brief communication

Statistical distribution of hydrophobic residues along the length of protein chains

Implications for protein folding and evolution

Stephen H. White and Russell E. Jacobs
Department of Physiology and Biophysics, University of California, Irvine, California 92717

ABSTRACT We consider in this paper the statistical distribution of hydrophobic residues along the length of protein chains. For this purpose we use a binary hydrophobicity scale which assigns hydrophobic residues a value of one and non-hydrophobes a value of zero. The resulting binary sequences are tested for randomness using the standard run test. For the majority of the 5,247 proteins examined, the distribution of hydrophobic residues along a sequence cannot be distinguished from that expected for a random distribution. This suggests that (*a*) functional proteins may have originated from random sequences, (*b*) the folding of proteins into compact structures may be much more permissive with less sequence specificity than previously thought, and (*c*) the clusters of hydrophobic residues along chains which are revealed by hydrophobicity plots are a natural consequence of a random distribution and can be conveniently described by binomial statistics.

INTRODUCTION

Similarities among families of protein sequences are seen at three levels of complexity: identical residues at identical positions, chemically similar residues at identical positions, and, at the most complex level, the conservation of physicochemical properties of entire proteins or protein segments (Nolan and Margoliash, 1968). At this highest level, the conservation of regions of hydrophobicity is particularly important because of their aggregation to form the interiors of globular proteins, the interfaces between subunits, or the lipid-spanning segments of membrane proteins (Kauzmann, 1959; Chothia, 1984; Engelman et al., 1986). The amino acid sequence of any protein reveals that hydrophobic residues tend to occur in clusters along the length of the chain. The resulting hydrophobic regions of amino acid sequences can be located by means of hydropathy plots (Rose and Roy, 1980; Kyte and Doolittle, 1982; Engelman et al., 1986) which are sliding-window averages or sums of amino acid hydrophobicity parameters (see Fig. 1 *A*). These plots, regardless of window size or hydrophobicity scale, are invariably "noisy" and suggest random fluctuations along the chain. Investigation of this noise led us to examine randomized membrane protein sequences which yield hydrophobicity plots such as those of Fig. 1 *B* for the *L* subunit of the photosynthetic reaction center of *Rb. sphaeroides*. Despite the detailed differences between the plots of the native and random sequences, they are remarkably similar in exhibiting peaks which could be interpreted as potential regions of transbilayer helices; the naive observer would certainly attribute the random plots to membrane proteins. These results suggested that hydrophobic residues might tend to be randomly distributed along the protein chains. We show in this paper that the distribution of the hydrophobic residues along the chain cannot be distinguished from that expected for a random distribution for the vast majority of the membrane and soluble proteins examined. The clustering observed is a consequence of the random distribution and can be described by simple binomial statistics. This suggests that functional proteins could have originated from random sequences. It also suggests that the folding of proteins into compact structures may be much more permissive with less sequence specificity than previously thought, consistent with recent theoretical studies of protein folding (Skolnick et al., 1989; Kolinski and Skolnick, 1989; Lau and Dill, 1990; Chan, H.S., and K.A. Dill, manuscript submitted for publication).

TESTS OF RANDOMNESS

We approached the question of the possible random distribution of hydrophobic residues by first adopting a *binary* hydrophobicity scale which assigns a value of 1 to hydrophobic residues and 0 to the others so that a protein sequence can be represented as a binary sequence. There is considerable debate (Engelman et al., 1986) as to the proper hydrophobicity scale for amino acids, but Phe, Met, Leu, Ile, Val, Cys, Ala, Pro, Gly, Trp, and Tyr are frequently considered as hydrophobic and to these we assign the value of 1. Different assignments can be made but these have no significant effect on our observations.

