CABIOS

Vol.6. no.3 1990
Pages 165 - 171

Identification of the most significant amphipathic helix with application to HIV and MHV envelope proteins

Ivan E.Auger* and Charles E.Lawrence

Abstract

Amphipathic helices, which play important roles in protein structure, occur in a wide variety of lengths. Yet existing methods employ fixed window lengths. We present a hierarchical procedure that identifies the Q most significant amphipathic helices regardless of length. Since the observed hydrophobicities are not normally distributed, test statistics usually employed for least-squares regression are inappropriate for assessing statistical significance of amphipathic helices. We show that an adjusted F statistic provides a good test. An application to the envelope protein of HIV finds an unexpected long amphipathic helix in gp41.

Introduction

In an amphipathic structure hydrophobic residues tend to cluster on one side with hydrophilic residues on the opposite side. An amphipathic α-helix is a structure that exhibits periodicity in hydrophobicity of $\sim$ 3.6 residues per turn. Amphipathic helices are important elements of protein structure. They can have functional roles as in the case of the spikes in viral envelope protein (Wilson *et al.*, 1981; Skehel *et al.*, 1982; de Groot *et al.*, 1987), and they can participate in T cell receptor binding (DeLisi and Berzofsky, 1985) and in signal transport functions (Heijne, 1986).

The two basic methods for detecting periodicity in hydrophobicity compare the sequence to a harmonic sequence (Cornette *et al.*, 1987). The first method uses the discrete Fourier transform to compute the so-called power spectrum or correlation function of the observed numerical sequence and a harmonic sequence. This approach has been widely used to detect amphipathic helices as illustrated by Eisenberg *et al.* (1984) for hydrophobic moments and by DeLisi and Berzofsky's (1985) procedure to detect T cell antigenic sites. The second method uses least-squares estimation. Cornette *et al.* (1987) have shown that these two approaches are closely related, and have discussed the advantages of each.

In native proteins, helices come in a wide range of lengths. For example, in influenza hemagglutinin HA2 there is a long helix of 60 residues, a shorter helix of 30 residues (Wilson *et al.*, 1981) and an amphipathic signal peptide of 10 residues (Auger and Lawrence, 1989). Existing methods employ windows with lengths set at fixed values, usually in the range of 7−20 residues. This restriction to fixed-length windows stems from the legitimate concern that existing measures of amphipathicity may not be length invariant. As we shall show, this is indeed the case since high amphipathic moments are more likely to arise by chance alone in short segments. This incompatibility of the available analytic techniques with the variability in length of natural helices has lead to considerable difficulty in sequence analysis (Heijne, 1986).

Below we examine the role of chance variation on measures of amphipathicity, develop an amphipathic measure that is length unbiased and present a method for identification of the Q most significant amphipathic helices. We illustrate the application of this method to an envelope protein of mouse hepatitis virus (MHV), a coronavirus, and to the envelope protein of human immunodeficiency virus (HIV).

Methods

A segment of observed hydrophobicities H_k, $k = p, \ldots, p + d$ exhibits amphipathic periodicity that can be described as:

$$H_k = \beta_0 + \beta_h \sin(k\cdot\omega + \phi) + \epsilon_k, \quad k = p, \ldots, p + d \quad (1)$$

where β_0 is a constant, β_h is the amplitude of the periodicity in hydrophobicity, ϕ is the phase, ω is $2\pi/3.6$ and ϵ_k is the error, i.e. the difference between the observed hydrophobicity (H_k) and the one predicted ($\hat{H}_k$) by the sine model. Least-squares estimates of the unknown parameters of equation (1) are obtained as follows by minimizing

$$\sum_{k=p}^{p+d} (H_k - \hat{H}_k)^2 = \sum_{k=p}^{p+d} \epsilon_k^2 \quad (2)$$

where

$$\hat{H}_k = \beta_0 + \hat{A}\sin(k\cdot\omega) + \hat{B}\cos(k\cdot\omega)$$

Now

$$\beta_h = \mathrm{sqrt}(\hat{A}^2 + \hat{B}^2) \text{ and } \phi = \arctan(-\hat{A}/\hat{B}).$$

