The Evolution of Proteins from Random Amino Acid Sequences. I. Evidence from the Lengthwise Distribution of Amino Acids in Modern Protein Sequences

Stephen H. White and Russell E. Jacobs*

Department of Physiology and Biophysics, University of California, Irvine, CA 92717, USA

Summary. We examine in this paper one of the expected consequences of the hypothesis that modern proteins evolved from random heteropeptide sequences. Specifically, we investigate the lengthwise distributions of amino acids in a set of 1,789 protein sequences with little sequence identity using the run test statistic ($r_o$) of Mood (1940, Ann. Math. Stat. 11, 367–392). The probability density of $r_o$ for a collection of random sequences has mean = 0 and variance = 1 [the N(0,1) distribution] and can be used to measure the tendency of amino acids of a given type to cluster together in a sequence relative to that of a random sequence. We implement the run test using binary representations of protein sequences in which the amino acids of interest are assigned a value of 1 and all others a value of 0. We consider individual amino acids and sets of various combinations of them based upon hydrophobicity (4 sets), charge (3 sets), volume (4 sets), and secondary structure propensity (3 sets). We find that any sequence chosen randomly has a 90% or greater chance of having a lengthwise distribution of amino acids that is indistinguishable from the random expectation regardless of amino acid type. We regard this as strong support for the random-origin hypothesis. However, we do observe significant deviations from the random expectation as might be expected after billions of years of evolution. Two important global trends are found: (1) Amino acids with a strong α-helix propensity show a strong tendency to cluster whereas those with β-sheet or reverse-turn propensity do not. (2) Clustered rather than evenly distributed patterns tend to be preferred by the individual amino acids and this is particularly so for methionine. Finally, we consider the problem of reconciling the random nature of protein sequences with structurally meaningful periodic "patterns" that can be detected by sliding-window, autocorrelation, and Fourier analyses. Two examples, rhodopsin and bacteriorhodopsin, show that such patterns are a natural feature of random sequences.

Key words: Protein evolution — Protein sequence analysis — Random protein sequences — Run test — Protein folding — Rhodopsin — Bacteriorhodopsin

Introduction

One of the enduring problems of protein evolution is the origin of the rich diversity in sequence, size, and structure of modern proteins. The conventional view is that modern proteins have evolved by means of gene duplication, exon shuffling, and random mutations from a small set of "starter sequences" with lengths of 30 to 50 amino acids (Eck and Dayhoff 1966; McLachlan 1972; Gilbert 1978; W. Doolittle 1978; Darnell 1978; R. Doolittle 1979; Blake 1983; Holland and Blake 1990; Dorit et al. 1990; Patthy 1991; Dorit and Gilbert 1991). A recent version of this hypothesis envisages modern exons

* Present address: Beckman Institute, Mail Stop 139-74, California Institute of Technology, Pasadena, CA 91125, USA
Offprint requests to: S.H. White
as descendents of the original genes encoding the starter sequences (Dorit et al. 1990; Dorit and Gilbert 1991). Regardless of the precise details, the underlying assumptions of this starter-set hypothesis are that (1) the first proteins were small and (2) the most likely way of increasing the lengths of protein sequences was by a gene duplication process.

An alternate view is that the earliest proteins originated from random sequences of amino acids (White and Jacobs 1990; Lau and Dill 1990; Shakhnovich and Gutin 1990a). Assuming the existence of a primitive system for the coding, translation, and synthesis of protein sequences, White and Jacobs (1990) suggested that there may have been a protein synthetic ‘‘big bang’’ which resulted in a very large number of random heteropeptide sequences. Because it is now known that the ability to fold into compact structure places few restrictions on the space of possible amino acid sequences (Chan and Dill 1990; Lau and Dill 1990; Shakhnovich and Gutin 1989; Shakhnovich and Gutin 1990a), a significant fraction of these sequences could have folded into protoproteins. The primitive proteins are assumed to be the members of this protogroup that could catalyze some advantageous chemical reaction or serve a useful structural role. Even if the probability for the formation of catalytic primitive proteins was exceeding low, the number of such proteins which could be produced from, say, one mole of random heteropeptide chains, could nevertheless be very large.

The random-origin hypothesis has two corollaries. First, the amino acid sequences of the primitive proteins would be examples of random sequences derived by random sampling of sequence space. Second, the lengths of the protoprotein sequences would be determined solely by the lengthwise statistics of the nucleic acid ‘‘begin’’ and ‘‘end’’ signals which delimited the primitive genes or their transcribed messages. That is, gene duplication would not be required to achieve a broad range of sequence lengths. If the random-origin hypothesis is correct, then modern random sequences should, to a first approximation at least, conform to relatively simple statistical descriptions. The purpose of the present paper is to investigate the first corollary by examining the lengthwise distributions of amino acids in modern protein sequences. The second corollary will be considered in a separate paper (S. H. White, submitted).

We divide statistical studies of protein sequences into two broad categories: local and global. Local studies are those concerned with the frequencies of occurrence of single amino acids or short oligopeptide sequences (two to 11 amino acids) within proteins whereas global studies are those concerned with entire sequences. Most of the local studies involve identifying single amino acids and oligopeptides which are biased toward specific structural features. Studies concerned with the prediction of secondary structure are good examples. (See reviews by Fasman (1989) and Garnier (1990).) The first study of oligopeptides was undertaken by Gomow and Ycas (1958) in an effort to break the genetic code using crypticographic and information theory analyses. Ycas (1958) concluded from the limited number of sequences then available that there was no evidence for the nonrandom occurrence of dipeptides. Later studies using larger databases have produced somewhat conflicting results but it is now fair to say that many oligopeptides have been observed which do not meet the random expectation (Black et al. 1976; Klapper 1977a,b; Saroff 1984; Vonderviszt et al. 1986; Doollittle 1989; Zielenkiewicz et al. 1988; Wilson et al. 1985; McCaldon and Argos 1988). Of these studies, that of McCaldon and Argos (1988) is particularly thorough and convincing.

Global approaches to protein sequence statistics tend to be concerned with the broader and less specific relationships between sequence and structure. For example, the early observations of Ycas (1958) led Gates and Fisher (1971) to assume that proteins were random heteropolymers in order to estimate the dimensions of soluble proteins based upon simple geometrical considerations (Fisher 1964) and observations regarding the frequency of hydrophobic residues. One of their important conclusions, subsequently confirmed by several laboratories (Lee and Richards 1971; Janin 1979), was that some of the hydrophobic residues had to be exposed on the surface of the protein. Janin (1979) showed that the mathematical relation between protein molecular weight and the numbers of residues buried and exposed accurately obeyed the model of Gates and Fisher (1971) and thereby supported implicitly the proposition that proteins may be examples of random heteropolymers. Ptitsyn (1985, 1987) and Finkelstein and Ptitsyn (1987) have shown that many of the basic folding patterns of proteins are natural consequences of random protein sequences. We came to a similar conclusion regarding membrane proteins (White and Jacobs 1990).

The work presented here is concerned with the lengthwise distribution of amino acids and falls in the global category because we are concerned with the general patterns of lengthwise distributions rather than specific amino acid arrangements within short oligopeptide sequences. We assess these general patterns using the run test which we used earlier (White and Jacobs 1990) to examine the statistics of the runs of a particular set of hydrophobic residues. We showed that the statistical distribution of the run test statistic for 5,247 known protein se-
quences in the 1988 Protein Identification Resource (PIR) database (George et al. 1986; Orcutt et al. 1983) was remarkably similar to that expected for an equivalent number of random sequences. The deficiency of the study was that only a single physical property was examined in a collection of sequences which had an uneven representation of superfamilies. In the present paper, we therefore apply the analysis to a representative set of 1,789 sequences with little sequence identity and extend it to include the amino acids individually and by classes defined according to physical properties and secondary structure propensity.

We find that any randomly chosen sequence in the representative collection (The Superfamily Set) has a 90% or greater chance of having a lengthwise distribution of residues that is indistinguishable from the random expectation regardless of amino acid type. This supports the random-origin hypothesis but does not by itself rule out the starter-set hypothesis. We also show that the mathematical means of the probability densities for an entire collection of sequences can deviate significantly from the random expectation. These differences reveal global restraints selected through protein evolution. The largest systematic discrepancies occur when the amino acids are considered individually. When considered by classes, the discrepancies are small enough that, to a first approximation, extant proteins may be considered as a collection of random sequences. We discuss the problem of reconciling sequence randomness with structurally meaningful periodic "patterns" that can be detected by auto-correlation and Fourier analyses. We show by means of example that such patterns are a natural consequence of random sequences. This helps explain how structure can arise from random sequences.