0006-3495/90/04/911/11 $2.00

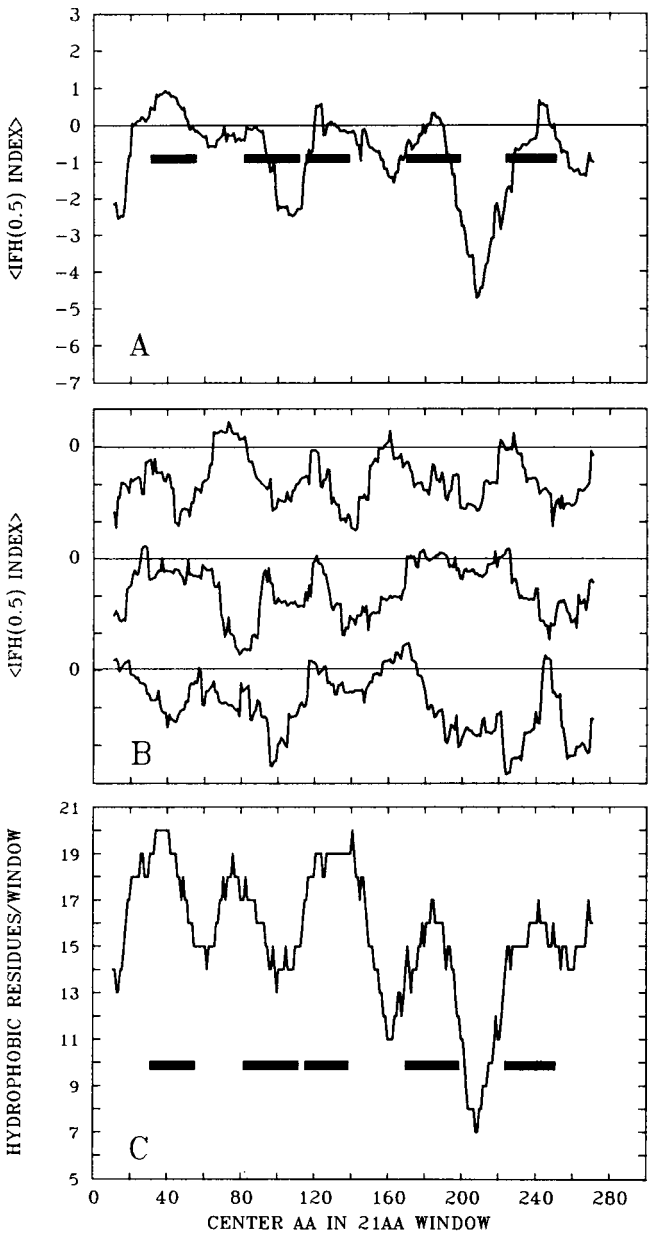

FIGURE 1 Hydropathy plots of the *L* subunit of the photosynthetic reaction center of *Rb. sphaeroides*. The plots in *A* and *B* are sliding window averages; *C* is a sliding window sum. (*A*) The hydrophobicity profile of the native sequence using the interfacial hydrophobicity index [IFH(h)] of Jacobs and White (1989) with a hydrogen bond parameter $h = 0.5$. This scale uses the bilayer interface as the reference "phase;" a single point above the zero line indicates a window of 21 residues with an average index value favoring insertion into a bilayer. (*B*) Hydrophobicity profiles of random *L* subunit sequences obtained by three consecutive randomizations of the native sequence. The sequences were randomized using a microcomputer spread-sheet program; a random number between 1 and 10^7 was assigned to each residue of the native sequence which were then rearranged in ascending or descending order of the random numbers. (*C*) Hydrophobicity plot of the *L* subunit using the *binary* hydrophobicity scale described in the text. The sliding window sum tells how many hydrophobic residues there are in each of the $L - W + 1$ windows (L = length of sequence; W = window width). This plot clearly reveals the five major hydrophobic domains associated with the five known helices of the subunit (Allen et al., 1987).

The sliding-window sum of binary hydrophobicity for the *L* subunit (Fig. 1 *C*) clearly reveals the hydrophobic peaks associated with the five observed transbilayer helices (Allen et al., 1987).

A random binary sequence (Bernoulli sequence) does not have a uniform distribution of ones and zeros along the length of the sequence. Rather, one finds clumps of ones (hydrophobes) of varying size separated from one another by one or more zeros (hydrophiles). The distribution of clump sizes is described by $c(n) = L \cdot f^n \cdot (1 - f)^2$, where $c(n)$ is the number of clumps containing n ones [$n \geq 1$], L the number of residues, $f = N/L$ the fraction of the residues which are hydrophobic, and N the number of "ones" in the sequence (Roach, 1968). The total number of clumps is $C = L \cdot f(1 - f)$. The observed and theoretical clump distributions for bacteriorhodopsin (BR) and bovine rhodopsin are shown graphically in Fig. 2. The hydrophobic clumps generally follow the theoretical distribution with large clumps being less likely than small ones. One can easily devise a chi-square test to compare the theoretical and observed clump distributions (Table 1). However, this is not a strong test of randomness because the clumps themselves might not be randomly distributed (all of the clumps might be near one end of the chain, for example). A random sequence should have properly intermixed clumps (runs) of ones and zeros of different sizes. The so-called *run* test (see e.g. Wani [1971]) considers simultaneously the sizes and distributions of runs of *both* the ones and zeros. If a binary sequence is random, the expected number of runs is $\mu = 2f(L - N) + 1$ with a standard deviation $\sigma = \sqrt{(\{2f(L - N)\ [2f(L - N) - 1\}/[L - 1])}$. If μ_0 is the observed number of runs and N and $L - N$ are each >10, the statistic $r_0 = (\mu - \mu_0)/\sigma$ will be normally distributed so that the probability $P(r > r_0)$ of a random sequence having r greater than the observed r_0 can be calculated.

