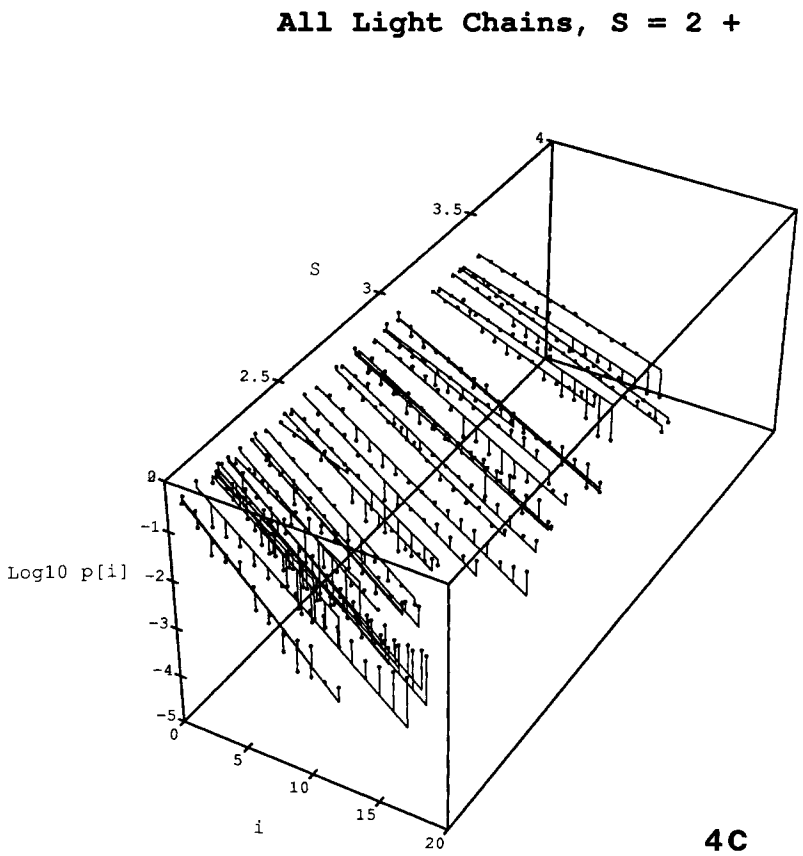


Fig. 4. Rank-frequency distributions. Weighted least-square fits to $\log_{10} p_i$ vs. i, all light chains. These are the same data as represented in Figure 3b, except that here each data point is connected to its regression line by a vertical line segment, and the S scale is expanded. **a.** $0 < S \leq 1$. **b.** $1 < S \leq 2$. **c.** $S > 2$.

CONCLUSIONS

Variability Measure

Despite the great insights into immunoglobulin structure and function afforded by Wu and Kabat's definition of variability,[6] we conclude that the information-theoretical entropy, S, as defined by Equation (2), is a better statistical measure of the "spread" of a distribution of amino acid types at a given site than is V_K. S exhibits desirable mathematical properties not exhibited by V_K. It also has an intrinsic meaning which is well correlated with what we intuitively mean when we say that a site is variable, and it affords a definite interpretation, some of whose implications we have explored.

Protein and nucleic acid sequences are accumulating at a fast rate, one that is bound to accelerate under the encouragement of the various genome projects. Although the goal of these projects is sometimes stated as the sequencing of a genome, in fact the idea of a genome sequence is more appropriate to an individual than to a species. One would like to know where genome sequences vary among individuals, where they vary among species, and by how much. To the extent to which this variability can be captured by a single number, there is much to recommend the information-theoretical entropy as the measure of choice. S can be used to measure variability in sequences of nucleic acids or codons as readily as in protein sequences.

While this work was in progress, we became aware of unpublished work by Dr. I. Pardowitz of the Max-Planck Institute for Experimental Medicine, Gottingen (personal communication). Dr. Pardowitz has also proposed the Shannon entropy as a measure of the variability of a position in a set of aligned sequences, as we have. He has used the S measure to analyze nucleotide sequences, rather than protein sequences, and has shown that the minimization of S can be used as a criterion both for sequence alignment and for the sorting of sequences into binary trees that reflect similarity. When this is done, it is possible to parse the total information in a given set of sequences into contributions from knowing the best consensus sequence, from knowing the binary tree, and from knowing the individual sequences in detail.

Rank-Frequency Distributions

We noted that S and V_K are highly correlated for the immunoglobulins, and we have traced this correlation to the fact that the function p_i (i) tends to be log-linear. We do not know why this relationship holds for the immunoglobulins, or whether it holds for other protein families. If the immunoglobulins are unique in this regard, then the high degree of correlation which they exhibit between V_K and the S-based measures is unlikely to persist in other families.