The effect of chance variation on these estimators can be assessed when the distribution of the estimators is known under the null model, H_0: $\beta_h = 0$. That is, the sequence shows just chance variation around a constant hydrophobicity, $H_k = \beta_0$

Laboratory of Biometrics, PO Box 509, Wadsworth Center for Laboratories and Research, New York State Department of Health, Albany, NY 12201, USA

**To whom reprint requests should be sent*

$+ \epsilon_k$. If the errors ϵ_k were normally distributed, then it can be easily demonstrated that the parameter estimators $\hat{A}$ and $\hat{B}$ would be normally distributed, and the corresponding F statistic:

$$F = \frac{\Sigma(\hat{H}_i - \bar{H})^2/2}{\Sigma(H_i - \hat{H}_i)^2/(n - 3)} \quad (3)$$

is F distributed with 2, n - 3 degrees of freedom (Draper and Smith, 1966).

This approach is one step beyond the hydrophobic moment (Eisenberg *et al.*, 1984) and the least-square power spectrum (LSPS) (Cornette *et al.*, 1987) in that it assesses the statistical significance of having observed such an amphipathic helix. We could similarly assess the statistical significance of LSPS since $LSPS/\sigma^2$ is χ^2 distributed with three degrees of freedom if the errors are $N(0,\sigma^2)$ (Draper and Smith, 1966). This would test whether all regression parameters are different from zero, but this is useful only when σ^2 is known.

The assumption of normality of the errors is clearly not justified in this setting since the hydrophobicities assume 20 discrete values. It is well known that in some circumstances the F statistic is very intolerant to departures from normality of the dependent variable. In a classic paper, Box and Watson (1962) show that the robustness of the F statistic to departures from normality depends on the distributional form of the independent variables. They further show that, under the assumption that the errors under the null are independently distributed with some common distribution, not necessarily normal, then the F statistic is approximately distributed as $F_{\delta\nu 1,\delta\nu 2}$, where ν_1 and ν_2 are the nominal number of degrees of freedom. In this application $\nu_1 = 2$, $\nu_2 = n - 3$. Also, we have:

$$\delta^{-1} = 1 + (n + 1)\psi/(n - 1 - 2\psi);$$

$$\text{where } \psi = \frac{n - 3}{2n(n - 1)} \cdot C_X\Gamma_Y$$

$\Gamma_Y = E[k_4/k_2^2]$, where k_2 and k_4 are the sample cumulants for the n values of Y and C_X is a multivariate analogue of k_4/k_2^2.

To examine the effects of the non-normality and the utility of the above approximation in this setting, we generated 100 000 random residue sequences with sizes m = 10, 15, 20, 30, 40, 50, 75 and 100. The amino acid frequencies used are those of Dayhoff *et al.* (1978). Since we are interested in the most significant ones, we compare in Table I the p values for both the unadjusted and adjusted F distributions with the observed ones at significance levels in the right tail, 0.05, 0.01, 0.005 and 0.001. Notice that the departures from normality have a substantial and segment-length-dependent bias in the F statistic. For example, at a nominal value of 0.005 the unadjusted F statistic is nearly twice the nominal value for segments of length 10 and gradually approaches the nominal value as the segment length increases. On the other hand, the adjusted results provide uniformly good agreement and are consistently superior to the unadjusted distribution. Thus, the unadjusted distribution has a substantial length bias, while the approximation shows little if any bias. To evaluate the validity of the approximation in the extreme right tail of the distribution, we have examined the distribution of the smallest p values. We generated 1000 replicates of the smallest p value of a segment of length 20 in a sequence of length 200. We find that these p values from the adjusted F statistic follow the expected extremel distribution (Galambos, 1978) while the unadjusted do not.

We also investigated the utility of this approximation for other fixed frequencies. These include a periodicity of 3.5 as suggested for heptad patterns in coiled coil helices (McLachlan and Stewart, 1975), a periodicity of 3.7 as suggested by Cornette *et al.* (1987) and periodicities of 2 and 2.3 which are commonly used for β-strands. In all cases, simulation results were comparable with those described above for period 3.6. Furthermore, simulations using alternate hydrophobicity scales (Cornette *et al.*'s PRIFT, 1987; Sweet and Eisenberg, 1983; Eisenberg and McLachlan, 1986) also confirmed the utility of the Box and Watson adjustments. The only exception was Hopp and Wood's scale (1981) where there was no need for an adjustment.