INTERPRETATION OF RUN TEST RESULTS

Before using the r_0 statistic, it is useful to consider its meaning and interpretation when applied to a collection of protein sequences whose members have a wide range of r_0 values. By determining r_0 for each sequence in the collection or in a fairly drawn sample of the collection, a density distribution $n(r_0)$ can be constructed which describes the number of chains having a particular r_0 as a function of r_0. Consider three types of collections. The

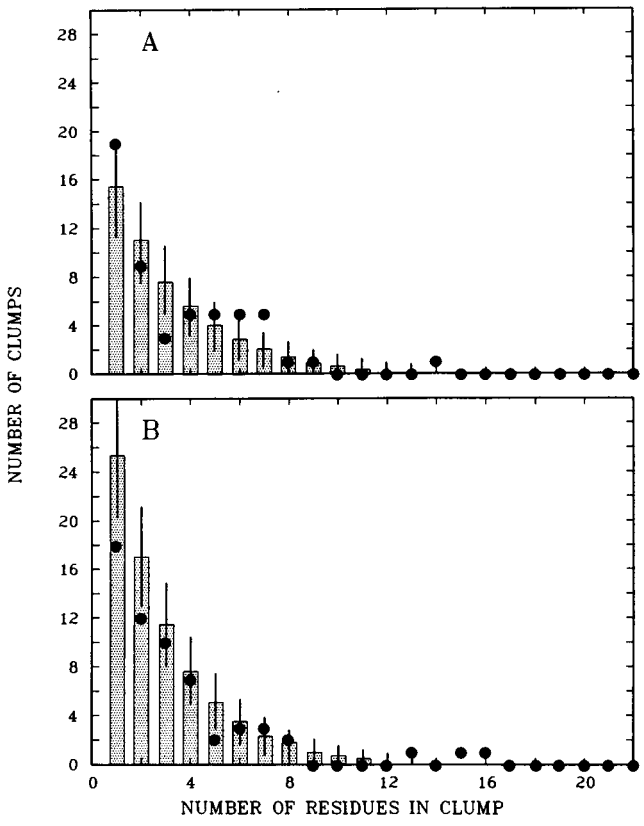

FIGURE 2 Plots of the distribution of clumps of the hydrophobic residues of bacteriorhodopsin (BR) (*A*) and bovine rhodopsin (*B*). The clump analysis (Table 1) indicates that the distribution of hydrophobic clump sizes cannot be distinguished statistically from that of a random chain. Stippled bars show the distribution expected for a random chain and the solid lines the expected standard deviations. The observed distributions of clumps for the native sequences are shown as solid points.

first and simplest is a large family of sequences *produced randomly* by a single process with each member having the same amino acid composition and length. Let each member of the family have length L_j and fraction f_j of hydrophobes. We shall call such a family a single-j collection. The $n(r_0)$ density distribution of the collection or a fairly drawn sample of it will be a Gaussian distribution with mean $\langle r_j \rangle = 0$ and standard deviation $\sigma_j = 1$ (the normal or so-called $N(0,1)$ distribution). The second and more complex type of collection is a multi-j collection consisting of several families of differing L_j and f_j which, again, have been produced randomly by a single process. The $n(r_0)$ density distribution of each family in the collection will be $N(0,1)$ and therefore the whole collection will be $N(0,1)$. The third and most complex type of collection is one consisting of members of differing L_j and f_j produced by one or more processes either randomly or nonrandomly. We shall call this type of collection a *heteronomous* collection. The $n(r_0)$ distribution in this case obviously cannot be easily defined.