Statistical Methods and Databases

Protein Databases. The primary database source is the Protein Identification Resource (PIR) Sequence Database produced by the National Biomedical Research Foundation (Washington, D.C.) [see Orcutt et al. (1983) and George et al. (1986).] PIR release number 21 (30 June 1989) was the principal database used and will be referred to as the 1989 PIR Database. However, we also examined the 1988 PIR Database (release number 16, 31 March 1988) and the 1990 PIR Database (release number 27, 31 December 1990). To form these databases we merged the so-called Annotated & Classified entries with the Preliminary entries. The total numbers of entries in the databases were 7,996 for 1988, 12,476 for 1989, and 20,354 for 1990.

Two problems arise if the databases are used in their entirety. First, not all of the sequences are complete; some are fragments or have uncertain amino acid assignments. We therefore rejected all sequences with the PIR identifier preceded by F1 (indicating fragment) or which contained B (Asx), Z (Glx), or X (unknown or nonstandard). These restrictions reduced the operational sizes of the databases to 5247 qualifying sequences for the 1988 PIR Database, 8752 for 1989, and 14,620 for 1990. A qualifying sequence is one which satisfies the mathematical criteria for the runs test; this generally means that the sequence must have a length of at least 21 amino acids and that there must be at least 10 amino acids to which a value of 1 can be assigned.

The second problem is that different superfamilies of sequences, based on sequence identity/homology, are not evenly represented. To reduce the possibility that this unevenness would bias the results, we constructed from the Annotated & Classified sections of the 1989 and 1990 PIR databases Superfamily Sets by extracting one qualifying sequence from each superfamily. We arbitrarily selected the first qualified sequence of each superfamily that was encountered. The 1989 Superfamily Set contained 1789 sequences and the 1990 Superfamily Set 2272. Because of the amount of work done prior to obtaining the 1990 version of the database, the 1989 Superfamily Set is the primary database discussed in this paper.

The Run Test for Sequence Randomness. We adopt as the null hypothesis the assertion that protein sequences are random; specifically, we consider the assertion that a sequence is indistinguishable from one expected by the random selection of amino acids (with replacement) from a pool containing the different kinds of amino acids in proportion to their frequencies of occurrence in the sequence. This is referred to as the independence random model by Karlin et al. (1991). In our experience the run test is, overall, the simplest and most sensitive general method for testing this assumption. A complete general theory of the statistical distribution of runs was developed by Mood (1940) and has been discussed didactically by David and Barton (1962). A protein consists of a sequence of 20 kinds of objects (amino acids) and one can consider the distribution of the runs of these objects using Mood's limit theorem. However, it is more useful and informative to simplify the problem by reducing the sequence to a binary sequence by assigning a value of 1 to the amino acids of interest and 0 to all others.

The theory reduces to a particularly simple and easily used form for binary sequences. (See, e.g., Wani 1971.) Consider a sequence of length \( L = a + b \) consisting of \( a \) 1's and \( b \) 0's with \( a > b \). If this is a sequence of independent Bernoulli trials (a Bernoulli sequence), then the expected total number of runs of 1's and 0's is

\[
\mu_{\text{ex}} = [2ab(a + b)] + 1
\]  

If one were to generate by a single random process a large number of sequences of length \( L \) with fixed \( a \) and \( b \), the observed number of runs \( \mu_{\text{obs}} \) would fluctuate about the expected value with a variance

\[
\sigma^2 = 2ab(2ab - a - b)(a + b)^2(a + b - 1)
\]  

Comparisons of \( \mu_{\text{obs}} \) with \( \mu_{\text{ex}} \) are conveniently made by means of the run test statistic given by

\[
r_0 = (\mu_{\text{obs}} - \mu_{\text{ex}})/\sigma
\]

If \( a \approx b > 10 \), then \( r_0 \) will be normally distributed with mean \( r_0 = 0 \) and \( \sigma^2 = 1 \) [the so-called \( N(0,1) \) distribution]. If \( a < b \) or
If $a + b < 20$, $r_{o}$ will not be normally distributed and the probability density must be calculated directly from combinatorial formulas. (See Appendix.) In this paper, only those protein sequences with $a$ and $b$ to be used in order to assure proper conditions for $r_{o}$ to be normally distributed when $a = b$. This is not a strong restriction when considering the distributions of individual amino acids which have a high frequency of occurrence or classes of amino acids such as the charged ones. It becomes somewhat restrictive when individual amino acids of low frequency are considered in that only longer sequences are used. A more difficult problem for low frequencies ($a < b$) is that $r_{o}$ is no longer N(0,1). As discussed in the Appendix, the distribution becomes bimodal in such cases. However, because $r_{o}$ inherently compares the observed with the expected number of runs, its mean value for a population indicates the general tendencies of the population independent of the shape of the probability density.

It is often convenient to consider the fraction $f$ of ones in a sequence rather than the number. In this case $f = a/L$ and equations (1) and (2) become

$$\mu_{f} = \frac{2f(1 - f)}{L + 1}, \quad (1a)$$
$$\sigma^{2} = \frac{2f(1 - f)(2f - 1)}{L - 1}, \quad \sigma = \frac{\sqrt{\sigma^{2}}}{L^{2}} \quad (2a)$$

**Interpretation of Run Test Results.** We show in Fig. 1 the application of the run test to 100 Bernoulli sequences of length 24 and $f = 0.5$ ($a = b = 24$) generated by randomly sorting 100 times a vector consisting of 12 ones followed by 12 zeroes. The run test statistic $r_{o}$ (eq. (3)) was determined for each sequence, the $r_{o}$ values were grouped together into bins (class intervals) of width $\Delta r_{o} = 0.5$, and the number density $n(r_{o})$ was determined. $n(r_{o})$ is shown in Fig. 1 as the solid squares; the vertical lines through points show the expected standard deviation of $n(r_{o})$ for multiple trials taken as $\sqrt{n}$. The shaded bars show the number density expected for the perfect N(0,1) distribution. A normal distribution with $r_{o} = 0.005$ and $\sigma = 1.08$ could be fitted to the experimental points by $\chi^{2}$ minimization, which yielded a reduced $\chi^{2}(r_{o})$ of 0.65. The value of $\sigma$ is somewhat greater than 1.00 but, as we discuss below, is within the range of standard deviations expected for multiple 100-sequence trials.

Three examples of binary sequences (not produced randomly) and their values of $r_{o}$ are shown in the bottom half of Fig. 1. The top sequence shows that if the 1's cluster together too much, i.e., if there are too few runs, then $r_{o}$ is negative. In this particular case $r_{o} = -4.17$ or 4.17$\sigma$ away from the expected value, indicating that the occurrence of this sequence by chance is very unlikely. The middle sequence with $r_{o} = +4.02$ can be considered random because $r_{o}$ is well within $1\sigma$ of the expected value. The bottom sequence consisting of alternating 1's and 0's has too many runs and with $r_{o} = +4.59$ is very unlikely to occur by chance. These examples show that if $r_{o}$ is large and negative, the 1's (i.e., amino acids of interest) are strongly clustered together whereas if $r_{o}$ is large and positive, the 1's do not cluster enough and are too evenly spaced along the sequence.

To make the decision as to whether or not a single sequence can be considered as random, one adopts the null hypothesis $H_{0}$ that the sequence is random. Because $r_{o}$ is normally distributed when $H_{0}$ is true, the two-tailed probability $P(r > r_{o})$ of a random sequence having a value of $r$ greater than the observed $r_{o}$ can be easily calculated. $H_{0}$ is rejected if $P(r > r_{o}) \leq \alpha$ where $\alpha$ is the level of significance which is frequently taken as 0.05. The practical implication of this choice is that for any single sequence (one not considered in the context of the whole population) $H_{0}$ is rejected if $|r_{o}| \geq 2$. The choice of $\alpha$ is an arbitrary one and there is obviously a hazard in drawing conclusions about the origins of a single sequence (i.e., from a random or nonrandom process). This is made clear by the results shown in Fig. 1 where all of the sequences were produced by a single random process; the sequences with $|r_{o}| \geq 2$ must be just as random as the others. It is therefore more meaningful to examine collections or families of sequences.

If the number density $n(r_{o})$ of a single homogeneous collection of sequences is N(0,1) as in the example of Fig. 1, one can conclude that the collection is equivalent to one produced by a single random process. A collection comprised of several homogeneous families, each produced by a single random process, will necessarily also be N(0,1) because each family is. In either case, all of the sequences must be considered as examples of random sequences. If the sequences in a collection were produced by a nonrandom process or by some combination of processes, then the number density $n(r_{o})$ will not be N(0,1). In general, it is difficult to say much about the exact nature of the processes involved in such complex cases without extensive analysis or simulation. The practical approach, and the one we adopt here,
is to revert to the single-sequence type of analysis: All of the sequences with values of $r_m$ between $+2$ and $-2$ will be considered to be mathematically random in that they have $r_m$ values expected for sequences produced by a random process. The fraction of the total number of sequences lying within this range will be used as a general measure of the randomness of the collection. In many cases, the distributions will tend to be $N(r_m;\sigma^2)$ so that the values of $r_m$ and $\sigma$ provide useful information about the general tendencies of the collection. For example, if $r_m < 0$ one can assume a general tendency for the collection to have fewer runs of 1’s and 0’s, so the clustering of the 1’s and 0’s is larger than expected.