In the next section we illustrate how these results can be employed to identify the Q most significant non-overlapping amphipathic helices.

Algorithm

To find the most significant amphipathic helix, we select the segments whose consonance with the model, equation (1), is least likely to arise by chance alone, as follows. For any given segment, starting at any residue p and continuing an arbitrary number of residues d, the probability of type I error is the

Table I.

Segment size	Observed significance level			
	0 05	0.01	0 005	0.001
10	0.04988	0.01153	0.00577	0.00112
	0.06016	0.01656	0.00942	0.00234
15	0 05035	0.01097	0.00582	0.00158
	0.05631	0.01367	0.00790	0.00237
20	0.04942	0.01100	0.00574	0.00164
	0.05379	0.01323	0.00714	0.00199
30	0 05092	0.01087	0.00571	0.00133
	0.05343	0.01214	0.00653	0.00161
40	0.05007	0.01037	0.00548	0.00119
	0.05204	0.01120	0.00593	0.00140
50	0.05094	0.01087	0.00562	0.00124
	0.05244	0.01152	0.00612	0.00132
75	0.05038	0.01025	0.00530	0.00114
	0 05131	0.01061	0.00574	0.00122
100	0.04909	0.00969	0.00511	0.00124
	0.05000	0.01018	0.00527	0.00133

Each cell contains the significance level computed by the adjusted-top/unadjusted-bottom F distribution for a given observed significance level.

probability of falsely rejecting the null hypothesis H_0, when it is true in favor of the alternative. As shown in the previous section, this probability is well approximated by evaluating the test F ratio against the F distribution with an adjusted number of degrees of freedom as follows:

$$\alpha(i,i+d) = P(\text{rejecting } H_0 \mid \beta_h = 0) = P \quad (f > F) \quad (4)$$

Since $\alpha(i,i+d)$ applies for arbitrary i and d, the most significant amphipathic helix is the segment with smallest probability of type I error, i.e.

$$p_{sig} = \min_{i,d \in S} \alpha(i,i+d)$$

where S is the set of starting points and lengths of interest. For example, to search an entire polypeptide of length n for amphipathic helices of length between 8 and 150 residues, S is the set, $7 \leq d \leq 149$, $i = 1, 2, \ldots, n-d$.

The algorithm shown in Figure 1 will identify the Q most significant non-overlapping segments.

The above algorithm first computes the least-squares estimates of model (1) and then it successively finds the Q most significant non-overlapping amphipathic segments. In order to efficiently compute step 1, we use the computation of the least-squares solution for segment H_i to H_j to compute the solution for segment H_i to H_{j+1}. This can be done very efficiently with techniques such as regression updating using Givens or Householder transformations (Seber, 1977), or by using the method of provisional means to compute the means, sum of squares and sum of cross-products (Herraman, 1968).

The above algorithm is hierarchical. As such, all segments included in previous steps remain in the solution. It may be the case that the top Q segments identified by this approach are not the best Q when they are considered in combination. We have previously described an algorithm that will yield the best Q segments regardless of order of entry (Auger and Lawrence, 1989). We need not repeat the algorithm here, since all that is required is the use of the p values from the adjusted F statistic in the objective function of the algorithm.

Under the null, the random variable p_{sig} is the smallest of a set of uniform random variables $\alpha(i,i+d)$. There are $O(n^2)$ of these uniform random variables. Many of these are observed from overlapping segments, since any segments within d residues of one another use common data. Thus, p_{sig} follows the distribution of the first-order statistic from a set of d-dependent uniform random variables. This situation is common in sequence analysis problems. Two types of methods have been developed for finding the critical values for such order statistics. The analytic approach employs asymptotic extremal theory (Karlin and Ghandour, 1985). The second set of methods employs random permutations (Karlin and Ghandour, 1985). In the applications presented below we employ the latter.