Now consider a collection of sequences encompassing a range of values of L_j and f_j. To make a decision about the randomness of a *single* sequence (in the absence of any other information), one adopts the null hypothesis that the sequence is random. The hypothesis is rejected if $P(r > r_0) \leq \alpha$ where α is a risk factor. One can arbitrarily set α but it is frequently taken as 0.05 (Wani, 1971). Now suppose that a large number of sequences, comprising a sample, is drawn from the collection so that $n(r_0)$ for the sample can be determined. In the absence of any additional information, there is no *a priori* reason to expect $n(r_0)$ of the sample to adhere to any particular mathematical form. The distribution could, for example, consist of a collection of apparently uncorrelated points; the random sequences are those for which $P(r > r_0) > \alpha$. If $n(r_0)$ is the

TABLE 1 Summary of the statistical analysis of the distributions of hydrophobic residues of selected proteins

Protein	L	f	Runs			Clumps			
			μ_o	μ	$P(r > r_o)$	C_o	C	$P(\chi^2 > \chi^2_o)$	n_h
PSRC: M subunit									
R. viridis	323	0.71	144	133.4	0.15	72	66.2	0.17	9
R. sphaeroides	307	0.70	130	129.0	0.90	65	64.0	0.41	8
PSRC: L subunit									
R. viridis	273	0.72	116	111.6	0.51	58	55.3	0.74	8
R. sphaeroides	281	0.73	115	111.0	0.54	58	55.0	0.69	9
PSRC: H subunit									
R. viridis	258	0.60	127	124.8	0.77	64	61.9	0.67	2
R. sphaeroides	260	0.63	119	122.6	0.63	60	60.8	0.73	4
Halobacteria opsins									
HR	274	0.71	104	113.4	0.16	52	56.2	0.73	8
BR	262	0.71	108	108.1	0.99	54	53.5	0.25	7

(continued)

TABLE 1 (continued)

Protein	L	f	Runs			Clumps			n_h
			μ_o	μ	$P(r > r_o)$	C_o	C	$P(\chi^2 > \chi^2_o)$	
β-Adrenergic receptors									
Hamster	418	0.58	177	204.8	<0.01*	89	101.9	<0.01*	3
Human	413	0.58	175	202.1	<0.01*	88	100.5	0.01*	3
α-Subunit nicotinic acetylcholine receptor									
Calf	437	0.58	208	214.4	0.62	104	106.7	0.30	3
Human	437	0.57	212	215.0	0.77	106	107.0	0.46	3
Other membrane proteins									
Sub. K receptor	384	0.67	155	171.0	0.06*	78	85.0	0.79	8
Rhodopsin (bovine)	348	0.67	121	155.0	≪0.01*	61	77.0	0.24	7
α-Helical proteins									
Horse Hb (β)	146	0.58	80	72.0	0.17	40	35.5	0.02*	1
Uteroglobin	91	0.56	45	45.8	0.86	23	22.4	0.94	0
Myohemerythrin	118	0.51	63	60.0	0.58	32	29.5	0.97	0
Myoglobin	153	0.52	85	77.3	0.21	43	38.2	0.25	0
Parvalbumin	108	0.54	53	54.7	0.74	27	26.8	0.97	0
β-Sheet proteins									
Super oxide dis.	151	0.54	80	76.1	0.52	40	37.6	0.27	1
Plastocyanin	99	0.59	54	49.0	0.30	27	24.0	0.27	1
Rubredoxin	174	0.55	96	87.1	0.17	48	43.0	0.37	1
Trypsin inhib.	181	0.54	88	91.0	0.65	44	45.0	0.24	1
Concanavalin A	237	0.50	134	119.5	0.06*	67	59.2	0.49	0
Subtilisin inhib.	113	0.65	60	52.7	0.13	30	25.8	0.94	2
α-Helix/β-sheet proteins									
Flavodoxin	138	0.54	71	69.5	0.79	36	34.2	0.99	1
Carboxy pep.	307	0.52	158	154.1	0.66	79	76.6	0.51	1
Adenylate kin.	194	0.50	104	98.0	0.39	52	48.5	0.27	0
Subtilisin	274	0.60	136	132.3	0.64	68	65.6	0.02*	3
Triose phos. ism.	248	0.55	126	123.6	0.76	63	61.3	0.51	1
Cysteine-rich proteins									
Phospholipase	120	0.58	67	59.6	0.17	33	29.3	0.01*	1
Crambin	46	0.67	23	21.2	0.54	11	10.1	0.52	1
Insulin	52	0.60	24	25.6	0.63	12	12.5	0.55	0
Wheat germ aglt.	171	0.64	70	80.0	0.10	35	39.5	0.27	3
Small metal-rich proteins									
Ferredoxin	98	0.51	49	50.0	0.84	25	24.5	0.91	0
High potential Fe	85	0.60	44	41.8	0.62	22	20.4	0.22	1
Cytochromes *c*									
Rb. sphaer. (*L*)	124	0.54	70	62.3	0.18	35	30.8	0.96	0
Rps. viridis. (*M*)	107	0.52	51	54.4	0.51	25	26.9	0.70	0
Rps. tenue (*S*)	92	0.62	40	44.4	0.33	20	21.7	0.06	1
D. vulgaris (*S**)	82	0.60	47	40.4	0.13	24	19.7	0.96	1
Bullfrog	104	0.53	48	52.3	0.34	24	25.9	0.82	0
Kangaroo	104	0.52	50	52.9	0.56	25	26.0	0.87	0
Human	104	0.52	48	52.9	0.33	24	26.0	0.53	0

