Title

Covariability of V3 Loop Amino Acids

Journal

AIDS Research and Human Retroviruses, 12(15)

ISSN

0889-2229 1931-8405

Authors

BICKEL, P. J.
COSMAN, P. C.
OLSHEN, R. A.
et al.

Publication Date

1996-10-10

DOI

10.1089/aid.1996.12.1401

Peer reviewed

AIDS RESEARCH AND HUMAN RETROVIRUSES
Volume 12, Number 15, 1996
Mary Ann Liebert, Inc.

Covariability of V3 Loop Amino Acids

P.J. BICKEL,[1] P.C. COSMAN,[2] R.A. OLSHEN,[3] P.C. SPECTOR,[1] A.G. RODRIGO,[4] and J.I. MULLINS[4]

ABSTRACT

We reanalyzed for covariability a set of 308 human immunodeficiency virus type 1 (HIV-1) V3 loop amino acid sequences from the B envelope sequence subtype previously analyzed by Korber *et al.*,[1] as well as a new set of 440 sequences that also included substantial numbers of sequences from subtypes A, D, and E. We used the measure employed by Korber *et al.*, essentially the likelihood ratio statistic for independence, plus two additional measures as well as clade information to examine the new set and both data sets simultaneously. We set forth the following conclusions and observations. The eight most highly connected sites identified through these statistical approaches included all of the six residues previously shown to have determining roles in structure, immunologic recognition, virus phenotype, and host range; each of the seven pairs of covariant sites found by Korber were signaled by our additional two measures in the set of 308 sequences, although 2 or 3 dropped out of the examination of the set of 440 when the requirement of stringent significance was applied for some or all of the three tests, respectively; using the same criteria, a total of 20 (including 5 Korber *et al.* pairs) or a total of 6 (including 4 Korber *et al.* pairs) were found when the set of 440 was added. Several limitations to statistical analysis of this type of HIV sequence data were also noted. For example, the data sets were, by historical necessity, collected haphazardly. For example, it was not possible to separate substantially sized groups out according to time of or since infection, disease status, antiviral treatment, geography, etc. There was also an enormous "wealth of significance" within the data. For example, for one measure the 440 data set showed 233 of the 465 pairs of sites with a likelihood ratio statistic of <0.001. Last, most sites had consensus amino acids in 80% or more of the sequences; hence, there was an absence of data on many combinations of amino acids. Given the observed linkage between sites shown to be covariable and those known to have critical biological function, the statistical approaches we and Korber *et al.* have outlined may find use in predicting critical structural features of HIV proteins as targets for therapeutic intervention.

INTRODUCTION

IT HAS BY NOW BEEN KNOWN for years that the envelope gene is one of the most variable parts of the HIV genome. Its V3 loop region has been sequenced and studied intensively in view of its immunogenicity and functional importance. Korber *et al.*[1] analyzed a set of 308 DNA sequences encoding the 31 V3 loop amino acids from the 1991 AIDS database[2] whose provenance is described in Ref. 1. Their goal was to identify pairs of sites where mutations would "with high confidence be identified as covarying." They advanced a set of seven pairs of covarying sites that seemed to merit further analysis. The covariation they observed statistically could be the result of biochemical interactions between the sites—constraints of protein structure of functional relation driven by selection, which are processes that one would wish to uncover. However, it was recognized that the statistical covariation could be the result of phylogenetic effects, "an evolutionary heritage from distinct founder viruses." That is to say, a group of sequences might be largely descendants of a single ancestral virus, and the appearance of a strong covariation between two sites might simply reflect that there was insufficient time to achieve much divergence in the independent evolution of those sites. To this should be added that the 1991 database is not a random sample from the population

[1]Department of Statistics, University of California, Berkeley, Berkeley, California 94720-3860.
[2]Department of Electrical and Computer Engineering, University of California, San Diego, San Diego, California 92093-0407.
[3]Departments of Health Research and Policy, Electrical Engineering, and Statistics, Stanford University, Stanford, California 94305-5092.
[4]Departments of Microbiology and Medicine, University of Washington, Seattle, Seattle, Washington 98195-7740.

of HIV viruses. We expect biases from several other sources. A possible source is unknown epidemiological clustering where several patients are infected by the same individual. An attempt was made to eliminate known epidemiological clustering. Also, within patients there can be substantial variability of viruses as a function of length of infection and of disease state. In the gathering of these databases, there was no careful effort made to ensure that sampling from different groups of individuals would be representative in terms of geography, disease status, treatment with antivirals, etc. In these various ways, covariability can be an artifact of sampling. Nevertheless, several of the pairs that were identified had been observed to covary in functional ways *in vitro*.[3–10] For a review of this work and, more generally, HIV sequence variability, see Ref. 11.

Our work can be seen in part as a follow-up to Ref. 1. Thus,

1. We reanalyze their 308 sequences with different statistical tools and examine what answers other measures of covariability give.
2. We analyze a new set of 440 sequences (provided by B. Korber, L.A.N.L. and Santa Fe Institute) with the same set of 31 residues from the 1993 AIDS database, using their measures and ours.
3. We see to what extent covariation persists and what new pairs of sites appear statistically covariant in the new set.

We also develop tools to explore interaction between groups (triples, quadruples, etc.) of sites. In particular, we explore

1. The extent to which particular sites are critical to interactions
2. The existence of cliques (sets of more than two sites that appear to act in concert)

We propose

1. Some broad conclusions on the Korber *et al.*[1] and our methodologies as applied to these data sets
2. Some covariable pairs that have stood a variety of tests and thus bear examination for biological function
3. Two new measures of covariability, one of which is of a type generalizable to assessing linkage disequilibrium

In particular,

1. We note that all but one of the six sites signaled as significant by Korber *et al.* and four to five of their seven pairs are singled out as significantly covariable by our new data and criteria.

2. We argue that the Korber *et al.* criteria are too extreme, potentially ruling as not significantly perfectly covariable pairs. Our criteria do not have this feature but still have a degree of arbitrariness. For a list of 20 distinguished pairs we require less significance than Korber *et al.* on at least one of three measures, coupled, however, with requirements that significance be present in both data sets and separately in clades (that are discussed below). We also focus on a sublist of six pairs for which significance by all three measures is realized.

3. We note biological evidence of importance and covariability for six of the eight sites on our short list. Details are given in the section Biological Correlates. However, our statistical techniques are employed purely as a hypothesis-generating procedure. In other words, the methods seek to highlight covariation between sites and not to explain the biological relevance of the covariation. As discussed below, to the extent that data are available on the biological significance of covarying sites, our methods do in fact identify these sites. But we indicate in our discussion why any statistical measures of covariability for data sets such as these can only be considered as pointers to possible biological activity. If it were possible to remove the various biases, for example owing to unknown stage of infection, epidemiological linkage, and the impact of antiretroviral therapies, then indeed a significant result statistically could only be due to either direct biochemical interaction of the sites, interaction of the sites with some selective pressure in the environment that acts in a correlated way at both sites, or random chance, which should only occur with probability equal to the significance level. However, given the limitations of these data sets, such conclusive results are currently not possible. But our goal in this work is also to introduce statistical techniques that may be useful to any researchers interested in sequence covariation. Such issues will become increasingly susceptible to analysis as better data become available, e.g., from patients repeatedly sampled over time. A supplementary analysis of these data using a modification of a clustering method that has proven successful in data compression and other engineering applications[12] will appear elsewhere. That approach to clustering is closely related to the CART[13] algorithms for classification and regression.