Comparing Means and Standard Deviations of Densities. We use the central limit theorem to examine the equivalence of sample means by defining the random variable $z = (r_m - r_{sm})/(\sigma_m/\sqrt{m})$ where $r_m$ and $\sigma_m$ are the mean and standard deviation of the parent population and $r_m$ and $\sigma_{sm}$ are the values for the derived population (Wani 1971; Barlow 1989). The distribution of $z$ for large $m$ is $N(0,1)$, so the two-tailed probability $P(|z| < d)$ that $|z|$ is smaller than $d$ standard deviations can be calculated. We take $d = 2$ so there is a 95% chance that $|z| < 2$. Therefore, the expected range of $r_m$ from the definition of $z$ is

$$|r_m - r_{sm}|/\sqrt{m}\sigma_m \leq 2$$

(4)

In the example shown in Fig. 1, there are $m = 100$ sequences, so the maximum allowable range of $r_m$ is $-0.2$ to $+0.2$. For the Superfamily Set of about 1,800 sequences, the range of $r_m$ is $-0.05$ to $+0.05$ assuming the parent set of proteins (the extant proteins) from which this sample is drawn is $N(0,1)$.

The allowable range of variation in $\sigma_m$ is calculated from a well-known theorem (Wani 1971; Barlow 1989) which states, in terms of the above definitions, that the function $x^2 = (m - 1)(\sigma_m)^2/(\sigma_{x^2})^2$ has a $\chi^2$ distribution with $(m - 1)$ degrees of freedom (Wani 1971). When the number of degrees of freedom is large (greater than 30), $\sqrt{(2\chi^2)}$ approaches a normal distribution with mean $\sqrt{(2k - 1)}$ and unit variance. To include 95% of random observations, we require that $\chi^2$ be within 2 standard deviations of the expected $\chi^2$ so that the allowable range of $\sigma_m$ under the null hypothesis that the populations have equivalent statistical properties is not rejected. The range of $\sigma_m$ is easily shown to be

$$\sigma_m/\sigma_{sm} = \sqrt{(2k - 1)} \pm 2\sqrt{(2k)}$$

(5)

Returning again to the example of Fig. 1, for $\sigma_{sm} = 1$ and $k = 99$, $\sigma_m$ may range from $0.86$ to $1.14$. The observed value is $1.08$. For the Superfamily Set of proteins under the null hypothesis that they are drawn from a parent population which is $N(0,1)$, the allowable range of $\sigma_m$ is $0.967$ to $1.033$ or $|\sigma_m - 1| \leq 0.033$.

Autocorrelation and Fourier Analyses. We will consider in the Discussion the relation between sequence randomness as judged by the run test and periodic patterns of amino acids detected by autocorrelation and Fourier analyses. Macchiato et al. (1985) have summarized the autocorrelation analysis of protein sequences and the problem of establishing the significance of autocorrelation orders relative to the random expectation. The autocorrelation coefficient of order $K$ for an ordered series of $L$ values $y_1, \ldots, y_j, \ldots, y_L$ is given by

$$r_K = \frac{\sum_{j=1}^{L-K} (y_j - y_{sm})(y_{j+K} - y_{sm})}{(1/L) \sum_{j=1}^{L} (y_j - y_{sm})^2}$$

(6)

where $y_{sm}$ is the mean value given by $(\Sigma y_j)/L$ and the maximum value of $K$ is $K_{max} = L/10$. The expected range of fluctuation of $r_K$ for random sequences which will include 95% of the observations is given by

$$r_K(0.95) = (-1 \pm 1.645 \sqrt{(L - K - 1)/L - K})$$

(7)

so that a value of $r_K$ outside this range indicates autocorrelation of the series at the $K$th order.

McLachlan and Stewart (1976) used Fourier analysis to find periodicities in α-tropomyosin and recognized that random sequences can have periodicities which must be taken into account during the assignment of significance levels to Fourier coefficients. This is a special case of so-called time-series analysis, which has been important in spectral analyses of communication signals for many years. (See, for example, Parzen 1967). The formulation of McLachlan and Stewart (1976) is particularly useful because they derived analytical expressions for random binary sequences. The complex Fourier transform of such a sequence is

$$C_j = (1/L) \sum_{j=1}^{L} y_j \exp(2\pi i j m/L)$$

(8)

where $j = 0, 1, \ldots, (L - 1)$. We will be particularly interested in the intensities

$$I_m = |C_m|^2$$

(9)

If the values of $y_j$ are either 0 or 1, so that the sequence consists of a 1’s and b 0’s, the intensities will be largest if $f = a/L = 0.5$ and will vanish if $a = 0$ or $b = 0$. For a random sequence with $a$ and $L$ large, McLachlan and Stewart (1976) show that the probability density $P(I_m)$ of the intensities $I_m$ will be exponential:

$$P(I_m) = (1/e) \exp(-I_m/J)$$

(10)

where $J^2$ is the variance of the intensities. Defining $I$ as the mean intensity, it can be shown that $J = I = f(1 - f)/L$. It is convenient to normalize the $I_m$ with respect to the mean intensity by defining $I_m' = I_m/I$ so that $I_m'$ has a mean and variance of 1. Strong intensities are considered to be those that are two or more standard deviations above the mean; i.e., $I_m' > 3$. We write equation (10) in the form

$$P(I_m') = (1/\beta) \exp(-I_m'/\beta)$$

(11)

which we fit to the observed distribution of intensities using $\chi^2$ minimization. If the sequence is "perfectly" random, then $\beta = 1$. 
Amino Acid Class Assignments


Computations. The database analyses were carried out using a Sun4/110 computer (Sun Microsystems, San Jose, CA). Programs were written in the C language that extracted the sequences from the databases, selected those that satisfied the criteria described earlier, calculated the run test $r_o$ values, and accumulated the data for the number density $n(r_o)$ histograms. The class intervals (bins) for the histograms were of size $\Delta r_o = 0.08$ except for $r_o > 3.24$ and $r_o < -3.24$ which were aggregated into single class intervals. The total number of class intervals was 83, giving $k = 80$ degrees of freedom. The time required to carry out these operations for a single class of amino acids for the entire 1990 PIR database was less than 5 min.

The shapes of the histograms were usually approximately Gaussian. The best Gaussian curves were fitted to the histogram data by $\chi^2$ minimization on an 80386 microcomputer using the program FitAll (v5.01) by MTR Software (Toronto, Ontario). Routines were written for FitAll in Turbo Pascal (v6.0, Borland International, Scotts Valley, CA), which determined the means and standard deviations of the Gaussians with their areas constrained to be equal to the number of sequences multiplied by the class interval size.

Results

Comparisons of the Databases

Two issues were of immediate concern to us. First, is the PIR database statistically stable? Second, what is the effect of the uneven representation of superfamilies in a database? To answer these questions, we examined the run test statistics of our various versions of the databases using HFOB-4 (A, V, I, L, F, C, M, W, Y, P, G), which we used in our earlier study (White and Jacobs 1990). The results are summarized in Table 1. The number densities of runs for the 1988, 1989, and 1990 PIR databases are shown graphically in Fig. 2A along with plots of the $N(0,1)$ density (dashed curve) and best-fit Gaussian $[N(r_o/\sigma)]$, solid curves). The gross similarities of the data as the size of the PIR database has increased are apparent. The data of Table 1 indicate that as the size has increased, $r_o$ has decreased from 0.21 to 0.17 while $\sigma$ has increased from 1.18 to 1.24. However, in all cases the number of sequences between $r_o = \pm 2$ remained steady at about 0.89, meaning that 89% of the sequences cannot be distinguished from the random expectation. The average length of the sequences has grown from 290 to 330 with the expansion of the database but the average fraction of the HFOB-4 residues has remained steady at 0.54.

The behavior of the 1989 and 1990 Superfamily Sets is shown in Fig. 2B and Table 1 where, again, the gross properties of the densities are seen to be very similar. The densities are similar to the complete PIR sets except that the $\sigma$ are significantly smaller and the densities are closer to normal as judged by the reduced $\chi^2$ values. The same general trends are seen in the lengths, fraction of hydrophobes, and mean $r_o$ as for the complete databases. The fraction of sequences which qualify as random has risen slightly to about 91% but is still smaller than the 95% expected for the perfect $N(0,1)$ distribution. One of the major reasons for this is the exceptionally large number of sequences with extreme $r_o$ values (points surrounded by rectangles) which is observed for all versions of the databases.