Fig. 1.

```
(* Step 1: compute probability of type I error *)
  for all (i,j), i< = j, i,j∈[1,n] do
    p_ij = α(i,j)
(* Step 2: find Q most significant segments *)
    amphsig[1. .n] = false
    psig = 0
    for k = 1 to Q do
      for i = 1 to n do
        if (.not.amphsig[i]) then
          for j = i + 1 to n do
            if .not.amphsig[j] then
              break
            else if psig[1] > p_ij then
              psig = (p_ij,i,j)
      amphsig[imax. .jmax] = true
      write k, 'th most significant is', psig
    end do
```

Applications

Many animal viruses obtain a lipid-containing membrane during maturation through a 'budding' process. These viruses direct the insertion of their own surface glycoproteins into the membrane envelope. Influenza haemagglutinin is the most well studied of these viral envelope proteins. This protein is composed of a fibrous stem section anchored into the membrane by a hydrophobic membrane-spanning region, and a globular domain distal to the membrane (Wilson *et al.*, 1981). The stem domain is composed of two amphipathic helices of 30 and 60 residues. The fibrous stem is formed by a coiled-coil interaction of these helices into a trimer. There is substantial evidence that

Table II.

Region of MHV envelope protein	Rank	Segment		p_{sig}	No. of segments more significant after 100 shufflings
		Boundaries	Length		
90a	1	971–1034	64	0.12405×10^{-7}	0
	2	799–843	45	0.15096×10^{-3}	11
	3	1296–1306	11	0.26194×10^{-3}	2
	4	1235–1248	14	0.63291×10^{-3}	4
	5	1071–1085	15	0.10547×10^{-2}	4
90b	1	8–19	12	0.87571×10^{-3}	87
	2	337–366	30	0.25073×10^{-2}	91
	3	593–636	44	0.36907×10^{-2}	90
	4	573–580	8	0.46280×10^{-2}	88
	5	239–247	9	0.52530×10^{-2}	84

many other viral envelope proteins exhibit a similar structure (Wiley, 1985). These proteins are frequently called spike proteins due to their appearance on electron micrographs. Thus, a search for amphipathic helices of variable length in the sequences of viral surface glycoproteins appears promising.

Application to a coronavirus

We have previously analyzed influenza hemagglutinin using another method (Auger and Lawrence, 1989). Since the substantive results of the two analyses are quite similar, we do not present them here. Instead, we first present an analysis of another well-defined viral surface glycoprotein. Mouse hepatitis virus (MHV-A59) is a coronavirus. Coronaviruses are positive-stranded, RNA-enveloped viruses. They infect human and domestic animals. An unusually large (~200 Å) spike protein (E2) projects from the virion surface. This spike protein mediates binding of the virion to cell surface receptors and is involved in host cell fusion. E2 is also the main target of the host immune response. A post-translational cleavage of the protein between residues 717 and 718 divides this protein into two halves, 90B and 90A. Cleavage is required for fusion activity (Sturman *et al.*, 1985). By analogy with influenza hemagglutinin the carboxyl half (90A), which contains the apparent transmembrane anchor segment, is generally acknowledged to constitute the stem portion. As an internal control we also applied the algorithm to the other structural proteins of MHV, E1 and N. None of these are expected to have long amphipathic helices.

Using an analysis focusing on heptad repeat patterns, de Groot *et al.* (1987) have proposed that there are two helices that combine together in a coiled coil to form the fibrous stem structure. Their model proposes that these helices form a dimer with a coiled coil in a manner similar to that of influenza hemagglutinin.

The most signifiant amphipathic helices for 90A and 90B are give in Table II along with the critical values derived from 100 random shufflings of the sequence. The most significant amphipathic helix is composed of 64 residues (971–1034). As indicated in Table II, this result is unlikely to have resulted from chance alone. The longer helix (948–1056) suggested by de Groot *et al.* covers this region. The second most significant helix (799–843) is also long. However, only the first amphipathic helix is statistically significant. The second helical structure (1209–1267) indicated by de Groot *et al.* (1987) is very close to the transmembrane region near the carboxyl terminus of the molecule. As indicated in Table II, we find no long helix in this region. The fourth most significant helix (1235–1248) is within the region, but is only 14 residues long. Application to the two other structural proteins of MHV (data not shown) showed no other significant amphipathic helices. Thus, we also find evidence to support the proposal that 90A has a helical fibrous structure. Furthermore, we find no significant amphipathic structures in 90B, the reputed globular half of the spike protein, or in any of the internal controls, structural proteins E1 and N.