The set of 440 sequences consists in part of 364 "nonembargoed" sequences from the set of 410 sequences described in Ref. 14. As described there, these constitute a mix of single sequences from an individual when only one was available, a randomly chosen one if only two were available, and a consensus if more than two were available. Experiments in which sequences were drawn from individuals who were known to be epidemiologically linked and who had genetically similar sequences were represented by only one sequence. To these 364 were added 76 sequences consisting of single sequences from unlinked individuals taken early in their infections.[15] This set of 440 and the 308 sequences studied by Korber *et al.*[1] had 152 in common (identical for the 32 residues considered). However, as will become evident, the two sets differed in many ways. Sequences from the LaRosa *et al.*[16] set were not included in the 440. Clades *A*, *C*, and *E*, are represented by 135 sequences in the 440 but did not exist in the 308; and the distributions of residues at many sites are very different. For instance, at position 24 the consensus amino acid (with 24%) of the 308 is aspartic acid, whereas in the 440 (with 41%) it is glutamic acid. This is not surprising. The epidemic is dynamic; and, of course, neither of these sets can be viewed as a random sample from the population of HIV viruses extant on or before 1991 and 1993, respectively. We discuss these matters further after we present our methodology and findings.

METHODOLOGY

Our approach, as in Ref. 1, starts with covariation. Our first step is to isolate pairs of sites i, j that appear to covary signif-

icantly. We use three statistics for this purpose. The first is the information theory-based M_{ij} of Korber *et al.*[1] Our second statistics, G_{ij}, was developed by Goodman and Kruskal[17] on the basis of a statistic introduced by L. Guttman. It has been applied with success in the social sciences. The third statistic, P_{ij}, focuses on covariability of a single pair of residues at a pair of sites. In statistical language, M_{ij} is the likelihood ratio statistic for testing the hypothesis of independence of two sites against arbitrary covariability. In information theory language M_{ij} is the mutual information at sites i and j. Qualitatively, M_{ij}, which is never negative, is large if any of a number of particular pairs of residues at i and j are favored relative to what would be expected from their marginal frequencies by chance (if i had nothing to do with j). M_{ij} is approximately equivalent for large sample sizes to the familiar Pearson's chi-square statistic χ^2_{ij} for testing independence. We used M_{ij} rather than χ^2_{ij} for comparability with Korber *et al.* G_{ij} is the reduction in the chance of guessing incorrectly that knowledge of the residue at i gives in guessing the residue at j and conversely. If i and j have nothing to do with each other, then $G_{ij} = 0$. Finally, in statistical language, $P_{ij}(a,b)$ is the likelihood ratio statistic for testing the hypothesis of independence of sites i and j against the alternative that a pair of residues (a,b) is favored relative to what would be expected from chance. $P_{ij}(a,b)$ has been used in the genetics literature to study linkage disequilibrium, where attention focuses on a particular pair of alleles.[18] P_{ij} is the largest of the $P_{ij}(a,b)$. Qualitatively, P_{ij}, like M_{ij}, tends to be large if there is covariance at i, j; but it focuses on situations where only one particular pair of amino acids exhibits covariance (although we do not know which pair that is). For instance, it is perfectly suited to picking out situations where if one site does not have the consensus amino acid, it is very likely that the other site will correspondingly be "forced" also to have a nonconsensus amino acid. The particular pair of amino acids a,b making P_{ij} largest can be viewed as playing the role of pairs having the highest specific information but the definition is different—see the Appendix.

Each of these measures focuses on a different aspect of covariability. As applied to our data, they frequently "light up" together.

Here are the definitions. For a set of N aligned sequences of the same length let

$$\hat{p}_{ij}(a,b) = \text{(number of sequences with residue } a \text{ at site } i, \text{ residue } b \text{ at site } j)/N$$

Then

$$M_{ij} = \sum_{a,b} \hat{p}_{ij}(a,b) \log\left[\frac{\hat{p}_{ij}(a,b)}{\hat{p}_i(a)\hat{p}_j(b)}\right]$$

where

$$\hat{p}_i(a) = \text{(number of sequences with residue } a \text{ at site } i)/N$$

For reference,

$$\chi^2_{ij} = N \sum_{a,b} [\hat{p}_{ij}(a,b) - \hat{p}_i(a)\hat{p}_j(b)]^2/\hat{p}_i(a)\hat{p}_j(b)$$

The statistic G_{ij} is given by

$$G_{ij} = \frac{1}{2} \frac{\sum_a \hat{p}_{ij}(a, \max) + \sum_b \hat{p}_{ij}(\max, b) - \hat{p}_i(\max) - \hat{p}_j(\max)}{1 - 1/2\,[\hat{p}_i(\max) + \hat{p}_j(\max)]}$$

where $\hat{p}_{ij}(a, \max) = \max_b \hat{p}_{ij}(a,b)$, $\hat{p}_i(\max) = \max_a \hat{p}_i(a)$, $\hat{p}_{ij}(\max, b) = \max_a \hat{p}_{ij}(a,b)$. As we mentioned, this awkward-looking quantity has a very nice interpretation. If we were asked to predict the residue at i with information only about frequency of residues at i, we would use the residue giving the modal (consensus) frequency $\hat{p}_i(\max)$. If we knew the residue b at j, then we would use the residue giving the modal (consensus) conditional frequency $\hat{p}_{ij}(\max, b)/\hat{p}_j(b)$. G_{ij} gives the average reduction in the chance of guessing incorrectly that knowledge of i gives for j and knowledge of i for i.

The statistic P_{ij} is given by

$$P_{ij} = \max_{a,b} P_{ij}(a,b)$$

where $P_{ij}(a,b)$ is the M_{ij} statistic obtained by replacing the 20-letter alphabet at site i by $a,\bar{a}$ and at site j by $b,\bar{b}$, where $\bar{a}$ is "not a" and $\bar{b}$ is "not b." While, as we stated, $P_{ij}(a,b)$ has appeared before in the literature,[18] to the best of our knowledge, P_{ij} is new here.

We use M, G, and P to create lists of sites that are candidates for covariability as follows:

1. For each data set of N sequences under consideration, the 308 of Ref. 1, the 440, and the envelope sequence subtypes (clades, see below), we generated 100,000 pseudo data sets of N sequences each by independently permuting the amino acids at each site. Compared to the "real" data set, these pseudo data sets have the same marginal probabilities for the amino acids at each site, but the information about covariation between sites, if any, is not maintained.

2. For the V3 loop of length 31 amino acids there are $(31 \times 30)/2 = 465$ pairs of sites to consider. For each of the pseudo data sets and each pair of sites i,j the corresponding M_{ij} is computed. For each pair of sites i,j, we count the number n_{ij} of such M_{ij} that are equal to or exceed the M_{ij} observed in the original data set. We examine the fraction $p_{ij} = n_{ij}/100{,}000$. The motivation for this is that if there were only chance variation (independence) between sites i and j, then the distribution of values of M, obtained from "real" data sets or from the pseudo data sets should be the same. Thus, we would not expect the "real" M_{ij} to be an extreme value among the 100,000 values. If n_{ij} is the number of permuted M_{ij} values that exceed the "real" one, then $(n_{ij} + 1)/(100{,}000 + 1)$ is the attained one-sided significance level for testing the hypothesis of independence of sites i and j against arbitrary covariability. Note that p_{ij} can be 0—the "real" M_{ij} can be larger than any of the 100,000 generated M_{ij}. This would correspond to a significance level of $(10^5 + 1)^{-1}$. If there were only chance variation between i and j we have observed something extremely rare. The same process is carried out for G_{ij} and P_{ij}. In the following, p_{ij} denotes the fraction $n_{ij}/100{,}000$ of pseudo scores that exceed the real score,

and it is also loosely used to denote the significance level $(n_{ij} + 1)/(100{,}000 + 1)$ of that event.