These results show that the PIR database is remarkably stable and is unlikely to change drastically as more sequences are added, and the results suggest that it represents a reasonable sampling of the extant proteins. Comparing the full PIR sets with the Superfamily Sets shows that the two types of sets are not much different and that uneven representation of the superfamilies is not a severe problem. We nevertheless felt it prudent to use the Superfamily Sets, which has the additional advantage of reducing computational time. The differences between the 1989 and 1990 Superfamily Sets are not great; the results reported here using the 1989 Set are unlikely to change much as the size of the Superfamily Set increases.

Table 1. Summary of the statistical properties of the databases

<table>
<thead>
<tr>
<th>Database</th>
<th>$N$</th>
<th>$f_{HF}$</th>
<th>$L$</th>
<th>$f_{x2}$</th>
<th>$r_m$</th>
<th>$\sigma$</th>
<th>$r_{x2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1988 PIR</td>
<td>5247</td>
<td>0.540</td>
<td>290</td>
<td>0.90</td>
<td>0.21</td>
<td>1.19</td>
<td>2.22</td>
</tr>
<tr>
<td>1989 PIR</td>
<td>8752</td>
<td>0.539</td>
<td>321</td>
<td>0.89</td>
<td>0.20</td>
<td>1.22</td>
<td>3.22</td>
</tr>
<tr>
<td>1990 PIR</td>
<td>14,620</td>
<td>0.542</td>
<td>330</td>
<td>0.89</td>
<td>0.17</td>
<td>1.24</td>
<td>3.79</td>
</tr>
<tr>
<td>1989 SF</td>
<td>2278</td>
<td>0.542</td>
<td>319</td>
<td>0.91</td>
<td>0.21</td>
<td>1.09</td>
<td>1.70</td>
</tr>
<tr>
<td>1990 SF</td>
<td>2272</td>
<td>0.544</td>
<td>323</td>
<td>0.91</td>
<td>0.19</td>
<td>1.12</td>
<td>1.59</td>
</tr>
</tbody>
</table>

* See Methods for descriptions of databases. $N$ is the number of sequences in the database. $f_{HF}$ fraction of HFOB-4 residues. $L$ average length of sequences. $f_{x2}$ fraction of sequences with $|r_m| < 2$. $r_m$ mean of the distribution. $\sigma$ the standard deviation of the distribution. and $r_{x2}$ the reduced $\chi^2$ (i.e., $\chi^2 + k$ where $k = 80$ is the number of degrees of freedom)
Distributions of the Amino Acids by Class

Figure 3 shows examples of the $n(r_o)$ densities for four classes of amino acids (HFOB-2, CHARG-A, ALFA, and VOL-3) and Fig. 2 the density for HFOB-4. The densities for the other classes have the same general appearances as these and, likewise, tend to be normal. The fits of the densities to $N(r_o, \sigma^2)$ for all of the classes are summarized in Table 2. The number of sequences used varies between classes because of the constraints placed on qualifying sequences. (See Methods.) The fraction $f_{\leq 2}$ of the sequences in each class with $|r_o| \leq 2.0$ (Table 2) is at least 0.91 for all classes.

Two questions arise concerning the fits of the data to $N(r_o, \sigma^2)$ distributions. First, is the true distribution actually normal and, second, are $r_m$ and $\sigma$ statistically different than 0 and 1, respectively? The first question can be answered by examining the $r_X^2$ value ($=r_X^2$) obtained from the fitting of the data to $N(r_m, \sigma^2)$ and the second by the criteria discussed in Methods. The data for the fits described in Table 2 have been placed in 83 class intervals so that the number of degrees of freedom $k = 80$. For $P(r_X^2 \geq r_X^2) > 0.05$, $r_X^2$ must be less than 1.30. By this criterion, only the classes HFOB-2, ALFA, BETA, VOL-1, and VOL-2 can be considered to be adequately described by a normal $N(r_m, \sigma^2)$ distribution. The reason the distributions are not generally normal, however, is obvious from the examination of Figs. 2 and 3, where it can be seen that there are far more points with $|r_o| > 3.24$ than expected by chance. If one ignores these points, the shapes of the densities are much closer to normal. Despite the fact that the majority of the densities are significantly different than normal, one must be impressed by the general appearance of the densities (Figs. 2 and 3) and by the fact that $f_{\leq 2} \geq 0.90$ (Table 2). It is reasonable to treat the densities as normal to a first approximation. Regarding the second question, we showed in Methods that the rejection of the null hypothesis that $N(r_m, \sigma^2) = N(0,1)$ required $|r_m| \geq 0.05$ and $|\sigma - 1| \geq 0.033$. Assuming the observed densities to be approximately normal, the classes that could be considered drawn from a randomly distributed set would be CHARG-N and VOL-3 and perhaps BETA and R-TURN. None of the sets qualify completely as being $N(0,1)$ although BETA comes closest overall.

The densities with the most deviant means are HFOB-4 with $r_m = 0.21$ and ALFA with $r_m = -0.37$ (uncertainties are $\pm 0.02$), which are about four and sevenfold greater than the significance criterion of eq. (4). The density for the ALFA set (Fig. 3C) is particularly striking in appearance because it is precisely normal in shape ($r_X^2 = 0.97$, Table 2) while its points are neatly shifted in a systematic way to the left. Negative values of $r_m$ indicate that the general preference of the population of proteins is for the amino acids of the class to cluster together more than chance would predict. Therefore, the ALFA amino acids have a very strong tendency to cluster. The large positive value of $r_m$ for the HFOB-4 set indicates that these amino acids have a strong anticlustering tendency. This appears to result from the inclusion of Gly and Pro in the set judging by the much smaller value of $r_m$ for the HFOB-3 set. The implication of this observation is...
that Gly and Pro tend not to be “natural partners” with the members of the HFOB-3 set. It is perhaps significant that Gly and Pro tend to be strong breakers of secondary structure.

CHARG-A, CHARG-P, VOL-2, and VOL-4 show strong clustering tendencies while BETA and R-TURN show only modest tendencies in this direction. It is interesting that CHARG-N shows no tendency toward clustering or anticlustering despite the clustering behavior of CHARG-A and CHARG-P. Also of interest is the fact that of all the classes, only the hydrophobes show anticlustering tendencies although HFOB-1 and HFOB-3 do this to a much smaller extent than HFOB-4. This suggests that runs of hydrophobes tend to be generally broken by other types of residues.

Distributions of the Individual Amino Acids
The $n(r_o)$ densities for the individual amino acids are strikingly different from those for classes of amino acids. As illustrated by the examples shown in Fig. 4, the densities can be generally characterized as bimodal with a sharp cutoff for $r_o \geq 2$. The extent and nature of the bimodality varies among the amino acids. The extremes of behavior are shown in Fig. 4 by Ala (little bimodality) and Phe (large bimodality); the most typical behavior is that of Asp (Fig. 4B). For the purpose of characterizing the bimodal behavior, we fitted a sum of two Gaussians to each density by $\chi^2$ minimization. These fits are summarized in Table 3 and shown by the solid curves in Fig. 4. The relative areas of the two Gaussians are indicated in Table 3 by the parameter $F_i = N_i/N$ where $N$ is the total number of sequences and $N_i$ the number of sequences associated with Gaussian $i$ ($i = 1$ or 2). The means of Gaussian 1 are all negative (average $r_{m1} = -0.48$) with typical $\sigma$ values of about 1 whereas they are positive (average $r_{m2} = 0.79$) with typical $\sigma$ of 0.25 for Gaussian 2. The average relative areas of the Gaussians are $F_1 = 0.63$ and $F_2 = 0.37$.