Application to HIV

The viral surface glycoprotein of human immunodeficiency virus (HIV) is synthesized as a gp160 precursor protein, and is subsequently cleaved into the exterior gp120 and integral membrane gp41 proteins. Given the importance of this virus for public health, it is not surprising that this sequence has been the subject of intensive analysis (Alizon *et al.*, 1986; Cease *et al.*, 1987; Modrow *et al.*, 1987). We have analyzed both gp120 and gp41.

Table III shows the five most significant helices in gp120 of the BH10 variant. Cease *et al.* (1987) used the AMPHI algorithm (DeLisi and Berzofsky, 1985; Margalit *et al.*, 1987; Spouge *et al.*, 1987) to identify antigenic sites in gp120. They found five potential antigenic sites, and they chose to study those from more conserved regions. The two most favorable sites are env T2 (residues 112–124) and env T1 (residues 428–443). They found that helper T cell immunity can be induced with short peptides for both of these. From Table III we find that the first and third most significant regions are 420–435 and 99–129. Thus, env T1 clearly overlaps with our most significant region, and env T2 is contained in the third most significant region. Random shuffling indicates that these regions are not statistically significant.

Our most surprising result is found in the analysis of gp41. As indicated in Table IV for the BH10 variant, we find a long amphipathic helix covering a long stretch between residues 785 to 854 near the carboxyl terminus of gp41. We found significant long amphipathic helices in several other variants of HIV-1, but this is expected since gp41 is highly conserved across the HIV-1 variants. Given this high degree of conservation, more

Table III.

Region of MHV envelope protein, BH10	Rank	Segment		P_{sig}	No. of segments more significant after 100 shufflings
		Boundaries	Length		
gp120	1	420–435	16	0.17332×10^{-2}	93
	2	201–209	9	0.26762×10^{-2}	75
	3	99–129	31	0.33788×10^{-2}	63
	4	2–18	17	0.41240×10^{-2}	46
	5	246–257	12	0.51292×10^{-2}	35

notable is the finding of a somewhat different pattern for variants BRU, HXB2 and HXB3. As illustrated for BRU in Table IV, we found two long, significant amphipathic helices near the carboxyl terminus of these variants. We also find a long amphipathic helix for HIV2's ROD variant (see Table IV). Given the 44.8% level of conservation (Guyader *et al.*, 1987) in the gp41 region between HIV-1 (BRU isolate) and HIV-2 (ROD isolate), this result is notable. As we have discussed above, such long amphipathic helices are observed in the stem portion of viral envelope glycoproteins.

The membrane-associated polypeptide gp41 has three possible membrane-spanning regions (Modrow *et al.*, 1987) (see Figure 2). In the following paragraph the residue numbers of the BH10 variant are used for illustrative purposes. The first two candidate membrane spanning regions, TM1 and TM2, follow directly after the cleavage site in a 70 residue segment (511–580), interrupted by a stretch of 10 hydrophilic residues. The third candidate region, TM3, is a segment of 26 hydrophobic amino acids (670–695), followed 26 residues downstream by a hydrophilic segment (722–745). Thus gp41 may cross the membrane once, twice or three times. Consequently, it is not clear whether the segment in which the long amphipathic helices are located, carboxyl to TM3, is inside or outside the membrane. There is experimental evidence for both locations.

Within the TM1 sequence is a peptide that shows sequence similarity to the paramyxovirus fusion peptides. It has been shown that deletion mutants that lack this segment interfere with membrane fusion (Kowalski *et al.*, 1987). On the other hand, the carboxyl terminus of gp41 does not appear to be necessary for fusion in that deletions in this region do not affect syncytium formation (Kowalski *et al.*, 1987). However, such deletions result in a poorly replicating virus (Terwilliger *et al.*, 1986). HIV replicates by budding at the cell surface. While the sequence of events is not fully understood, evidence indicates that the budding process involves the interaction of the cytoplasmic domain of viral surface glycoproteins and components of the virus in the cytoplasm (Wiley, 1985). Thus, the carboxyl end of gp41 may be involved in the budding process, and thus be at the interior.