We use this methodology, as did Korber *et al.*, for obtaining significance probabilities, rather than the chi-square approximation for M_{ij} and similar approximations derivable for G_{ij} and P_{ij}, for two reasons:

1. It is well known that in the case of M_{ij} (and χ^2_{ij}) the chi-square approximation can be poor for sparse tables and large values of the statistics even if the sample size N is large.[19]
2. We are making statements about many pairs at the same time. The scheme we describe enables us to obtain measures of the simultaneous validity of these statements. Of course, we still need to note that even if everything is happening according to chance, the chance of at least one statement being wrong is much higher than the chance of a particular statement being wrong. That is the reason for working at significance levels such as 1/100,000 for the full set of sites.

RESULTS

Initial list of covariable pairs

At this point the following phenomena can be noted.

1. There is a huge amount of "statistical structure." A very large number of site pairs have $p_{ij} = 0$ for one of M, G, P (Table 1). Histograms of the p_{ij} for G and the 440 data set reveal that on the order of 50% of the pairs have $p_{ij} \leq 0.001$. That is, at most 100 of the 100,000 pseudo data sets gave a value as large as or larger than the observed! The value 0.001 is usually taken as an adequate level of statistical significance. It ensures that, on the average, only 1 time in 1000 will we call something significant when in fact it happened by chance alone. However, if a statistical test yields a false claim of significance only 1 time in 1000, still, if we run the test thousands of times, sooner or later there will be false claims. This is the problem of making calls of significance *simultaneously* for many pairs. In our case, we examine 465 pairs. Calling 50% of the pairs covarying is unreasonable since we expect that some of those claims arise simply from this multiplicity of tests. Still, this is a remarkably large number of covarying pairs; if no sites were truly covariable, then chance alone would not be expected to produce a single pair with $p_{ij} \leq 0.001$! It seems reasonable to restrict ourselves to the site pairs for which $p_{ij} = 0$. Since each has a significance level of 10^{-5}, the chance that we make a false covariability call for *any* of the 465 pairs is $<10^{-5} \times 465$, that is to say, <0.005. Thus a restriction to this group means we have a simultaneous 0.005 significance level.

TABLE 1. NUMBER OF SITE PAIRS WITH $p_{ij} = 0$

Data set	M	P	G	MPG[a]	M *or* P *or* G
308	63	58	52	36	81
440	134	125	117	95	152

[a]*MPG* means *M* and *P* and *G*.

Korber *et al.* arrived at 7 pairs of sites for the 308 data set as follows. They used 1,000,000 rather than 100,000 pseudo random data sets and required that, for a pair of sites i,j to be declared covariable, M_{ij} must be larger than the largest value observed among the pseudo random data sets, not only for i,j but for all other pairs of sites as well. A motivation for this is the desire to avoid having any false-positive pair on the list. If all sites had identical compositions, then the criterion of requiring that a pair should score higher than the random scores generated from all pairs would make sense. Since they made simultaneous significance statements for 465 pairs of sites, using 1,000,000 sets assured them that their chances of declaring significance where there was none were no greater than about 0.0005. We used 100,000 data sets generally to save on computing time, since it seemed to make no difference in the final sets of candidates we proposed. For example, with 1,000,000 pseudo random data sets our figures for M and the 440 data set in Table 1 would change from 134 to 121, and all the pairs we included for their M significance would still be included.

2. As Table 1 shows, from 44 to 62% of the pairs of sites that have $p_{ij} = 0$ for G or M or P have $p_{ij} = 0$ for all three measures. We can draw the comforting conclusion that a substantial proportion of the pairs that covary tend to do so as measured in any of these three intuitively plausible although hardly distinct ways.

3. By all three measures the amount of structure revealed by the 440 is considerably greater than what is revealed by the 308. This is not simply the familiar story of more effects detected with larger samples. Twice the number of pairs called significant by M or G were still observed when the 440 were reduced to the 288 sequences that did not appear among the 308. This is consistent with other observations[11,20] of the increasing complexity of the viral quasispecies, in view of the fact that the 440 contained mainly sequences acquired between 1991 and 1993.

Lists of the site pairs with $p_{ij} = 0$ counted in Table 1 and indeed software with which we did nearly all computations are available from the authors (e-mail: spector@stat.berkeley.edu).

Pruning the list of covariable pairs

When the modal frequency (frequency of the consensus amino acid) at either member of a pair of sites is high, the statistics G, P, M that quantify covariability tend to have small values that are sensitive to changes in a small number of sequences. In particular, highly significant values can be a consequence of alignment errors or even typographical errors in a few sequences. Korber *et al.*[1] met this issue by their tremendously restrictive criterion that their observed M values be larger than anything seen in the 1,000,000 pseudo random data sets at any pair of sites. We argue in the Appendix that this is too restrictive. Indeed, we begin by restricting our subsequent analyses to site pairs whose observed G or M or P statistics have $p_{ij} = 0$. This restriction helps but is insufficient to eliminate pairs for which changes in a few sequences could create large changes in significance—see the Appendix. Some progress comes from adding the requirement, as we do, that the numerical value of the test statistic be in the top quartile of its set of values for the data set in question. This is, of course, in the same spirit as Korber *et al.*'s restriction to the maximum

being exceeded for all pairs of sites.[1] However, it is far weaker, excluding at least some bizarre tables. For instance, association of site 0 (in which residue C is known not to vary) with site 17 would owe to what may be a typographical error in which two sequences are ascribed Q at 0. This association is excluded. We also restrict to pairs satisfying these criteria for *both* data sets. Our logic is that if two sites covary significantly in both the 308 and the later 440, this should point more toward fundamental biological covariation.

The resulting list of 60 pairs for which significance was found by any of the three criteria is given in Fig. 1. In Table 2a we give the 23 sites involved and their connectivities, the numbers of sites to which they are significantly connected. We view this as our basic, too-large list of suspects, but which we nonetheless offer for examination by those with other evidence. Subsequently, we discuss some site pairs that appear to be significant in one but not the other data set.

One possibility for pruning this list is to consider the 21 pairs that had $p_{ij} = 0$ for *G and M and P*, again for both data sets. Our logic here is that being signaled by *M* and *G* and *P* is a strong indication of structure. The weak point is that, as statistics, *G*, *M*, and *P* each point to departure in senses that were described in our discussion of methodology. And any departure could come from fundamental biology. This list and the corresponding 13 sites are also given in Fig. 1 and Table 2b.