We show in the Appendix that this bimodal behavior is expected when the fraction $f$ of amino acids in a sequence assigned a value of 1 is small ($f < 0.2$). It results from the fact that there is a relatively small probability of strong clustering when the den-
Table 2. Summary of run test $n(r_o)$ distributions for the Superfamily Set of proteins

<table>
<thead>
<tr>
<th>Class</th>
<th>$N$</th>
<th>$f_{2.2}$</th>
<th>$r_m$</th>
<th>$\sigma$</th>
<th>$r^2_{\chi}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>HFOB-1</td>
<td>1751</td>
<td>0.91</td>
<td>0.05</td>
<td>1.11</td>
<td>1.64</td>
</tr>
<tr>
<td>HFOB-2</td>
<td>1771</td>
<td>0.91</td>
<td>-0.03</td>
<td>1.11</td>
<td>1.21</td>
</tr>
<tr>
<td>HFOB-3</td>
<td>1781</td>
<td>0.91</td>
<td>0.08</td>
<td>1.12</td>
<td>1.40</td>
</tr>
<tr>
<td>HFOB-4</td>
<td>1789</td>
<td>0.91</td>
<td>0.21</td>
<td>1.09</td>
<td>1.70</td>
</tr>
<tr>
<td>CHRG-A</td>
<td>1727</td>
<td>0.93</td>
<td>-0.14</td>
<td>1.03</td>
<td>1.44</td>
</tr>
<tr>
<td>CHRG-P</td>
<td>1631</td>
<td>0.93</td>
<td>-0.15</td>
<td>1.04</td>
<td>1.70</td>
</tr>
<tr>
<td>CHRG-N</td>
<td>1490</td>
<td>0.93</td>
<td>-0.01</td>
<td>0.98</td>
<td>1.64</td>
</tr>
<tr>
<td>ALFA</td>
<td>1775</td>
<td>0.93</td>
<td>-0.37</td>
<td>1.03</td>
<td>0.97</td>
</tr>
<tr>
<td>BETA</td>
<td>1739</td>
<td>0.94</td>
<td>-0.07</td>
<td>1.01</td>
<td>0.84</td>
</tr>
<tr>
<td>R-TURN</td>
<td>1751</td>
<td>0.94</td>
<td>-0.06</td>
<td>1.03</td>
<td>1.44</td>
</tr>
<tr>
<td>VOL-1</td>
<td>1766</td>
<td>0.93</td>
<td>-0.04</td>
<td>1.06</td>
<td>1.11</td>
</tr>
<tr>
<td>VOL-2</td>
<td>1736</td>
<td>0.94</td>
<td>-0.19</td>
<td>1.02</td>
<td>1.28</td>
</tr>
<tr>
<td>VOL-3</td>
<td>1724</td>
<td>0.94</td>
<td>0.00</td>
<td>1.01</td>
<td>1.50</td>
</tr>
<tr>
<td>VOL-4</td>
<td>1726</td>
<td>0.93</td>
<td>-0.13</td>
<td>1.05</td>
<td>1.45</td>
</tr>
</tbody>
</table>

* See footnote for Table 1

The bimodal nature of the densities, the average fraction $f_{2.2}$ of the total number of sequences which are mathematically random is 0.94 (range: 0.89–0.97). This observation and the similar one for classes of amino acids strongly suggest that, overall, the preferred lengthwise distributions of amino acids are those commensurate with the random expectation. However, as for the classes, some of the individual amino acids show a strong global tendency to cluster or to avoid clustering. These tendencies can be inferred from the mean values of $r_o$ ($r_{max}$ in Table 3), recalling that $r_o$ in eq. (3) is zero when the observed number of runs equals the expected number of runs. The statistical significance ($\alpha = 0.05$) of the values of $r_{max}$ in Table 3 can be estimated from eq. (4); for $N = 1000$, $|r_{max}| > 0.06$ for the mean to be significantly different from zero. By this criterion, eight amino acids (Ala, Arg, Gln, Glu, His, Lys, Met, Ser) show a strong tendency to cluster while only one (Ile) shows a strong tendency against clustering. Excluding Ser and Arg, the remaining strongly clustering members of the ALFA set of amino acids (A, Q, L, E, H, M, K), which we showed earlier to have exceptionally strong clustering tendencies as a group. Thus, except for Leu, all of the members of the ALFA set show a strong tendency to cluster individually as well as a group. The clustering tendency of Met with $r_{max} = -0.75$ ($|r_{max}| > 0.09$ for significance) is exceptionally strong. Interestingly, Met also has the highest $\alpha$-helix propensity on the Levitt (1978) scale.

Discussion

General Statistical Features of Protein Sequences

The vast majority of the protein sequences, considered individually, have lengthwise distributions of amino acids that cannot be distinguished from the random expectation. We indicated earlier that we use the term random to mean a sequence of independent events. Such sequences are referred to as being stochastic and it is therefore reasonable to refer to protein sequences in this way. Despite the stochastic nature of the individual sequences, the data for the collection of proteins considered as a whole show that certain types of lengthwise distributions have been subjected to selective pressure. This is evident in three ways from the data presented. First, there are far more proteins with $|r_o| > 3.24$ than would be expected by chance and the number grows in proportion to the database size. This presumably reflects special advantages and properties of these proteins. Second, the mean values of $r_o$ for nine of the single amino acids are statistically very different from zero. Third, the densities for classes of amino acids show that certain classes (such as the ALFA Set) prefer to cluster while others (such as the HFOB-4 Set) prefer to be broken up in a somewhat even fashion. This result suggests the run test as a statistical method for finding classes of amino acids whose clustering may be structurally advantageous but which have not previously been apparent. It is feasible to perform the run test on a sample of a hundred or so proteins of known structure to look at all possible combinations of amino acids to find those that prefer to cluster (unpublished experiments).

While the ALFA Set of amino acids shows a strong and systematic tendency to cluster, the BETA and R-TURN sets show this tendency to a much smaller degree. This is somewhat surprising because the BETA and R-TURN sets were selected by the same statistical criteria as the ALFA Set (Levitt 1978). Charged residues as a group (CHARG-A) also have a tendency to cluster. Interestingly, CHARG-P (positive) amino acids have a tendency to cluster but CHARG-N (negative) do not. All of the positively charged amino acids have a strong tendency to cluster individually (Table 3) while only Glu of the negatively charged ones shows this tendency. The VOL-2 (P, V, D, N, and
Fig. 4. Examples of \( n(r_i) \) densities for four amino acids. The solid lines represent the fit of a sum of two Gaussian curves to the data and the dotted line the \( N(0,1) \) distribution. Alanine (panel A) and methionine (panel D) are the extreme examples; aspartate (B) and phenylalanine (C) show the typical behaviors. The two-Gaussian fits for all of the amino acids are summarized in Table 3. All of the densities are strongly bimodal with a sharp cutoff for \( r_i > 2 \), but the fraction of sequences with \( |r_i| < 2 \) is 90% or greater. The number of sequences used is determined by the mathematical criteria of the run test. (See text.) The sharp cutoff and bimodal behavior are expected for sequences in which the fraction of amino acids assigned a 1 is smaller than 0.1 to 0.2. (See Appendix and Fig. 7.)

E) and VOL.4 (F, K, Y, R, and W) sets also show a moderately strong tendency to cluster. The latter result is not surprising in view of the tendency of positively charged amino acids to cluster but it does suggest a possible preferred association of Phe, Tyr, and Trp with positively charged amino acids. Because the negatively charged amino acids do not have a strong tendency to cluster, the tendency of the VOL.2 set to cluster suggests a tendency of Pro, Val, and Asn to associate with negatively charged amino acids.

Besides revealing global tendencies of classes of amino acids for the Superfamily Set of proteins, the run test analysis appears to provide a convenient means of characterizing any set of protein sequences statistically. The collection of statistical parameters for the classes can logically serve as a means of "fingerprinting" a set of proteins, and we suggest that the parameters of the Superfamily Set can serve as reference standards. This may be particularly useful for assessing the extent to which a collection of proteins of known structure is representative of the Superfamily Set and, perhaps, the extant proteins. Preliminary examinations of several relatively large sets of proteins with structures known to high resolution reveal significant deviations from the Superfamily Set (unpublished observation).