The appearance of this relationship is reminiscent of Zipf's law.[18] Zipf found that many phenomena, such as the rank-frequency distribution of various parts of speech and the sizes of manufacturing enterprises, follow a log–log curve with a characteristic slope of negative one; however, some of his examples (Chinese characters and Gothic root morphemes) exhibit log-linear rank-frequency diagrams. Zipf's attempt to explain these distributions is qualitative rather than quantitative in nature and is difficult to state precisely enough to test; however, two possible origins for the phenomena we have observed come to mind.

First, it is possible that the systematic variation of the p_i represents a thermodynamic equilibrium; that is, that it reflects an underlying Boltzmann distribution. If this assumption is made, that is, if we assume

$$\frac{p_i}{p_j} = e^{-\Delta G_{ij}/RT} \tag{13}$$

where ΔG_{ij} is the free-energy difference between states i and j, it is easy to show that Equation (7) implies that the energy levels are equally spaced. In this interpretation, one can either view all the sites as exhibiting the same set of levels (in which case the more variable positions are characterized by higher temperatures), or one can view the temperature as constant (in which case the more variable sites have the more closely spaced levels).

In the most literal interpretation of this picture, what we have been calling S corresponds to a thermodynamic entropy, k* corresponds to a partition function, and the ΔG represent real free-energy differences in some sort of mean-field sense, that is, averaged, for a given site, over all the possible occupancies of the other sites. This energy could imaginably be a structural energy, a free energy of formation from biological precursors, or even a free energy of binding to a "mean-field antigen." This interpretation is reminiscent of the work of Bryant and Lawrence,[19] who were able to show that, despite the fixed arrangement of charged residues in any single native protein, the spatial disposition of charged residues in a large ensemble of proteins of known structure exhibits a distribution reflecting the well-known modifications of Coulomb's law in common use in protein modeling. The thermodynamic hypothesis, however, does not explain the log-linearity of our observed rank-frequency distributions; we can think of no reason to believe that the energy levels associated with type substitutions should in general be equally spaced.

Another possibility entirely is that the origin of these distributions is dynamic, rather than energetic. The sequences in the immunoglobulin database reflect both evolution in the large sense and the history of exposure of typical (we hope!) individuals to antigens. It is possible—even likely—that the statistics of appearance of types at positions is dominated by the mechanisms involved in evolution and expression, rather than by equilibrium energetics. Jerne's network theory of the immune response[20] seeks to account for the antibody repertoire of a single individual, not the pooled repertoire of many individuals or even many species. According to this theory, however, the levels of specific antibody types arise from a complex set of interactions involving other antibodies as well as the individual's history of antigen exposure. Thus, for a single individual, at least, Jerne's theory would seem to favor a dynamic, rather than an energetic explanation of the distributions of amino acid types in hypervariable domains.

The two sorts of explanation could be related. Yano and Hasegawa,[21] Volkenshtein,[22] and, more recently, Schneider[23] have discussed the question of whether the sequence entropy tends to increase with time under the action of evolutionary dynamics, in analogy to the second law of thermodynamics.

We should also point out that site variability in proteins can have two different origins or "meanings." The variability of the CDR sequences of the immunoglobulins is the result of evolutionary selection: the CDRs "need to be" variable in order for the immune system to function, or, to put it differently, a mechanism for the expression of variability has proved useful for higher organisms. On the other hand, high variability can also come about as a result of evolutionary neutrality: some positions presumably exhibit a broad range of amino acid types because it matters little what happens there.

Recent molecular genetic studies (e.g., the well-known work of Lim and Sauer[24]) indicate that proteins are able to accommodate, both structurally and functionally, a far greater variety of mutations than occur naturally. This seems to us to argue against a thermodynamic or "equilibrium" picture, though it could be that the most variable sites approach equilibrium and that the conserved sites are in some sense kinetically trapped. In any case, the phenomenological statistics of variability as well as the dominating mechanism (energetic or dynamic) governing it could depend on both the nature of the variability (selected for or neutral) and the context of its generation (evolutionary or developmental).

We note that the immunoglobulins are expressed by unique mechanisms, and we are as reluctant to assume that what we have found will apply to other protein families as we are to claim a complete understanding of what we have observed.