TABLE 2. NUMBER OF CONNECTIONS AND SITES FOR SIGNIFICANT PAIRS: 308 AND 440

No. of connections	*Site*
a. *M* or *P* or *G*	
15	11
10	4, 19
8	5, 15
7	9, 10, 30
6	24
5	7, 12, 17, 26
4	18, 25
3	28
2	6, 21, 23, 31
1	8, 13, 27
b. *MPG*	
7	11
5	4, 9, 15
4	5, 19
2	7, 10, 17, 24, 26
1	12, 18

Covariable pairs within clades

A variety of other possibilities exist for pruning the initial large list down to a size that is of use to experimentalists. However, one observation leads us to a different approach. The wealth of observed covariation may in large part be due to phylogenetic effects. The characteristics of a shared ancestor themselves presumably sometimes reflect functional or structural significance. This would not be true for regions of the genome that are under no selection. Even when it is true, covariability that persists even when shared ancestry is partially accounted for seems more likely to be due to functional or structural factors. Korber, Myers, and others,[2,14] on the basis of a phylogenetic analysis based on long (883) site stretches of the *env* gene, have produced largely consistent trees ending in seven clades: *A, B, C, D, E, F*, and the aberrant subgroup *O*. The geographical clustering of these clades is consistent with the history of the epidemic. We are not entirely comfortable with the precision of the fine detail of these or any other phylogenies for HIV. Nevertheless, it seems reasonable that the effect of shared ancestry can be reduced by examining covariability within clades determined by similarity over long stretches of the HIV genome. Korber *et al.* took this factor into account in part of their analysis by their classification of amino acid pairs (in their Table 1) as interesting according to both "predictiveness" and "frequency." We divided the 440 sequences into the same clades and repeated our analyses for the three statistics within each. It is possible, of course, that there are subclades within the clades—that some sequences within a clade share a common ancestor that is closer than the purported ancestor of the entire clade. Attempts to subdivide the clades in this way would result in too few sequences for any remaining analysis of covariation. Even as it is, the number of sequences per clade varies from 248 for *B* to 8 for *O*. For the *C, F*, and *O* groups (which are not present in the 308) our ability to detect covariability (the power of the test based on G, M, or P) is reduced. We therefore limited ourselves to clades *A, B, D*, and *E* since *A, D*, and *E* have 40–50 sequences per clade. Within that group we arbitrarily noted pairs in our lists of 60 and 21 pairs, respectively, that had, for one of our statistics, values of $p_{ij} \leq 0.01$ in *B*, and had $p_{ij} < 0.05$ in at least one of clades *A, D*, and *E*.

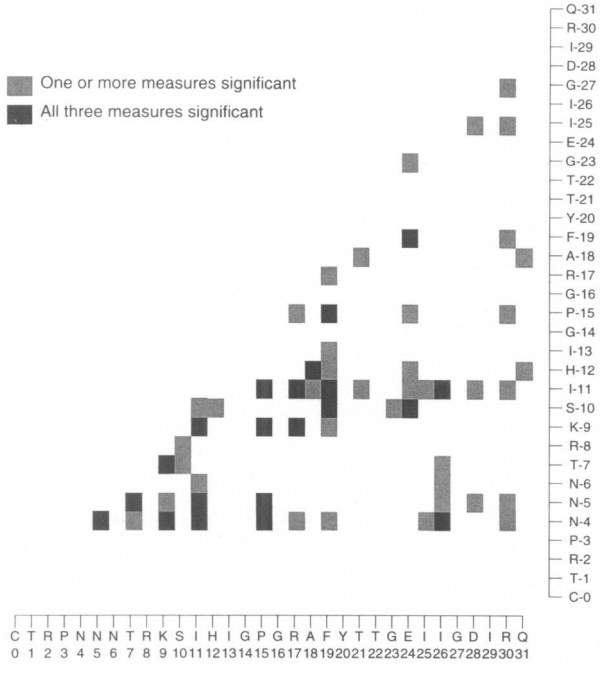

FIG. 1. Significant connections for both data sets, with the constraint that to be significant not only must $p_{ij} = 0$, but also the $\sqrt{}$value of the statistic must lie in its upper quartile for the data set.

Figure 2 gives:

1. The 6 pairs and 8 sites that are G and M and P significant and continue to be significant throughout the clades
2. The 19 pairs and 15 sites that are G or M or P significant and continue to be significant throughout the clades.

In Table 3b we give the P_{ij} statistic and the pair of residues yielding the maximum value of $P_{ij}(a,b)$ as well as the consensus pair of residues for each of the 20 pairs. The rest of our discussion refers primarily to the pairs of Fig. 2.

Biological correlates

There are interesting biological correlates for the most highly connected sites. Numbers 5 and 7 are potential N-linked glycosylation sites and have been implicated in immune escape by Davis *et al.*[3] Substitutions at sites 10 and 12 have been implicated as determinants of cell tropism.[7–10] Korber *et al.*[1] and de Jong *et al*[6] linked simultaneous mutations at site 10 in conjunction with mutation at the block of sites 21–24 to conversion from a non-syncytium-inducing, low-replicating to a syncytium-inducing, high-replicating phenotype. Chesebro *et al.*[7] showed that a single change at site 12 from S to H created noninfectious virus and that altering site 12 in conjunction with sites 20–29 caused a phenotype switch from T tropic to macrophage tropic. Ghiara *et al.*[21] found no substitution for phenylalanine possible at site 19 in a crystallographic study in which substitutions at sites such as 10 that strongly covary with 19 are forbidden. It is gratifying that all six of these sites—5, 7, 10, 12, 19, and 24—appear in our shortest list of eight sites.

Among the pairs in Fig. 2, only two pairs, 5–28 and 12–19, were not significant for all three statistics in at least one data set. Both of these were signaled by M only, which suggests that the Korber *et al.* statistic is the most sensitive of the three.

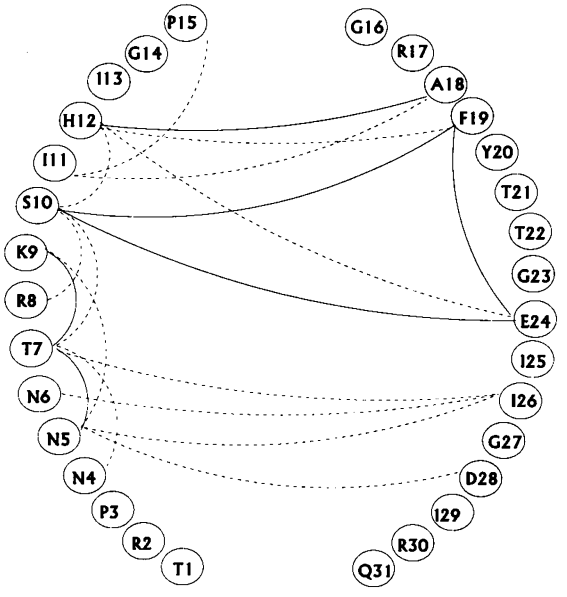

FIG. 2. Significant connections. (—) Significant for all three statistics, both data sets, and in clades as indicated in text. (– – –) Significant for at least one statistic, both data sets, and in clades as indicated in text.

An analysis of cliques

A *clique* is a set in which each site covaries with the other clique members. Table 3 presents one clique of four sites {10,12,19,24}, and two cliques of three sites {5,7,9} and {5,7,26}. At this point we address such questions as, "Is the observed covariability between 19 and 24 possibly an artifact (spurious correlation) of actual covariability between 10 and 19 and 10 and 24?" Such questions may be answered by fitting a second-order log-linear model to the three-way contingency table corresponding to the three sites. This can be done (see the Appendix) provided there are sequences that exhibit all possible pairs of residues for each pair of sites. But there are not. For instance, both A and Y appear repeatedly for 10 and 19, respectively; but they never appear together. To reduce these problems we can distinguish less finely between residues by creating a few broad categories into which we group the 20 amino acids. A natural possibility is to classify according to the four types of side chains (positive, negative, polar, nonpolar) or to the two classes (hydrophobic, hydrophilic). Another possibility is to use the data to reduce our alphabet. The P statistic provides us with pairs of residues that with extreme covariation may or may not be the consensus (modal) pair. For each site we can use an alphabet consisting of the consensus amino acid, other amino acids that have been members for that site of a pair of amino acids maximizing $P_{ij}(a,b)$, and all amino acids without this property lumped together. Typically, this amounts to creating classes out of the most frequent and next most frequent amino acid at a site and lumping others together. We applied these types of analyses to the three cliques above.