Regarding the behavior of the single amino acids, eight of them (A, R, Q, E, H, K, M, S) show strong clustering tendencies while only Ile shows strong anticlustering. All of the ALFA set except for Leu are included with the clusterers, indicating most of the ALFA set members prefer clustering individually as well as a group. It is interesting in this regard that Ile, the lone anticlusterer, has a strong \( \beta \)-sheet propensity. The exceptional members of the group of individual clusterers are Arg and Ser in that they are not members of the ALFA Set; Arg is neutral for \( \alpha \)-helix formation and Ser is a strong breaker according to the Levitt (1978) scale. The behavior of Met is particularly interesting in that it has both the strongest clustering tendency and \( \alpha \)-helix propensity. Because Met appears to play a special role
<table>
<thead>
<tr>
<th>AA</th>
<th>Len.</th>
<th>Freq.</th>
<th>( r_{nn} )</th>
<th>( f_{z,2} )</th>
<th>( F_1 )</th>
<th>( r_{m1} )</th>
<th>( \sigma_1 )</th>
<th>( F_2 )</th>
<th>( r_{m2} )</th>
<th>( \sigma_2 )</th>
<th>( r_{max} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala</td>
<td>1301</td>
<td>398</td>
<td>0.090</td>
<td>-0.23</td>
<td>0.93</td>
<td>0.84</td>
<td>-0.35</td>
<td>0.96</td>
<td>0.16</td>
<td>1.00</td>
<td>0.29</td>
</tr>
<tr>
<td>Arg</td>
<td>1129</td>
<td>426</td>
<td>0.068</td>
<td>-0.15</td>
<td>0.94</td>
<td>0.78</td>
<td>-0.33</td>
<td>0.93</td>
<td>0.22</td>
<td>0.89</td>
<td>0.24</td>
</tr>
<tr>
<td>Asn</td>
<td>901</td>
<td>487</td>
<td>0.052</td>
<td>+0.06</td>
<td>0.97</td>
<td>0.61</td>
<td>-0.32</td>
<td>0.86</td>
<td>0.39</td>
<td>0.87</td>
<td>0.24</td>
</tr>
<tr>
<td>Asp</td>
<td>1037</td>
<td>456</td>
<td>0.060</td>
<td>-0.04</td>
<td>0.94</td>
<td>0.74</td>
<td>-0.18</td>
<td>0.83</td>
<td>0.26</td>
<td>0.91</td>
<td>0.26</td>
</tr>
<tr>
<td>Cys</td>
<td>257</td>
<td>704</td>
<td>0.036</td>
<td>+0.02</td>
<td>0.93</td>
<td>0.44</td>
<td>-0.46</td>
<td>1.58</td>
<td>0.56</td>
<td>0.54</td>
<td>0.12</td>
</tr>
<tr>
<td>Gln</td>
<td>851</td>
<td>499</td>
<td>0.048</td>
<td>-0.29</td>
<td>0.91</td>
<td>0.64</td>
<td>-0.64</td>
<td>0.95</td>
<td>0.36</td>
<td>0.79</td>
<td>0.19</td>
</tr>
<tr>
<td>Glu</td>
<td>1144</td>
<td>428</td>
<td>0.069</td>
<td>-0.18</td>
<td>0.93</td>
<td>0.77</td>
<td>-0.36</td>
<td>0.95</td>
<td>0.23</td>
<td>0.98</td>
<td>0.23</td>
</tr>
<tr>
<td>Gly</td>
<td>1203</td>
<td>421</td>
<td>0.076</td>
<td>-0.03</td>
<td>0.94</td>
<td>0.79</td>
<td>-0.22</td>
<td>0.92</td>
<td>0.21</td>
<td>0.99</td>
<td>0.26</td>
</tr>
<tr>
<td>His</td>
<td>471</td>
<td>631</td>
<td>0.031</td>
<td>-0.19</td>
<td>0.92</td>
<td>0.41</td>
<td>-1.02</td>
<td>0.76</td>
<td>0.59</td>
<td>0.65</td>
<td>0.16</td>
</tr>
<tr>
<td>Ile</td>
<td>1097</td>
<td>438</td>
<td>0.064</td>
<td>+0.10</td>
<td>0.97</td>
<td>0.72</td>
<td>-0.09</td>
<td>0.81</td>
<td>0.28</td>
<td>0.96</td>
<td>0.26</td>
</tr>
<tr>
<td>Leu</td>
<td>1415</td>
<td>379</td>
<td>0.100</td>
<td>+0.04</td>
<td>0.95</td>
<td>0.57</td>
<td>-0.38</td>
<td>1.00</td>
<td>0.43</td>
<td>0.69</td>
<td>0.59</td>
</tr>
<tr>
<td>Lys</td>
<td>1106</td>
<td>425</td>
<td>0.069</td>
<td>-0.09</td>
<td>0.94</td>
<td>0.68</td>
<td>-0.34</td>
<td>0.91</td>
<td>0.32</td>
<td>0.81</td>
<td>0.33</td>
</tr>
<tr>
<td>Met</td>
<td>525</td>
<td>604</td>
<td>0.032</td>
<td>-0.75</td>
<td>0.89</td>
<td>0.28</td>
<td>-1.57</td>
<td>0.92</td>
<td>0.72</td>
<td>-0.21</td>
<td>0.41</td>
</tr>
<tr>
<td>Phe</td>
<td>843</td>
<td>506</td>
<td>0.046</td>
<td>+0.00</td>
<td>0.97</td>
<td>0.63</td>
<td>-0.37</td>
<td>0.89</td>
<td>0.37</td>
<td>0.81</td>
<td>0.18</td>
</tr>
<tr>
<td>Pro</td>
<td>981</td>
<td>466</td>
<td>0.059</td>
<td>+0.03</td>
<td>0.94</td>
<td>0.64</td>
<td>-0.31</td>
<td>1.03</td>
<td>0.36</td>
<td>0.86</td>
<td>0.22</td>
</tr>
<tr>
<td>Ser</td>
<td>1244</td>
<td>413</td>
<td>0.074</td>
<td>-0.15</td>
<td>0.93</td>
<td>0.80</td>
<td>-0.32</td>
<td>0.98</td>
<td>0.20</td>
<td>0.96</td>
<td>0.21</td>
</tr>
<tr>
<td>Thr</td>
<td>1128</td>
<td>438</td>
<td>0.064</td>
<td>+0.01</td>
<td>0.96</td>
<td>0.73</td>
<td>-0.20</td>
<td>0.87</td>
<td>0.27</td>
<td>0.97</td>
<td>0.26</td>
</tr>
<tr>
<td>Trp</td>
<td>210</td>
<td>795</td>
<td>0.023</td>
<td>+0.07</td>
<td>0.95</td>
<td>0.22</td>
<td>-1.40</td>
<td>0.95</td>
<td>0.78</td>
<td>0.55</td>
<td>0.14</td>
</tr>
<tr>
<td>Tyr</td>
<td>687</td>
<td>545</td>
<td>0.040</td>
<td>-0.05</td>
<td>0.95</td>
<td>0.57</td>
<td>-0.48</td>
<td>0.92</td>
<td>0.43</td>
<td>0.77</td>
<td>0.18</td>
</tr>
<tr>
<td>Val</td>
<td>1235</td>
<td>413</td>
<td>0.073</td>
<td>-0.02</td>
<td>0.96</td>
<td>0.72</td>
<td>-0.33</td>
<td>0.92</td>
<td>0.28</td>
<td>0.93</td>
<td>0.35</td>
</tr>
</tbody>
</table>

* See footnote for Table 1. In addition, “Len.” is the average length, “Freq.” is the frequency (probability) of occurrence of the amino acid in the population, and \( r_{nn} \) is the mean value of \( r_c \) for the population. The values of \( r_{m1} \) which are significantly different than zero are italicized. \( N \) varies quite significantly because of the requirement that there be at least 10 amino acids of the chosen type in the sequence; it is therefore inversely related to amino acid frequency. The distributions were fitted using the sum of two normal distributions \( N(r_{m1}, \sigma_1) \) and \( N(r_{m2}, \sigma_2) \) with relative areas \( F_1 \) and \( F_2 \) (i.e., \( F_1 + F_2 = 1 \)). \( r_{max} \) is the maximum obtainable value of \( r_c \) calculated from the average lengths and frequencies (see Appendix); note that \( r_{m2} \) is always smaller than \( r_{max} \).

in the sequence-independent recognition of nonpolar protein surfaces (Gellman 1991), its strong global clustering tendency is particularly intriguing.

There are clearly many nonrandom aspects of the Superfamily Set of proteins but these appear as rather diffuse tendencies of the whole population. The predominant feature of the population is that the lengthwise distributions that are most common are those consistent with distributions classified as random by the run test. Whether the proteins started their existence in that condition or evolved to it, the clear message is that the random-like distribution is the preferred one. Therefore, many lengthwise patterns of amino acids which are structurally and functionally important may have the mathematical characteristics of random sequences. There is no reason a priori to expect all important structural signals to have statistically unusual patterns. This has two important implications. First, it will be difficult to devise successful statistics-based prediction algorithms which implicitly or explicitly use the random expectation as the reference state except in those cases for which the structural or functional signals have very nonrandom characteristics. Karlin and his colleagues (Karlin and Altshul 1990; Karlin et al. 1991) have developed elegant algorithms of this type. But, it seems clear from the work presented here that many signals will not be of this type. The most successful statistical algorithms are likely to be those that can distinguish distributions whose properties are known or suspected in advance from physicochemical or other considerations. Second, these results provide an additional insight into the difficult nature of the secondary structure prediction problem. The highly stochastic nature of sequences will cause them to appear to be noisy in secondary structure prediction algorithms and this must certainly contribute to prediction inaccuracy. However, it must be emphasized that the “sequence noise” is not the noise we usually think of which carries no useful information. Much of what appears to be noise may actually be signal. We suggest that the sequences are rich in information and that the simultaneous presence of many “signals” gives the appearance of noise. The prediction problem is that of disentangling the various signals.