On the other hand, it has been observed that monoclonal antisera against the hydrophilic sequence (735–752) give cell surface labeling and are neutralizing. This suggests that the segment carboxyl to TM3 (695–856) is on the exterior, where a role as a fibrous spike is more plausible.

Viral surface glycoproteins are anchored in the membrane by a transmembrane hydrophobic peptide, which is terminated by a cytoplasmic domain. In most of these proteins the cytoplasmic domain is short. HIV and other related lentiviruses are unusual in that they have a large peptide in the region that appears to be a cytoplasmic domain. The structure and function of these large domains is poorly understood. Our finding that a large proportion of this domain contains a significant amphipathic helix may be taken either as supporting evidence that this domain is on the exterior, as indicated by antibody studies, or that this peptide functions in budding activity.

Table IV. Significant amphipathic helices in gp41 envelope protein

Variant of HIV envelope protein, gp41	Rank	Segment boundaries	Segment Length	p_{sig}[a]
BH10[b]	1	785–854	70	$0.115499497 \times 10^{-6}$
BRU[c]	1	826–860	35	$0.126961561 \times 10^{-6}$
	2	748–814	67	$0.637363147 \times 10^{-5}$
ROD/HIV2	1	758–856	99	$0.302024638 \times 10^{-8}$

[a]Zero segments more significant after 100 shufflings in all cases.
[b]We find nearly identical results for the variants ARV2, WMJ2, RF/HAT, ZAIRE6, PV22, MAL, BH8 and EL1.
[c]We find nearly identical results for HXB2 and HXB3.

Discussion

Amphipathic helices play an important role in protein structure. These structures are known to occur in a wide range of lengths, yet existing methods are centered on the use of fixed window lengths. We have presented a method that identifies the most significant amphipathic helix, regardless of length.

A number of other features have been suggested as indicative of helices, such as the capping pattern at their termini (Presta and Rose, 1988; Richardson and Richardson, 1988). The method developed here can be extended to incorporate such additional features by adding terms to the regression model. Since this avenue provides a means to the simultaneous incorporation of multiple effects, it is an important area for further research.

A key to the development of the method we presented was our finding that the F distribution, with suitable adjustment for the non-normality of the data, provided a good approximation to the distribution of the test F statistic. This produces a good estimate of the probability of type I error. It is important to point out that the usefulness of the adjusted F distribution of Box and Watson is dependent on the distributional characteristics of the independent variables. However, as reported above, our simulation results show that this approximation is useful for many commonly used frequencies and hydrophobicity scales.

Under the null hypothesis, we assume that the hydro-

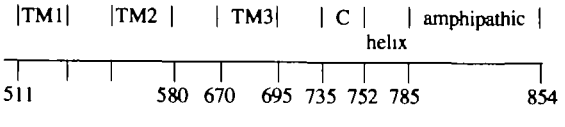

Fig. 2. gp41 of BH10 variant. TM1, TM2 and TM3 are the three transmembrane regions. C is the region that has been cell surface labeled by monoclonal antisera.

phobicities of the residues are independent. In the case of DNA sequences, it is well known that dinucleotide frequencies deviate from the expected under independence. Markov chain models have been used to describe this departure. The order of the Markov chain depends on the sequence (Fuchs, 1980; Blaisdell, 1986), and it has been observed to be from 1 to 4. The literature on Markov effects for proteins is limited. However, there is one directly germane report. Lipman and Pastor (1987) examined the correlation between the hydrophobicity of a residue and its neighbors 1–4 positions apart, and found no correlation in hydrophobicity when the full-length proteins were considered. However, when a sample is composed exclusively of α-helices of proteins with known structure, they found a correlation in the hydrophobicity of residues spaced four apart. This is the effect that we have exploited above.