Analysis of sites 10, 12, 19, and 24 with a data determined alphabet (given in the Appendix) revealed that, for the 308, the 10–19, 12–19, and 10–24 connections had $p < 0.01$; 12–24 had $p = 0.05$; while the 19–24 connection was not significant. On the other hand, for the 440, 10–19, 10–24 and 12–24 had $p < 0.001$, while 10–12 had $p = 0.03$; and 12–19 was not significant. The only clear indication is the persisting strength of the 10–19 and 10–24 connections.

Analysis of sites 5, 7, and 9 for the 308 gave 5–7 a $p < 10^{-4}$ for 308 and 440, with 7–9 equally extreme for the 440 but not the 308, and 5–9 mildly significant for both.

Analysis of sites 5, 7, and 26 both gave strong readings for 5–7, 5–26 with the 440 and only mild significance for 5–26 in the 308, and 7–26 in both. The second-order analyses here suggest that the observed higher order interactions may be genuine.

Reconciliation with Korber et al.

All of the pairs suggested as significantly covarying by Korber *et al.*[1] in the 308 sequences were signaled as such by G and P as well. But 3 of these, 23–24, 12–24, and 12–23, were not signaled by G and P in the 440. One, 12–23, was also not signaled by M and, if the criterion larger than M at any pair of sites is applied, the 23–24 connection is also eliminated. Using our criterion of M or G or P in both data sets and significance in the clades, we found that the five pairs not involving site 23, namely 10–12, 10–24, 12–18, 12–24, and 19–24, are retained. As we noted earlier, de Jong *et al.*[6] found biological covariation between site 10 and a group of sites including both 23 and 24. Our analysis suggests that it may only

TABLE 3. THE SIGIFICANT PAIRS: 308 AND 440, *M* OR *P* OR *G* AND IN CLADES

a. Number of Connections and Sites

No. of connections	*Sites*
5	7, 10
4	5, 12
3	19, 24, 26
2	9, 11, 18
1	4, 6, 8, 28

b. *P* Statistics, Significance, Determining Pair of Residues, and Consensus Pair[a]

Pair	*P308*	*sig. 308*	*P440*	*sig. 440*	*308 pair*	*440 pair*	*Consensus pair*
4–7	0.11405	0.00000	0.140577	0.00000	Y–T	Y–T	N–T
5–7*	0.103002	0.00000	0.234929	0.00000	N–T	N–T	N–T
5–9	0.114984	0.00000	0.096428	0.00000	N–Q	N–K	N–K
5–26	0.057838	0.00094	0.256108	0.00000	N–I	N–I	N–I
5–28	0.041566	0.00678	0.055771	0.00001	N–D	N–D	N–D
6–26	0.078047	0.00001	0.108245	0.00000	N–I	N–I	N–I
7–26	0.050493	0.00168	0.166351	0.00000	T–I	T–I	T–I
7–9	0.093647	0.00000	0.175587	0.00000	T–K	T–K	T–K
7–10	0.068062	0.00004	0.117923	0.00000	R–H	T–R	T–S
8–10	0.052759	0.00136	0.117297	0.00000	R–S	R–S	R–S
10–12*	0.091336	0.00000	0.076898	0.00000	S–N	G–H	S–H
10–19*	0.155750	0.00000	0.153428	0.00000	G–V	S–F	S–F
10–24*	0.111502	0.00000	0.157668	0.00000	R–Q	S–D	S–E,D
11–15	0.230370	0.00000	0.210741	0.00000	T–P	T–P	I–P
11–18	0.107030	0.00000	0.045988	0.00008	L–R	V–T	I–A
12–18*	0.368264	0.00000	0.241172	0.00000	T–V	T–V	H–A
12–19	0.068594	0.00014	0.029221	0.01187	R–F	R–F	H–F
12–24	0.096354	0.00000	0.036567	0.00230	R–R	H–E	H–E,D
19–24*	0.091741	0.00000	0.123356	0.00000	F–R	F–D	F–E,D

[a](1) The six *M* and *G* and *P* pairs are indicated by asterisks; (2) P308 and P440 are the values of *P* for the 308 and 440 data sets; (3) sig. 308 and 440 are the significance probabilities for *P* in these data sets; (4) 308 and 440 pairs are the residue pairs corresponding to the maximal P_{ij}. Consensus pair is the consensus pair with a switch in site 24 from the 308 to the 440.

be the 10–24 connection that really matters. It should also be noted that the signaling criterion used by us, although not as extreme as that of Korber and Myers (which is less than 10^{-6}), is still very demanding. The *p* value for *M* at 12–23 in the 440 is 2×10^{-5}!

To sum up, the restrictive *M* and *P* and *G* significance for both data sets and clades criterion leads to the retention of four of the seven Korber *et al.* pairs. It adds the glycosylation pair 5–7 and also 10–19. The less stringent *M* or *P* or *G* significance for both data sets and clades criterion leads to the retention of the five pairs cited above. In addition, 15 pairs are added that, among other things,

1. Complete the 10–12–19–24 clique
2. Bring in connections of the glycosylation sites 5 and 7 not only to neighboring sites on the left side of the loop but also to site 26 on the right side

In Fig. 3 we exhibit the 20 pairs that had $p_{ij} = 0$ for *G* or *M* or *P* in the 308 but not the 440. The only sites not appearing in Fig. 2 are 1, 20, and 22, although new connections appear; and, of course, the two Korber *et al.* pairs eliminated by the 440 set are here.

New covarying pairs in the 440

In Fig. 4 we give the 83 pairs in the 440 that had $p_{ij} = 0$ for *G* or *M* or *P*, but did not appear in the 308. Sites not appearing in both the 308 and 440 are 1, 20, 22, and 29. Sites 1, 20, and 22 appear to be coupled with different partners in the 308 and 440 and thus figure as "new" sites of interest in both. All these sites have connectivities of 1 or 2 only and do not seem worth pursuing.

DISCUSSION

The statistical analysis we have made is both limited and somewhat inconclusive. Although we point to some aspects of V3 loop variation that currently generalize, on the whole it supports the pessimistic views of Wain-Hobson[20] regarding the difficulty of finding persistent features of the genotype of the virus or at least of the V3 loop.

Perhaps the greatest limitation of the analysis, as we noted earlier, is that the data are haphazard rather than representing samples drawn from the population of HIV viruses extant up to 1991 and 1993, respectively. As Korber *et al.*[1] noted, this

can introduce a number of biases: "Founder virus effects," unknown epidemiological clustering, unrepresentative sampling from different groups of individuals in terms of geography, disease status, treatment with antivirals, etc. On top of this is the dynamic nature of the epidemic, with the virus evolving in response to immune challenges as it spreads into new populations. Thus, any statistical tools we use, significance probabilities, estimates of interaction strength in second-order log-linear models, are used as a guide to importance and not as confirmation.

A second major limitation is the "wealth of significance" that has caused us (and Korber *et al.*) to prune on the basis of fairly arbitrary thresholds in order not to have too many candidate sites to consider. As we have noted, for G in the 440 data sets more than 233 of the 465 pairs of sites have $p_{ij} \leq 0.001$, a value that is usually taken as extremely significant. Therefore pairs such as 12–23 that fail our significance test in the 440 cannot really be ruled out.

Another related major source of difficulty is the sparsity of the data. Many sites in the variable V3 loop are not obviously particularly variable. Most sites have the consensus amino acid 80% or more of the time, with 5–10 other amino acids each appearing rarely. There are two consequences. One we have noted before is that the study of the relationship of more than pairs of sites is made difficult by the absence of data on many combinations of amino acids. Another is that a small number of cases can make big changes in the observed significance of the covariability of a pair of sites. Examples are discussed in the Appendix.