Protein sequences obtained from the National Biomedical Research Foundation's Protein Identification Resource Sequence Database (Washington, DC). The letters in parentheses next to the bacterial cytochromes *c* indicate the size classes of Dickerson (1980). n_h is the estimated upper limit of the number of possible transbilayer helices (see text). In the clump analysis, C_o is the observed total number of clumps and C the number theoretically expected. The distribution $c_o(n)$ of clumps of size n observed were compared with the theoretically expected distribution $c(n)$ by means of the $\chi^2 = \Sigma\{[c_o(n) - c(n)]^2/c(n)\}$ statistic with 11 class intervals ($n = 1..10, n \geq 11$) and 10 degrees of freedom. $P(\chi^2 > \chi^2_o)$ is the probability that a random sequence will have a larger χ^2 than that observed. In the runs analysis, μ_o is the observed number of runs and μ the expected number. $P(r > r_o)$ is the probability that a value of r greater than $r_o = (\mu - \mu_o)/\sigma$ will be obtained with a random sequence. σ is the theoretical standard deviation (see text). Following the usual practice, we assume that values of $P(\chi^2 > \chi^2_o) \leq 0.05$ or $P(r > r_o) \leq 0.05$ are indicative of nonrandom distributions. *Proteins with $P(\chi^2 > \chi^2_o)$ or $P(r > r_o) \leq 0.10$.

$N(0,1)$ distribution and if the sample can be shown to be fair, then one could reasonably assume that the parent constitutes a randomly produced multi-j family. In any case, one would take the sequences with $P(r > r_0) \leq \alpha$ as having nonrandom characteristics. Another possible outcome is a normal distribution of $\sigma_j \neq 1$ centered at $\langle r_j \rangle \neq 0$. Even though this describes a very particular sample (and parent collection if the sample is fair), one would nevertheless conclude that those sequences with $P(r > r_0) > \alpha$ could not be distinguished from random. The interpretation of the distribution function describing $n(r_0)$ is unclear in this case. The fact that it is not $N(0,1)$ could mean that the collection was not fairly sampled, that a variety of processes (random and nonrandom) created the collection, or that a combination of the two prevails.

Finally, consider a collection of sequences described by the $N(0,1)$ distribution further. Suppose the collection contained 10^3 members. About 600 of these would have $P(r > r_0) > 0.32$ ($|\sigma_j| = 1$) and we would conclude that all of these sequences were random. Now, even though all of these particular sequences may be random, there will nevertheless be considerable variations in sequence. Thus, if the sequences all represented functional proteins, it is possible that there would be considerable diversity in three-dimensional structure and in function arising from sequence-dependent interactions despite the fact that all of the sequences would be considered random from the mathematical point of view.

We conclude from this discussion that (*a*) $P(r > r_0)$ is the principal and only practical criterion for judging randomness of an individual sequence. (*b*) The interpretation of the shape of the $n(r_0)$ distribution for a sample drawn from a collection is problematic. If it is $N(0,1)$, the sample has the statistical characteristics of a randomly produced collection. (*c*) The $n(r_0)$ of the sample can be assumed to represent the distribution of the parent collection only if the sample is fairly drawn. If it is fairly drawn and $n(r_0)$ is not $N(0,1)$, then it is reasonable to conclude that the collection itself is heteronomous. (*d*) There can be considerable sequence diversity among chains which are equivalent in terms of randomness.