As the HIV example illustrates, it is not always necessary to show statistical significance. In those cases where there is prior evidence that there are or should be important amphipathic helices, as in the case of gp120, the experimentalist needs help in identifying the best candidate sequences for further study. In such circumstances, the statistical significance of the identified sequence is of only marginal relevance to the pursuit of amphipathic helices. On the other hand, when there is no such *a priori* basis for the existence of amphipathic helices, as in the case of gp41, then statistical significance is crucial. In this case, without such a result there is little reason to pursue the finding experimentally.

Our applications to date have shown the ability of the algorithm to identify long amphipathic helices in the stem portions of viral spike proteins, where they are widely suspected of playing a major functional role. We have also illustrated the ability of the method to identify such a structure where none was suspected previously, in spite of extensive analysis.

Acknowledgements

We are grateful to Dr L.Sturman for many valuable comments and discussions on the applications to MHV and HIV.

References

Alizon,M., Wain-Hobson,S., Montagnier,L. and Sonigo,P. (1986) Genetic variability of the AIDS virus: nucleotide sequence analysis of two isolates from African patients. *Cell*, **46**, 63–74.

Auger,I. and Lawrence,C. (1989) Algorithms for optimal identification of sequence neighborhoods. *Bull. Math. Biol.*, **51**, 39–54.

Blaisdell,B.E. (1986) A measure of the similarity of sets of sequences not requiring sequence alignment. *Proc. Natl. Acad. Sci. USA*, **83**, 5155–5159.

Box,G.E.P. and Watson,G.S. (1962) Robustness to non-normality of regression tests. *Biometrika*, **49**, 93–106

Cease,K.B., Margalit,H., Cornette,J.L., Putney,S.D., Robey,W.G., Ouyang,C., Streicher,H.Z., Fischinger,P.J., Gallo,R.C., DeLisi,C. and Berzofsky,J.A. (1987) Helper T-cell antigenic site identification in the acquired immunodeficiency syndrome virus gp120 envelope protein using a 16-residue synthetic peptide. *Proc. Natl. Acad. Sci. USA*, **84**, 4249–4253.

Cornette,J.L., Cease,K.B., Margalit,H., Spouge,J.L., Berzofsky,J.A. and DeLisi,C. (1987) Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J. Mol. Biol.*, **195**, 659–685.

Dayhoff,M.O., Schwartz,R.N. and Orcutt,B.C. (1978) A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, New York, Vol. 3, pp. 345–352.

de Groot,R.J., Luytjes,W., Hirzinek,M.C., VanderZijst,B.A., Spaan,W.A. and Lenstra,J.A. (1987) Evidence for a coiled-coil structure in the spike proteins of coronaviruses. *J. Mol. Biol.*, **196**, 963–966.

DeLisi,C. and Berzofsky,J.A. (1985) T-Cell antigenic sites tend to be amphipathic structures. *Proc. Natl. Acad. Sci. USA*, **82**, 7048–7052.

Draper,N.R. and Smith,H. (1966) *Applied Regression Analysis*. John Wiley, New York.

Eisenberg,D. and McLachlan,A.D. (1986) Solvation energy in protein folding and binding. *Nature*, **319**, 199–203.

Eisenberg,D., Weiss,R.M. and Terwilliger,T.C. (1984) The hydrophobic moment detects periodicity in protein hydrophobicity *Proc. Natl. Acad. Sci. USA*, **81**, 140–144.

Fuchs,C. (1980) On the distribution of the nucleotides in seven completely sequenced DNAs. *Gene*, **10**, 371–373

Galambos,J. (1978) *The Asymptotic Theory of Extreme Order Statistics*. John Wiley, New York.

Guyader,M., Emerman,M., Sonige,P., Clavel,F., Montagnier,L. and Alizon,M. (1987) Genome organization and transactivation of the human immunodeficiency virus type 2. *Nature*, **326**, 662–669.

Heijne,G. von (1986) Mitochondrial targeting sequences may form amphiphilic helices. *EMBO J.*, **5**, 1335–1342.

Herraman,C. (1968) Sums of squares and products matrix. *Appl. Statist.*, **17**, 289–292.