So where does this leave us? All the Korber *et al.* sites and connections other than those with 23 continue to be signaled. The technology they initiated, somewhat modified, confirms known biologically significant sites such as 5, 7, and 10 and points to others such as 19, 24, and 26.

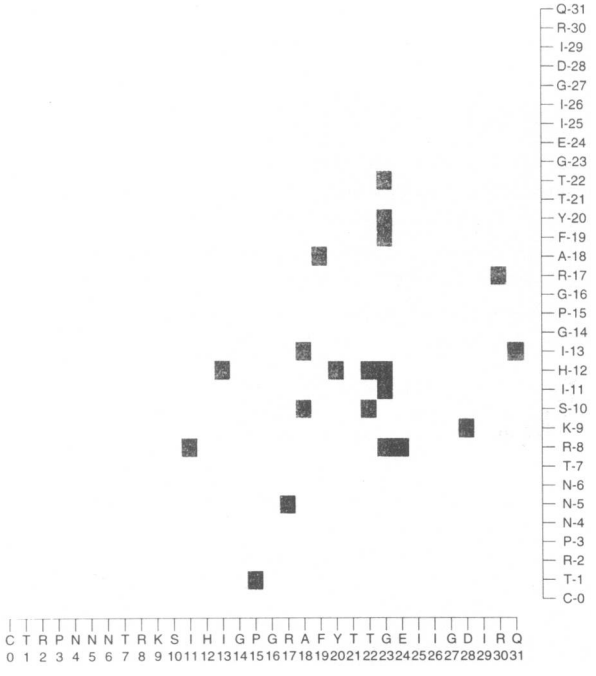

FIG. 3. Pairs significantly covarying in the 308 only, for G or M or P, with the constraint that the value of the statistic must lie in its upper quartile for the data set.

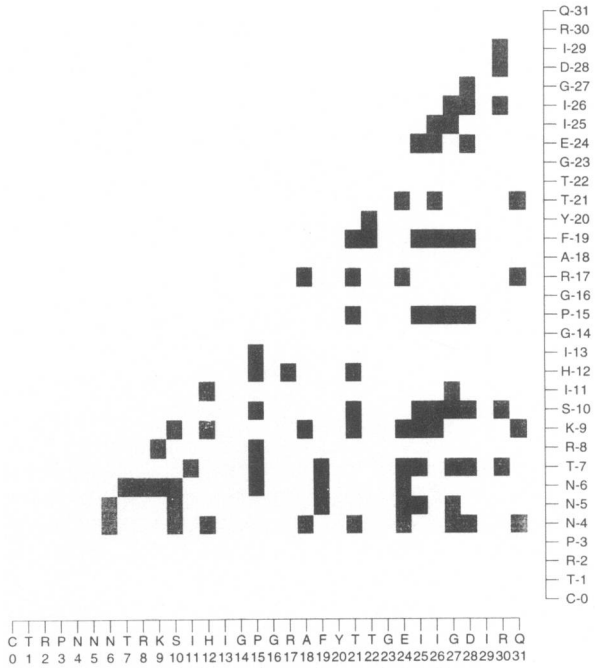

FIG. 4. Pairs significantly covarying in the 440 only, for G or M or P, with the constraint that the value of the statistic must lie in its upper quartile for the data set.

APPENDIX

Further discussion of the P_{ij} *statistic*

Although the P_{ij} statistic has not to our knowledge been introduced previously, it is based on a classical test for the hypothesis of independence against the alternative of a particular kind of "quasi independence"—see Refs. 22 and 23. Specifically, if $p_{ij}(x,y)$ denotes the probability (population frequency) of amino acid x at i and y at j and $p_i(x)$ is the probability of amino acid x at i then $P_{ij}(a,b)$ is the likelihood ratio test statistic for the hypothesis of independence,

$$H: p_{ij}(x,y) = p_i(x)p_j(y) \quad \text{for all} \quad x,y$$

under the blanket assumption that, for functions f and g,

$$p_{ij}(x,y) = f_i(x)g_j(y) \quad \text{for} \quad x \neq a \quad \text{and} \quad y \neq b \qquad (1)$$

The hypothesis simply specifies that (1) applies to $x = a, y = b$ as well. This reflects the view that, if there is covariability, it occurs only for a particular pair of amino acids, i.e., that there has been possible "slippage" from the hypothesis of independence only for a,b.

The statistic P_{ij} corresponds to the likelihood ratio statistic if in Eq. (1) we do not specify which pair (a,b) has "slipped."

Note that this idea can be extended to construct other plau-

sible statistics. For instance, we may make the assumption that Eq. (1) holds provided that $x \neq a$. This leads to the statistic

$$P_{ij}(a) = \sum_y [\hat{p}_j(y) - \hat{p}_{ij}(a,y)] \times \log\left\{\left[1 - \frac{\hat{p}_{ij}(a,y)}{\hat{p}_j(y)}\right][1 - \hat{p}_i(a)]^{-1}\right\} + \sum_y \hat{p}_{ij}(a,y) \log\left\{\frac{\hat{p}_{ij}(a,y)}{\hat{p}_i(a)\hat{p}_j(y)}\right\} \quad (2)$$

The corresponding statistics $\max_a P_i(a)$ and $\max_b P_j(b)$ are well worth pursuing in linkage studies when mutation at one site results in selective pressure not only at its own but also at related loci. $P_{ij}(a)$ may be thought of as a measure of the effect on prediction at site j of knowledge of amino acid a at site i. As such, it measures somewhat the same features as the specific information at j given residue a at site i of Korber *et al.* Their expression $I(S_a)$, is, in our notation,

$$I(S_a) = \sum_y \hat{p}_{ij}(a,y) \log\left[\frac{\hat{p}_{ij}(a,y)}{\hat{p}_i(a)}\right] - \sum_y \hat{p}_i(a)\hat{p}_j(y) \log \hat{p}_j(y)$$

which differs from the second term in Eq. (2) by

$$\sum_y [\hat{p}_{ij}(a,y) - \hat{p}_i(a)\hat{p}_j(y)] \log \hat{p}_j(y)$$

Returning to P_{ij} we note that it has the advantage over G_{ij} and M_{ij} of pointing to the pair of residues that seems to be statistically most covariant. In turn, that enables us when considering more than two sites simultaneously to use the data to reduce the amino acid alphabet at the sites under consideration and examine covariance between collections of sites as we will indicate. It is also worth noting that the extremal 2×2 table to which the P_{ij} statistic leads us visually highlights the strength of the relationship between a and b at sites i and j. Here is the 440 10–19 extremal table that illustrates our point:

	F	$\bar{F}$	
S	275 (241)	66 (100)	341
$\bar{S}$	36 (70)	63 (29)	99
	311	129	440

This table presents observed and (expected) numbers of cases in extremal P for sites 10 and 19. Categories are S (serine), S[m] (not S), F (phenylalanine, *F[m]* (not F).

Fitting second-order models

Given 3 sites and on the order of 10 amino acids appearing at each site in a set of sequences, we have approximately $10^3 = 1000$ distinct possible combinations of amino acids at the 3 sites. Given that we have even in our largest data set 440 sequences, we do not see many possible combinations. In fact, given that the consensus frequency at each site is typically not less than 70% and the high degree of covariability observed, we typically see fewer than 100 of the 1000 possible combinations.