RUN TEST RESULTS

Examples of the application of the run test to a wide variety of membrane and soluble proteins are shown in Table 1. The soluble proteins were chosen to cover the spectrum of structural classes described by Richardson (1981). Taking $P(r > r_0) \leq 0.05$ as the criterion for nonrandomness (Wani, 1971), the data of Table 1 show that among membrane proteins BR is clearly random ($P = 0.99$), whereas bovine rhodopsin is clearly not random ($P < 0.0001$). Other membrane proteins which give strong evidence of nonrandomness are certain membrane receptors (β-adrenergic and substance K receptors but not the nicotinic acetylcholine (ACh) receptor α-subunit). As for the soluble proteins, none have values of $P(r > r_0) \leq 0.05$; only concanavalin A has a sufficiently small $P(r > r_0)$ (=0.06) that it might be nonrandom. That is, most of the soluble proteins appear to have their hydrophobic residues distributed along the chain in a way that cannot be distinguished from random. Taken together, the results indicate that the random character of the distribution of hydrophobes in soluble proteins tends to be independent of structural classes defined by Richardson (1981).

The list of proteins in Table 1 is too small to be certain that the random nature of the hydrophobe distribution is a general feature of proteins. Further, the soluble proteins were selected from a special group of proteins comprised of those of known three-dimensional structure. We therefore performed the run test on all of the qualifying proteins of the entire set of protein sequences contained in the 1988 National Biomedical Research Foundation's Protein Identification Resource (PIR) data base (Orcutt et al., 1983; George et al., 1986). The "qualifying proteins" were those with N and $L - N$ greater than 10 whose complete sequences were known unambiguously. Of the more than 7,200 proteins and protein fragments in the data base, 5,247 qualified for the run test. These ranged in length from 22 to 2,700AA (mean = 290) with $f = N/L$ ranging from 0.15 to 0.87 (mean = 0.54). We found that 59.5% of the sequences have $|r_0| < 1.0$ ($P[r > 1] = 0.32$) and 88.8% have $|r_0| < 2.0$ ($P[r > 2] = 0.05$) (see Fig. 3). One can thus state with reasonable confidence that at least 60–80% of the known sequences have hydrophobic residue distributions with random characteristics. We could find no apparent correlation between r_0 and f or L.

The $n(r_0)$ distribution of the qualifying proteins in the data base was established by sorting the sequences by r_0 values into bins of width 0.08 except for $|r_0| > 3.0$ which are lumped into single bins (*circled points*, Fig. 3). The resulting density plot (*solid squares*) and the $N(0,1)$ distribution expected for a random multi-j family (*solid curve*) are shown in Fig 3. Considering the diversity of the data base and its likely heteronomous nature, it is mildly surprising that $n(r_0)$ has the general form of $N(0,1)$; we would have been less surprised by a scatter plot of uncorrelated points. We fitted a Gaussian curve to $n(r_0)$ using χ^2 minimization (*dashed curve*, Fig. 3) and found $\langle r_0 \rangle = -0.21 \pm 0.02$ with $\sigma = 1.67 \pm 0.02$; the reduced-χ^2 is 1.7. Thus, the actual distribution is wider and slightly skewed toward negative values of r_0. The latter observation means that the sequences in the data base tend to have slightly more runs ($\mu_0 > \mu$) than expected for a completely random set. Whether this is a general feature