ACKNOWLEDGMENTS

We would like to thank T.D. Schneider, T.T. Wu, I. Pardowitz, Stephen H. Bryant, and Charles E. Lawrence for helpful conversations, and the Pew Foundation, the National Institutes of Health (GM44336), and Barnard College for support.

REFERENCES

1. Shenkin, P.S., Yarmush, D.L., Fine, R.M., Wang, H., Levinthal, C. Predicting antibody hypervariable loop conformation. I. Ensembles of random conformations for ringlike structures. Biopolymers 26:2053–2085, 1987.
2. Fine, R.M., Wang, H., Shenkin, P.S., Yarmush, D.L., Levinthal, C. Predicting antibody hypervariable loop conformation. II. Minimization and molecular dynamics studies of MCPC603 from many randomly generated starting conformations. Proteins 1:342–362, 1986.
3. Kabat, E.A., Wu, T.T., Reid-Miller, M., Perry, H.M., Gottesman, K.S. Sequences of Proteins of Immunological Interest. 4th Ed. Bethesda, Maryland: U.S. Department of Health and Human Services, 1987.
4. Brucolleri, R.E., Haber, E., Novotny, J. Structure of antibody hypervariable loops reproduced by a conformational search algorithm. Nature 335:565–568, 1988.
5. Chothia, C., Lesk, A.M. Canonical structures for the hypervariable regions of immunoglobulins. J. Mol. Biol., 196: 901–917, 1987.
6. Wu, T.T., Kabat, E.A. An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. J. Exp. Med. 132:211–249, 1970.
7. Shannon, C.E., Weaver, W. "The Mathematical Theory of Communication." Urbana: University of Illinois Press, 1949.
8. Garnier, J., Levin, J.M. The protein structure code: What is its present status? Comp. Appl Biosci. 7:133–142, 1991, and references cited therein.
9. Stormo, G.D. Computer methods for analyzing sequence recognition of nucleic acids. Annu. Rev. Biophys. Chem. 17:241–263, 1988.
10. Schneider, T.D., Stormo, G.D., Gold, L.D., Ehrenfreucht, A. Information content of binding sites on nucleotide sequences. J. Mol. Biol., 188:415–431, 1986.
11. Berg, O.G., von Hippel, P. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. J. Mol. Biol., 193: 723–750, 1987.
12. Berg, O.G., von Hippel, P. Selection of DNA binding sites by regulatory proteins. II. The binding specificity of cyclic AMP receptor protein to recognition sites J. Mol. Biol., 200:709–723, 1988.
13. Jores, R., Alzari, P.M., Meo, T. Resolution of hypervariable regions in T-cell receptor β chains by a modified Wu–Kabat index of amino acid diversity. Proc. Natl. Acad. Sci. U.S.A. 87:9138–9142, 1990.
14. Bevington, P.R. "Data Reduction and Error Analysis for the Physical Sciences." New York: McGraw-Hill, 1962.
15. Davidson, N. "Statistical Mechanics." New York: McGraw-Hill, 1962.
16. Segal, D.M., Padlan, E.A., Cohen, G.H., Rudikoff, S., Potter, M., Davies, D.R. The three-dimensional structure of a phosphorylcholine-binding mouse immunoglobulin Fab and the nature of the antigen combining site. Proc. Natl. Acad. Sci. U.S.A. 74:4298–4302, 1974.
17. Press, W.H., Flanner, B.P., Teukolsky, S.A., Vetterling, W.T. "Numerical Recipes." Cambridge: Cambridge University Press, 1986.
18. Zipf, G.K. "Human Behavior and the Principle of Least Effort." New York: Hafner, 1972.
19. Bryant, S.H., Lawrence, C.E. The frequency of ion-pair substructures in proteins is quantitatively related to electrostatic potential: A statistical model for nonbonded interactions. Proteins 9:108–119, 1991.
20. Jerne, N.K. Idiotypic networks and other preconceived ideas. Immunol. Rev. 79:5–24, 1984.
21. Yano, T., Hasegawa, M. Entropy increase of amino acid sequences in protein. J. Mol. Evol. 4:179–187, 1974.
22. Volkenshtein, M.V. Mutations and the value of information. J. Theor. Biol. 80:155–169, 1979.
23. Schneider, T.D. Information and entropy of patterns in genetic switches. In: "Maximum-Entropy and Bayesian Methods in Science and Engineering," Vol. 2. Erickson, G.J., Smith, C.R. (eds.) Dordrecht: Kluwer Academic Publishers, 1988: 145–154.
24. Lim, W.A., Sauer, R.T. Alternative packing arrangements in the hydrophobic core of λ repressor. Nature 339:31–36, 1989.