We want to examine the possibility that covariation of sites i and j is a consequence of this mutual covariation with site k, i.e., that given the amino acid at site k, the amino acids at i and j vary independently. In view of our remarks above we cannot hope to examine all triples of amino acids. However, we can try to use a device that is standard in analyses of discrete data. We consider modeling $p_{ijk}(a,b,c)$, the probability that amino acid a occurs at i, b at j, and c at k by

$$\log p_{ijk}(a,b,c) = h_{ij}(a,b) + h_{ik}(a,c) + h_{jk}(b,c) \quad (3)$$

where h_{ij}, h_{ik}, and h_{jk} are arbitrary functions on the pairs of possible amino acids subject only to the constraint that

$$\sum_{a,b,c} p_{ijk}(a,b,c) = 1.$$

This restriction reduces the number of parameters that need to be fitted from the data. For instance, if $A = 10$ amino acids appear at each site, "only" $3A(A-1) = 270$ need to be fitted. If we could then test the hypothesis that, in this context, $h_{ij}(\cdot,\cdot) \equiv 0$, which corresponds to i and j independent given k, we could then use the resulting χ^2 tests and p values crudely with small p values as real evidence of covariability, and large p values as supporting the hypothesis. See Chapter 7 of Ref. 23 for a discussion of such methods.

Unfortunately, this program is not applicable to most triples of sites since in fact we do not have the observations requisite to fit the 270 or so parameters. As a consequence we reduce the "alphabet size" at each site in one of two ways. If the sites have charged consensus residues we use the charge alphabet, +, −, P, NP, where +, − refer to the charges on amino acid side chains when it is charged; P represents polar, and NP is nonpolar. For 3 sites this immediately reduces the number of possible combinations to 64, and all 64 parameters or at least a second-order model with 36 parameters can usually be fit. For sites that rarely exhibit charged side chain amino acids we use the P statistic to produce smaller data-determined alphabets. For example, consider the clique of sites 10, 12, 19, and 24, which are significantly covariant under any of the three statistics and for both data sets. If we consider the extremal P tables for each of these sites, and all pairings among the short list of 12, we find these amino acids occurring:

10 G, R, S
12 T, H
19 F, L, V
24 D, E

Not surprisingly, the modal (consensus) residue at the site for each data set appears in each list. In fact, S, G, and R are, in order, the three most frequent residues for site 10; F and L are, in order, the two most frequent for site 19; and E and D are, respectively, the consensus amino acids for site 24 in the 308 and 440 data sets. We can now analyze the relationship be-

tween sites 10, 12, 19, and 24 as in expression (3) above by replacing the amino acid alphabet by the four-letter alphabet: {G, R, S, other} for 10, {T, H, other} for 12, {F, L, V, other} for 19, and {D, E, other} for 24. These analyses have the unsatisfactory aspect that the alphabet depends on the data set. On the other hand, in practice it appears to boil down to using as distinct categories the residues of each site in order of frequency. This poses problems when examining the dynamic AIDS epidemic but should be reasonably stable for many other situations.

Sensitivity of results to small numbers of cases

We noted earlier that if the modal (consensus) frequencies of one of a pair of sites was high, highly significant values of our statistics ($p_{ij} = 0$) could be turned to low significance by changing the residue values for a small number of cases. Essentially this occurs when the values of the statistic G, M, or P is small in absolute magnitude, although large relative to what could have been given the marginal frequencies of residues at the two sites. We try to eliminate this situation in our short list by insisting on P, M, and G values in the top quartile of their distribution. We do not know the frequency of such misclassifications due to misalignment or mistyping, but it would certainly seem reasonable to be suspicious of strong covariation that is determined to be such by five or fewer cases.

Here is an example, drawn from our list of 20 pairs of Table 3, that shows what can happen.

The observed extremal 5 versus 26 table for the 308 is

	N	$\bar{N}$	
I	232	15	247
$\bar{I}$	45	16	61
		(6.14)	
	277	31	

Here, I is isoleucine, N is asparagine; and $\bar{I}\bar{N}$ are not I and not N, respectively; and the number in parentheses is "expected." This table has a significance probability $< 10^{-5}$ for P.

If five cases are moved from the I,N cell, maintaining the marginals we obtain,

227	20
50	11

the significance probability of the table becomes at least 0.01, which is above the median significance probability of 0.001 obtained for all pairs. On the other hand, the observed extremal 5 vs. 26 table for the 440 is

	N	$\bar{N}$
I	373	6
$\bar{I}$	29	32 (5.27)

This again has P with significance $<10^{-5}$. However, one would have to move more than 10 cases from the (I, N) cell to create a change in significance as drastic as before. Such analyses using the six pairs in Table 3 that are G and M and P significant show this kind of robustness of the relationship for both the 308 and 440 data sets. The situation for the other statistics M and G is more difficult to analyze. As we have seen with 5–26, P robustness may not hold for some of the remaining G or M or P pairs and at least one of the data sets. We intend to investigate such questions of robustness elsewhere.

Extreme thresholds can be misleading

Korber *et al.*[1] used as a threshold of significance the maximum value of the M statistics obtained in any of their 1,000,000 pseudo data sets for any pair of sites. We can illustrate that this rule is misleading by considering the P statistic, for which we can compute in closed form the maximum value it can attain for specified marginal frequencies.

Specifically, if the extremal P table is (when the entries are normalized by N) the number of cases given by

x	$\beta - x$
$\alpha - x$	$1 - \alpha - \beta + x$

where $\alpha \leq \beta \leq 1/2$, then the maximal value of P is $-\beta \log \beta - (1 - \alpha) \log(1 - \alpha) + (\beta - \alpha) \log(\beta - \alpha)$. If $\alpha = \beta = 1/2$ this achieves its maximum possible value of log2.

For the 440 the maximum value of P observed over all 100,000 pseudo data sets and pairs of sites is 0.052. On the other hand, the pair 10–27 exhibits a P of 0.021, which exceeds the maximum value of P for that pair over all pseudo data sets. In fact the maximum possible value of P at that pair of sites is 0.026. Thus we have a clear example in which setting the threshold value on the basis of the maximum value of the statistic observed at all pairs of sites for all pseudo random data sets is equivalent to setting an impossible goal.

ACKNOWLEDGMENTS

Research by P.J. Bickel was supported in part by Grant NSF DMS95-04955, NSA MDA-904-94-H-2020; research by R.A. Olshen was supported in part by Grant NIH CA59039-20 to Stanford University; research by A.G. Rodrigo and J.I. Mullins was supported in part by Grant NIH AI32885-4 to the University of Washington.

We are greatly indebted to Dr. Bette Korber for comments and help coming perilously close to coauthorship. We also thank Professor Leo Goodman for corrections and advice and Ms. Bonnie Chung and Ms. Debbie Haaxman for their considerable technical assistance. Any failings of this paper are, however, undoubtedly our own.