Hopp,T.P. and Woods,K.R. (1981) Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci. USA*, **78**, 3824–3828.

Karlin,S. and Ghandour,G. (1985) Comparative statistics for DNA and protein sequences: single sequence analysis. *Proc. Natl. Acad. Sci. USA*, **82**, 5800–5804.

Kennedy,R.C., Kanda,P., Dreesman,G.R., Eichberg,J., Chanh,T.C., Ho,D.D. and Sparrow,J.T. (1987) Properties of synthetic peptides that identify neutralizing epitopes on the HIV envelope glycoprotein. In *Vaccines 87*. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, pp. 250–255.

Kowalski,M., Potz,P., Basiripour,L., Dorfman,T., Goh,W.C., Terwilliger,E., Dayton,A., Rosen,C., Haseltine,W. and Sodroski,J (1987) Functional regions of the envelope glycoprotein of human immunodeficiency virus type I. *Science*, **237**, 1351–1355.

Lipman,D.J. and Pastor,R.W. (1987) Local sequence of hydrophobicity and solvent accessibility in soluble globular proteins. *Biopolymers*, **26**, 17–26.

Margalit,H., Spouge,J.L., Cornette,J.L., Cease,K.B., DeLisi,C. and Berzofsky,J.A. (1987) Prediction of immunodominant helper T cell antigenic sites from primary sequence. *J. Immunol.*, **138**, 2213–2229.

McLachlan,A.D. and Stewart,M. (1975) Tropomyosin coiled-coil interactions: evidence for an unstaggered structure. *J. Mol. Biol.*, **98**, 293–304.

Modrow,S., Hahn,B.H., Shaw,G.M., Gallo,R.C., Wong-Staal,F. and Wolf,H. (1987) Computer-assisted analysis of envelope protein sequences of seven human immunodeficiency virus isolates: prediction of antigenic epitopes in conserved and variable regions. *J. Virol.*, **61**, 570–578.

Presta,L.G and Rose,G.D (1988) Helix signals in proteins. *Science*, **240**, 1632–1641.

Richardson,J.S. and Richarson,D.C. (1988) Amino acid preferences for specific locations at the ends of α helices. *Science*, **240**, 1648–1652

Seber,G.A.F. (1977) *Linear Regression Analysis*. John Wiley, New York.

Skehel,J.J., Bayley,P.M., Brown,E.B., Martin,S.R , Waterfield,M.D., White,J.M., Wilson,I.A. and Wiley,D.C. (1982) Changes in the conformation of influenza virus hemagglutinin at the pH optimum of virus-mediated membrane fusion. *Proc. Natl. Acad. Sci. USA*, **79**, 968–972.

Spouge,J.L., Guy,H.R., Cornette,J.L., Margalit,H., Cease,K., Berzofsky,J.A. and DeLisi,C. (1987) Strong conformational propensities enhance T cell antigenicity. *J. Immunol.*, **138**, 204–212.

Sturman,L.S , Ricard,C.S. and Holmes,K.V. (1985) Proteolytic cleavage of the E2 glycoprotein of murine coronavirus. Activation of cell-fusing activity of virons by trypsin separation of two different 90K cleavage products. *J. Virol.*, **56**, 904–911.

Sweet,R.M. and Eisenberg,D. (1983) Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure. *J. Mol. Biol* **171**, 479–488.

Terwilliger,E., Sodroski,J.G., Rosen,C.A. and Haseltine,W.A. (1986) Effects of mutations within the 3' orf open reading frame region of human T-cell

lymphotropic virus type III (HTLV-III/LAV) on replication and cytopathogenicity. *J. Virol.*, **60**, 754–760.
Wiley,D.C. (1985) Viral membrane. In Fields,B.N. (ed.), *Virology* Raven Press, New York, pp. 45–67.
Wilson,I.A., Skehel,J.J. and Wiley,D.C. (1981) Structure of the haemagglutinin membrane glycoprotein of influenza virus at 3Å resolution. *Nature*, **289**, 366–373.

Received on July 12, 1989; accepted on February 28, 1990

Circle No. 1 on Reader Enquiry Card