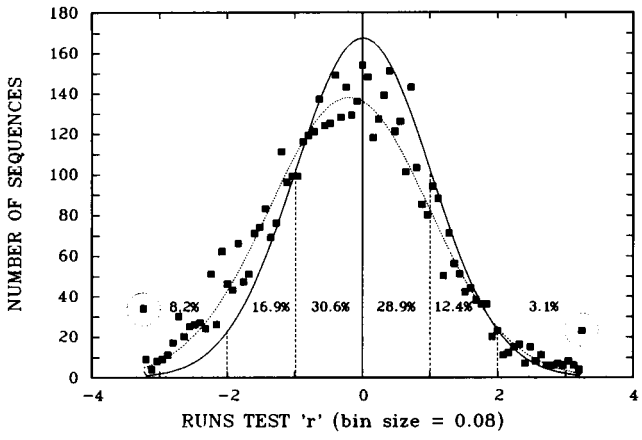

FIGURE 3 The frequency of appearance of proteins with a particular run-test r_0 in a data base of 5,247 protein sequences. The observed frequencies are indicated by the points (*solid squares*). A normal distribution of mean zero and standard deviation one is shown as the solid curve. A Gaussian curve (*dashed curve*) was fit to the data by minimization of χ^2. The density distribution observed is a characteristic of the data base; its relation to the extant proteins is unknown (see text). It is clear, however, that at least 60% of the qualifying proteins (see text) in the data base have distributions of hydrophobic residues along the lengths of the sequences which cannot be distinguished from that expected for a random distribution.

of the extant proteins or only a bias of the sampled collection cannot be known at present. No definite conclusions can be drawn from the shape of $n(r_0)$ because the accuracy of the sampling of the extant proteins by the data base is unknown. If the sample is assumed to be accurate, a reasonable conclusion is that the collection was not produced by a simple random process. However, because the shape of $n(r_0)$ is Gaussian-like and centered close to $r_0 = 0$, one must assume that random processes played a major role in the creation of the extant proteins. In any case, it is absolutely certain that there is an extraordinarily large number of known proteins with random characteristics.

IMPLICATIONS OF A RANDOM HYDROPHOBIC DISTRIBUTION

We conclude that the distribution of hydrophobic residues along the lengths of a large number of protein chains cannot be distinguished from that expected for a random distribution. This suggests that for many protein chains the ability to fold into a compact (though not necessarily functional) structure is a deeply "embedded" property; the statistical distribution of hydrophobic residues may suffice to determine the basic folding characteristics of a chain. The statistical nature of the distribution may explain why the folded conformations of proteins are much more highly conserved than are the primary sequences (Creighton, 1984) and perhaps why signal sequences have such a small information content (von Heijne, 1988). One should *not* conclude that specific interactions are unimportant in determining the final three-dimensional structure of the protein. However, in many cases these interactions may only "fine-tune" the structure (ignoring major interactions such as disulfide bonds). Skolnick and his colleagues have reached a similar conclusion based upon Monte Carlo studies of the folding of "generalized" four-bundle α-helical and six-strand β-barrel proteins (Skolnick et al., 1989; Kolinski and Skolnick, 1989). Experimentally, random copolymers of amino acids have been observed to form compact structures (Rao et al., 1974).

The idea of random protein chains at first seem contrary to the notion that evolutionary forces have highly selected functional proteins. However, it has been known for more than 20 years that neutral substitutions in evolutionarily related proteins are scattered randomly along the chain in accordance with Poisson statistics except for perhaps 5–20% of the residues which are invariant and strongly conserved (Fitch and Margoliash, 1967*a*; Jukes, 1969). Because so few of the residues of many proteins fall into the latter class, it is reasonable that a protein chain has random characteristics. An ancestral protein could have been initially nonrandom and subsequently randomized by neutral mutations. However, there is no reason to suppose *a priori* that an ancestral protein sequence could not have been selected from an ensemble of sequences with random characteristics. Indeed, considering that there are 10^{112} possible sequences for a 100-residue protein with a typical soluble protein amino acid composition, a random ancestral protein seems more likely. The fact that chains which presently have random characteristics can fold into functional proteins is consistent with this point of view. Fitch and Margoliash (1967*b*) have estimated the ancestral amino acid sequence of cytochrome *c* from the phylogenetic tree. The run test on this sequence gives $P(r > r_0) = 0.25$ which is quite comparable with the cytochromes *c* of Table 1. These ideas suggest a biological "minimalist" approach for the genetic preservation and improvement of a protein sequence: only very selected regions of a sequence, such as a haem binding site, for example, need be accurately maintained. As we discuss next, if the average hydrophobicity and chain length of the protein are preserved, the fundamental folding characteristics are likely to be preserved as well. Dill and his colleagues have recently put forward similar ideas, based upon thermodynamic analyses of protein folding (Lau and Dill, 1990; Chan, H.S., and K.A. Dill, submitted for publication). They suggest that there is an extraordinarily large class of protein sequences which can fold into compact structures

and that functional proteins may have originated from random sequences. Our findings support this point of view as do those of Zielenkiewicz et al. (1988) who found that repetitions of tetrapeptides along chains are random.