REFERENCES

1. Korber BTM, Farber RM, Wolpert DH, and Lapedes AS: Covariation of mutations in the V3 loop of human immunodefi-

ciency virus type 1 envelope protein: An information theoretic analysis. Proc Natl Acad Sci USA 1993;90:7176–7180.
2. Myers G, Korber BTM, Berzofsky JA, Smith RF, and Pavlakis GF (eds): *Human Retroviruses and AIDS*. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, New Mexico, 1991.
3. Davis D, Stephens M, Willers C, and Lachmann PJ: Glycosylation governs the binding of antipeptide antibodies to regions of hypervariable amino acid sequence within recombinant gp120 of human immunodeficiency virus type 1. J Gen Virol 1990;71:2889–2898.
4. Korber B, Wolinsky S, Haynes B, Kunstman K, Levy R, Furtado M, Otto P, and Myers G: HIV-1 intrapatient sequence diversity in the immunogenic V3 region. AIDS Res Hum Retroviruses 1992;8:1461–1465.
5. Hwang SS, Boyle TJ, Lyerly HK, and Cullen BR: Identification of the envelope V3 loop as the primary determinant of cell tropism in HIV-1. Science 1992;253:71–74.
6. de Jong J-J, Goudsmit J, Keulen W, Klaver B, Krone W, Tersmette M, and de Ronde A: Human immunodeficiency virus type 1 clones chimeric for the envelope V3 domain differ in syncytium formation and replication capacity. J Virol 1992;66:757–765.
7. Chesebro B, Wehrly K, Nishio J, and Perryman S: Macrophage-tropic human immunodeficiency virus isolates from different patients exhibit unusual V3 envelope sequence homogeneity in comparison with T-cell-tropic isolates: Definition of critical amino acids involved in cell tropism. J Virol 1992;66:6547–6554.
8. Fouchier RAM, Groenink M, Kootstra NA, Termsette M, Huisman HG, Miedema F, and Schuitemaker H: Phenotype-associated sequence variation I the third variable domain of the human immunodeficiency virus type 1 gp120 molecule. J Virol 1992;66: 3183–3187.
9. Westervelt P, Trowbridge DB, Epstein LG, Blumberg BM, Li Y, Hahn BH, Shaw GM, Price RW, and Ratner L: Macrophage tropism determinants of human immunodeficiency virus type 1 *in vivo*. J Virol 1992;66:2577–2582.
10. Milich L, Margolin B, and Swanstrom R: V3 loop of the human immunodeficiency virus type 1 Env protein: Interpreting sequence variability. J Virol 1993;67:5623–5634.
11. Seiller-Moiseiwitsch F, Margolin BH, and Swanstrom R: Genetic variability of the human immunodeficiency virus. Annu Rev Genet 1994;28:559–596.
12. Cosman PC, Gray RM, and Olshen RA: Evaluating quality of compressed medical images: SNR, subjective rating, and diagnostic accuracy. Proc IEEE 1994;82(6):919–932.
13. Breiman L, Friedman JM, Olshen RA, and Stone C: *Classification and Regression Trees*. Wadsworth, Belmont, California, 1984.
14. Myers G, Korber B, Wain-Hobson S, Smith RF, and Parlakis GN: *Human Retroviruses and AIDS*. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, New Mexico, 1993.
15. Kuiken CL, Zwart G, Gaan E, Coutinho RA, van der Hoek JAR, and Goudsmit J: Increasing antigenic and genetic diversity of the HIV-1 V3 domain in the course of the AIDS epidemic. Proc Natl Acad Sci USA 1993;90:9061–9065.
16. LaRosa GJ, Davide JP, Weinhold K, Waterbury JA, Profy AT, Lewis JA, Langlois AJ, Dresman GR, Boswell RN, Shadduck P, Holley LH, Karplus M, Bolognesi DP, Mathews TJ, Emini EA, and Putney SD: Conserved sequence and structural elements in the HIV-1 principal neutralizing determinant. Science 1990;249:932–935.
17. Goodman LA and Kruskal WH: *Measures of association for Cross Classification*. Springer-Verlag, New York, 1979.
18. Weir BS and Cockerham CC: Testing hypotheses about linkage disequilibrium with multiple alleles. Genetics 1978;88:633–642.
19. Everitt BS: *Analysis of Contingency Tables*, 2nd Ed. Chapman & Hall, London, 1992, pp. 37–40.
20. Wain-Hobson S: Is antigenic variation of HIV important for AIDS and what might be expected in the future? In: *The Evolutionary Biology of Viruses* (Morse S, ed.). Raven Press, New York, 1994.
21. Ghiara JB, Stura EA, Stanfield RL, Profy AT, and Wilson IA: Crystal structure of the principal neutralization site of HIV 1. Science 1994;264:82–85.
22. Goodman LA: Some multiplicative models for the analysis of cross classified data. In: *Proc. VIth Berkeley Symposium on Math. Statistics and Probability* (LeCam L, *et al.*, eds.), Vol. 1. University of California Press, Berkeley, California, 1972, pp. 649–696.
23. Agresti A: *Categorical Data Analysis*. John Wiley & Sons, New York, 1992.

Address reprint requests to:
Peter J. Bickel
Department of Statistics
University of California, Berkeley
Berkeley, California 94720-4735

This article has been cited by:

1. N. Pfeifer, T. Lengauer. 2012. Improving HIV coreceptor usage prediction in the clinic using hints from next-generation sequencing data. *Bioinformatics* **28**:18, i589-i595. [CrossRef]
2. V. Dahirel, K. Shekhar, F. Pereyra, T. Miura, M. Artyomov, S. Talsania, T. M. Allen, M. Altfeld, M. Carrington, D. J. Irvine, B. D. Walker, A. K. Chakraborty. 2011. From the Cover: Coordinate linkage of HIV evolution reveals regions of immunological vulnerability. *Proceedings of the National Academy of Sciences* **108**:28, 11530-11535. [CrossRef]
3. Sarah K Ho, Elena E Perez, Stephanie L Rose, Roxana M Coman, Amanda C Lowe, Wei Hou, Changxing Ma, Robert M Lawrence, Ben M Dunn, John W Sleasman, Maureen M Goodenow. 2009. Genetic determinants in HIV-1 Gag and Env V3 are related to viral response to combination antiretroviral therapy with a protease inhibitor. *AIDS* **23**:13, 1631-1640. [CrossRef]
4. C. Ahn, F. Seillier-Moiseiwitsch, G. G. Koch. 2008. Predictive tests for linked changes. *Statistics in Medicine* **27**:23, 4790-4804. [CrossRef]
5. Peter B. Gilbert, Vladimir Novitsky, Max Essex. 2005. Covariability of Selected Amino Acid Positions for HIV Type 1 Subtypes C and B. *AIDS Research and Human Retroviruses* **21**:12, 1016-1030. [Abstract] [Full Text PDF] [Full Text PDF with Links]
6. N Hoffman. 2003. Covariation of amino acid positions in HIV-1 protease. *Virology* **314**:2, 536-548. [CrossRef]
7. A. Gregory DiRienzo, Victor DeGruttola, Brendan Larder, Kurt Hertogs. 2003. Non-parametric methods to predict HIV drug susceptibility phenotype from genotype. *Statistics in Medicine* **22**:17, 2785-2798. [CrossRef]
8. P. J. Bickel, K. J. Kechris, P. C. Spector, G. J. Wedemayer, A. N. Glazer. 2002. Finding important sites in protein sequences. *Proceedings of the National Academy of Sciences* **99**:23, 14764-14771. [CrossRef]
9. S Crowder. 2001. Covariance analysis of RNA recognition motifs identifies functionally linked amino acids. *Journal of Molecular Biology* **310**:4, 793-800. [CrossRef]
10. Mark R. Segal, Michael P. Cummings, Alan E. Hubbard. 2001. Relating Amino Acid Sequence to Phenotype: Analysis of Peptide-Binding Data. *Biometrics* **57**:2, 632-643. [CrossRef]
11. Lynn Milich, Barry H. Margolin, Ronald Swanstrom. 1997. Patterns of Amino Acid Variability in NSI-like and SI-like V3 Sequences and a Linked Change in the CD4-Binding Domain of the HIV-1 Env Protein. *Virology* **239**:1, 108-118. [CrossRef]