One-Dimensional Randomness and Three-Dimensional Structure

The inescapable conclusion of this article is that most protein sequences are mathematically random in one dimension as judged by the run test. However, even the most casual observer of protein sequences will find apparently periodic or quasi-
Fig. 5. The periodicities of the distribution of hydrophobic residues (HF-OB-4 Set) along the sequence of bacteriorhodopsin as revealed by hydrophobicity plots (A), autocorrelation coefficients (B), and Fourier intensities (C). In panel A, the dashed curve is a 30-order Fourier reconstruction of the hydrophathy plot (solid curve) determined by a 21AA sliding window sum of the binary representation of the sequence. Bacteriorhodopsin is random based upon its run test statistic of -0.009 but it nevertheless has periodicities as the Fourier analysis of random sequences suggests it should. (See McLachlan and Stewart 1976.) The Fourier intensities of random sequences are expected to be exponentially distributed (eqs. 10 and 11) and they are as shown in panel D. The expected value of \( B \) in eq. (11) for a random sequence is 1; the observed value for bacteriorhodopsin is 0.953. This figure should be compared to Fig. 6 for bovine rhodopsin, which is highly nonrandom. The behaviors of the two sequences are qualitatively very similar.

periodic patterns of classes of amino acids along sequences. The clearest examples are for hydrophobic residues whose periodicity is revealed by hydrophobicity plots (Rose 1978; Rose and Roy 1980; Engelman et al. 1986). More subtle but equally important periodicities for other properties are invariably revealed for any protein sequence by autocorrelation (Zimmerman et al. 1968; Macchiato and Tramontano 1985) and Fourier (McLachlan and Stewart 1976; Eisenberg et al. 1982) analyses. These periodicities are very real and indicate the basic topologies and structural patterns of proteins in three dimensions. The question thus arises as to whether or not such periodic patterns, observed for virtually all proteins, are inconsistent with the idea of random sequences. The answer must obviously be no because patterns are observed and the sequences are mostly random by the criteria used in this paper. We have previously discussed in detail the origin from random sequences of periodic patterns of hydrophobic residues in the context of transmembrane helices and the secondary structure of soluble proteins (White and Jacobs 1990). To emphasize this point, it is useful and informative to consider hydrophobicity plots, autocorrelation coefficients, and Fourier analyses for two specific proteins.

As we did earlier (White and Jacobs 1990), we consider the distribution of HF-OB-4 residues along the lengths of the sequences of bacteriorhodopsin (BR) (Khorana et al. 1979) and bovine rhodopsin (RHO) (Nathans and Hogness 1983). Both are seven-helix membrane proteins with about the same fraction (≈0.7) of hydrophobic residues. However, BR is perfectly random with \( r_0 = -0.009 \) while RHO is highly nonrandom with \( r_0 = -4.126 \). The results of the hydrophobicity plot, autocorrelation, and Fourier analyses of these proteins based upon the HF-OB-4 binary hydrophobicity scale are shown in Figs. 5 and 6. Despite the difference in the randomness, the general features of the analytical re-
The periodicities of the distribution of hydrophobic residues (HFOB-4 Set) along the sequence of bovine rhodopsin as revealed by hydrophobicity plots (A), autocorrelation coefficients (B), and Fourier intensities (C). In panel A, the dashed curve is a 30-order Fourier reconstruction of the hydropathy plot (solid curve) determined by a 21AA sliding window sum of the binary representation of the sequence. Bovine rhodopsin is very nonrandom based upon its run test statistic of $-4.126$ but its overall behavior is very similar to that of bacteriorhodopsin (Fig. 5). Note that, as for random sequences, the Fourier intensities (panel D) are exponentially distributed (eqs. 10 and 11) but the value of $\beta = 0.865$ in eq. (11) is smaller than the value of 1 expected for a random sequence. This figure should be compared to Fig. 5 for bacteriorhodopsin, which is random. The behaviors of the two sequences differ only in quantitative details.

1. The middle sequence with $r_o = 0.42$ was chosen because the linear pattern of hydrophobes is that expected for the perfect amphipilic helix. Note in this sequence that there tends to be a $\ldots 110011 \ldots$ pattern which is interrupted occasionally by 11011 or 00100 patterns that break the symmetry to give the sequence its random character. This is, of course, only one of a very large number of sequences classified as random by the run test but it demonstrates that random sequences can have periodicities with structural significance in three dimensions. We have observed in extensive simulations using random sequences, however, that the probability of obtaining sequences with a strong amphipilic helix pattern increases drastically in the vicinity of $r_o = 0$. In fact, for sequences of length 12 to 24 and $f = 0.5$, the probability of an amphipilic helix pattern is about 8–12% (unpublished observations). Furthermore, 67% of the random sequences of length 12 had statistically significant peaks using the Fourier-based hydrophobic-moment analysis of Eisenberg et al. (1984) although only 12% had a peak at 100° characteristic of amphipilic $\alpha$-helices. From the point of view of the evolution of amphipilic $\alpha$-helices, this is a very large number.

The results of the autocorrelation analysis of BR and RHO are shown in Figs. 5B and 6B, respectively, where the dotted lines show the expected range of fluctuations in the correlation coefficients.
$r_K$ for random sequences (see Methods). Note that only one of the $r_K$'s is outside the range for BR while eight are out of range for RHO. Also notice that the $r_K$'s fluctuate rather evenly for BR whereas there is a clear systematic change in $r_K$ with increasing $K$ for RHO. These differences are related to the degree of randomness of the sequences and emphasize the difficulty of predicting structure from autocorrelation analysis. This can reveal distinctive and structurally important features of very nonrandom sequences but not necessarily of the random ones, which constitute the bulk of extant protein sequences. Simulations show, however, that random sequences will usually have a few significant values of $r_K$ as expected from the fact that 5% of them should be outside the 95% limits. For BR with $K_{\text{max}} = 27$ we should have observed one $r_K$ outside and the range and we do ($K = 26$, Fig. 5B).

The results of the Fourier analyses (see Methods) are shown in panels C and D of Figs. 5 and 6. In the C panels, the first 50 normalized Fourier intensities $I_m'$ (eqs. 8 and 9) are plotted against order number ($m$). The dotted horizontal lines indicate the expected mean value of 1 for the $I_m'$ and the solid horizontal lines indicate the 2σ significance levels. Note that there are several significant (but not strongly so) intensities for BR (Fig. 5C) as expected from the fact that the $I_m$'s are exponentially distributed. This behavior is much more apparent for RHO in Fig. 6C where the intensities are seen to be much stronger than for BR as expected from the hydrophobicity analyses in the A panels of Figs. 5 and 6.

The probability (number) densities of the Fourier intensities are plotted as solid points in the D panels of Figs. 5 and 6 along with the best-fit exponential distributions (eqs. 10 and 11). Both curves are accurately exponential with $r = \chi^2$ close to 1 but have β parameters (eq. 11) that are quite different. For BR $\beta = 0.953$, which is close to the value expected for a random sequence whereas for RHO the value is 0.865. This difference comes about because of the stronger intensities associated with the nonrandom RHO sequence.

Taken together, the results of Figs. 5 and 6 reveal very similar behaviors for both random and nonrandom sequences; the differences are only quantitative. The main point is that periodic behavior is a characteristic feature of random sequences and should not be interpreted as indicative of nonrandomness. Of course, these examples make clear that the periodicities are much stronger for nonrandom sequences than for random ones. However, one must keep in mind that more than 90% of the known protein sequences qualify as random, so the strengths of the periodicities will tend to be closer to those of BR rather than RHO.

The formation of secondary and tertiary structure from random sequences is easier to understand given the periodic patterns of amino acids inherent to random sequences. These patterns give rise to three-dimensional structures which seem to be very nonrandom. Of course, our judgment of three-dimensional structure is based upon the assumption that the α-helix and β-sheet motifs which characterize so many proteins are examples of nonrandom behavior. The results of Chan and Dill (1990), however, make it clear that most compact structures of proteins involve sheet and helix structures because of the steric constraints inherent to polypeptide chains. Thus, sequences which are random in one dimension must form helical and sheet structures simply because of steric constraints, implying that secondary structure itself is not necessarily a manifestation of nonrandom behavior.

**Implications for Protein Evolution**

Because the PIR database appears to be statistically stable based upon its growth from 5,287 qualifying sequences in 1988 to 14,620 in 1990, we suggest that the PIR database is a reasonable sample of the extant proteins. Given that assumption, we can assert that 90% of the extant proteins are likely to have lengthwise distributions which cannot be distinguished from the random expectation. This result is entirely consistent with the random-origin hypothesis (White and Jacobs 1990; Lau and Dill 1990; Shakhnovich and Gutin 1990a) and supports the idea that the ability to fold into compact structures in an unexceptional characteristic of heteropolymers as suggested by the work of several laboratories (Lau and Dill 1990; Shakhnovich and Gutin 1990b). Although the modern proteins do not have precisely $N(0,1)$ densities as expected for a purely random set of sequences, the densities for the classes of amino acids are, in our opinion, sufficiently close to $N(0,1)$ that the hypothesis that primordial proteins arose from random polypeptide sequences cannot be rejected. Densities which are close to $N(0,1)$ would be expected if primitive genomic nucleic acid and transcription-translation chemistry randomly produced the first proteins during the earliest era of the biosphere. If this scenario is correct, then one can liken the observed $n(r_0)$ densities of the modern proteins to the 2.7K cosmic background radiation of the present universe in that they are both the remnants of primordial processes.

Although the analysis presented is consistent with the random-origin hypothesis, it does not by itself entirely rule out the starter-set hypothesis. Because neutral amino acid mutations are known to occur randomly (Fitch and Margoliash 1967; Jukes 1969), the sequences evolving from the starter-set
Fig. 7. The basic statistical features expected for binary sequences that have a small number 1's. A The dependence of the maximum obtainable value of $r_n$ as a function of length and fraction $f$ of 1's. (See Appendix.) The values of $f$ for each curve are indicated. B A comparison of the actual and simulated $n(r_n)$ densities for Phe. The simulation was performed as described in the Appendix.

sequences would be expected to become randomized even if the initial proteins were nonrandom (Eck and Dayhoff 1966; McLachlan 1972). But why are the observed distributions Gaussian-shaped curves centered close to $r_n = 0$? If sequences drifted into the $-2 < r_n < 2$ region through randomization, there is no reason to believe that the modern distributions should be Gaussian-like. The simplest explanation for the shapes of the curves is that the primordial proteins were a random sample of sequence space.

The starter-set hypothesis is weakened relative to the random-origin hypothesis by the need to invoke a specific series of events, i.e., gene duplication and recombination (Doolittle 1979; Holland and Blake 1990) accompanied by random mutations, to account for the evolution of new protein species. This makes a rapid early expansion of the starter-set seem unlikely. With the random-origin scenario, a great variety of primitive proteins with diverse functions could have arisen nearly simultaneously on an evolutionary time scale. Indeed, the rich diversity of protein sequences which is possible in the random-origin hypothesis is perhaps its most compelling feature. About $10^{112}$ unique sequences can be formed from 100 amino acids chosen to reflect the amino acid composition typical of soluble proteins. Because this number is so large (the age of the universe is about $10^{17}$ s, Peebles et al. 1991), it would be more efficient to sample sequence space randomly by a shotgun approach than by beginning with a starter set of short protein sequences from limited regions of the space.

A great evolutionary advantage would have been gained in a feeble biosphere if the distribution of the lengths of the initial set of proteins was broad because this would assure a large number of initial folding motifs. Thus, a crucial question for the random-origin hypothesis concerns the lengths of the primitive protein sequences. Specifically, can the distribution of the lengths of the modern proteins be explained by means of a stochastic process which does not invoke specific mechanisms such as gene duplication or exon splicing? To answer this question, the distributions of the lengths of the sequences in the superfamily and other sets of proteins have been examined and found to conform to simple statistical rules. It can be demonstrated that the observed length distributions can be explained quantitatively by a simple model that assumes (1) a uniform random distribution of "begin" and "end" signals along the primitive genome and (2) a linear dependence of protein stability on sequence length (S.H. White, submitted).

Acknowledgments. We are grateful for the advice and criticisms of the "Wednesday Afternoon Group" consisting of Prof. Howard Tucker, Prof. Mark Finkelstein, Mr. Norbert Schumacher, and Mr. Les Vernon of the Mathematics Department. We thank the referees for their helpful comments and criticisms. The research was supported by the National Science Foundation (DMB-8807431).

Appendix: The Behavior of the Run Test Statistic for Binary Sequences When the Number of Ones Is Small Compared to the Number of Zeros

As in Methods, we consider a sequence of length $L$ comprised of $a$ 1's and $b$ 0's and show for $a \ll b$ that (1) there is a strong upper limit on the run test statistic $r_n$ and (2) the $n(r_n)$ density has a bimodal shape. Complete treatments of the statistics of runs in binary sequences have been given by Wani (1971) and, particularly, by David and Barton (1962). In brief, the number of runs is described by the random variable $\mu$, which has the range 2, 3, $\ldots, q$ where $q = 2a + 1$ if $a < b$. The probability function $P(\mu)$ for an even number of runs ($\mu = 2k$) differs from that for an odd number of runs ($\mu = 2k + 1$). For an even number

$$P_\mu(\mu) = \frac{(2 \cdot a^{a-1} \cdot b^{b-1} \cdot C_{a-1})^k \cdot C_a}{(a+b)^{a+b}}$$

(A1)
where \( C_n \) is the binomial coefficient \( \binom{x}{y} = \frac{x!}{(x-y)!} \). For an odd number of runs (David and Barton 1962), the probability function is

\[
P_f(n) = P_f(n - 1) \cdot (a + b - 2k)/2k \tag{A2}
\]

so \( \Sigma P_f(n) = \Sigma P_f(n) + \Sigma P_f(n) = 1 \). A few trial calculations for \( a = b \) quickly show that \( 2P_f = 2P_f \). This can be understood intuitively by recognizing that an even number of runs can occur only if the sequence begins or ends (but not both) with \( 1 \), which greatly lowers the probability of an even number of runs. The expected number of runs \( \mu \), and its variance \( \sigma^2 \) given by eqns. (1) and (2) in Methods may be derived using (A1) and (A2).

The probability of obtaining clusters of 1's by chance is small if the 1's are highly diluted by 0's. This follows from eq. (1) of Methods where it can be seen that for \( a = b, \mu_{n,2} = 2a + 1 \), which is also the maximum number of runs obtainable. The maximum value of \( \mu_{n,2} \) can therefore be calculated from eq. (3) by using eqs. (1a) and (2a) of Methods it is easy to show that

\[
r_{\text{max}} = 2fL/\sigma \tag{A3}
\]

\( r_{\text{max}} \) is plotted as a function of \( L \) with \( f/\sigma \) as a parameter in Fig. 7A.

For the typical amino acid frequency of 5%, \( r_{\text{max}} \) increases very slowly with increasing length so that for run the number of runs would be a sharp cut-off in the region of \( < r < 2 \). This behavior is apparent for the densities plotted in Fig. 4. We estimated \( r_{\text{max}} \) for each amino acid using the average frequency and length of the population. These values are listed in Table 3 and one should note that in all cases \( r_{\text{max}} > r_{\text{max}} \).

Determination of the exact shape of the \( n(r) \) densities for each amino acid by analytical methods is difficult because of the wide range of frequencies and lengths in the populations of sequences. However, one can establish an approximate shape for the independence random model by simulation. We did this for the 843 sequences used for examining the runs of Phe which have a mean frequency of 4.60% and average \( r_{\text{max}} = 0.00 \) (Table 3) in the following way. We first sorted the sequences by length into bins of width 100 AA and determined for each bin the number of sequences and their average f. Using eqs. (A1) and (A2), we then calculated the expected numbers of sequences having particular numbers of runs. A vector of 843 simulated values of \( r_{\text{max}} \) could then be produced for statistical comparison to the actual vector of 843 values. Because the simulated data lacked the richness of variety arising from the wider range of frequencies and length of the actual data, we sorted the \( r_{\text{max}} \) values into bins of width 0.50 rather than the 0.08 used for Fig. 4 and Table 3. The \( n(r) \) densities for the actual data and the simulation are compared in Fig. 7B. While the simulation is not perfect, it does show the basic bimodal behavior observed for the actual data.

The simple mean and standard deviation for the simulated data are \( -0.02 \) and 1.02 respectively compared to \( 0.00 \) and 0.99 for the actual data. We conclude for Phe that the distribution of \( r_{\text{max}} \) values cannot be distinguished from the random expectation allowing for differences in the details of the densities of frequencies and lengths. However, we list in Table 3 the simple mean values, \( r_{\text{max}} \) of \( r_{\text{max}} \) for each amino acid, and it should be noted that only 11 of the amino acids have means sufficiently close to 0 to be considered consistent with the random expectation. Thus, while we expect the \( n(r) \) densities to be qualitatively similar to one another and to the simulation of Fig. 7B, there will clearly be quantitative differences due to nonrandomness and to differences in the densities of the frequencies and lengths of each population.

References


Fisher HF (1964) A limiting law relating the size and shape of protein molecules to their composition. Proc Natl Acad Sci USA 51:1285–1291


Klapper MH (1977b) The independent distribution of amino acid near neighbor pairs and in polypeptides. Biochem Biophys Res Comm 78:1018–1024

Received November 4, 1991/Revised June 27, 1992