HOW STATISTICS CAN DETERMINE FUNDAMENTAL FOLDING PROPERTIES

Two very fundamental observations describe gross protein morphology: hydrophobic residues form the core of soluble proteins (Kauzmann, 1959) and chain turns occur at their surfaces at local minima in lengthwise hydrophobicity (Rose, 1978). The amino acids most likely to occur at highly exposed surface positions are Gly, Pro, Asn, Asp, and Ser (Hopp, 1985) which are the same residues most likely to participate in reverse turns *and* to disrupt secondary structure (Levitt, 1977). Thus, a general characteristic of folded soluble proteins is the tendency for the core to be formed by relatively hydrophobic segments of α-helix and/or β-strands as elegantly demonstrated by Rose and Roy (1980). White and Jacobs (1990) have shown that a similar situation exists for the membrane proteins of known structure where the predominant secondary structure is α-helix and the protein 'core" is buried in the bilayer interior. Ptitsyn and Finkelstein (1980) have noted that whereas the type of secondary structure may be determined by the α-helix or β-sheet propensities of the constituent residues, the *dimensions* of the secondary structural elements will be determined by the continuous hydrophobic surfaces which can be formed. Interestingly, these authors have considered in a limited way the folding of random copolymers and how statistical considerations lead to reasonable estimates of the average size of secondary structure elements in soluble proteins (Finkelstein and Ptitsyn, 1987).

It is apparent, then, that the regions of sequences rich in hydrophobes will determine the basic topology and dimensions of both membrane and soluble protein. These regions are conveniently revealed by hydrophobicity plots and we therefore consider sliding-window sums of binary hydrophobicity from a statistical point of view. This will make it possible to show how chain length, average hydrophobicity, and the statistical distribution of hydrophobes can determine basic folding properties. We specifically consider membrane proteins but the basic arguments are easily extended to buried secondary structure of soluble proteins. Consider a contiguous sequence of W residues in a random sequence where W is the width of the window used in a sliding window sum. The probability of finding n hydrophobic residues in the window is simply the binomial probability $P(n,W) = \{W!/[(W-n)!n!]\} \cdot f^n(1-f)^{W-n}$. A sliding window sum along a chain which results in $N_w = L - W + 1$ windows should produce a distribution of window occupancies given by $P(n,W) \cdot N_w$. The results for the sliding window sums for BR and bovine rhodopsin are shown in Fig. 4. It is clear that the underlying distribution of the windows is binomial as expected. One can calculate the expected standard devia-

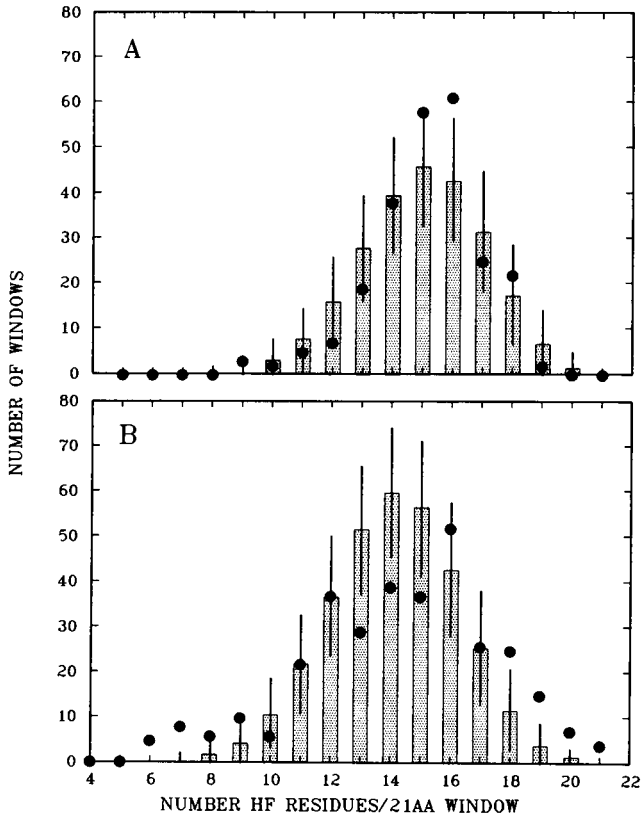

FIGURE 4 Plots of the frequency of appearance of 21AA windows containing different numbers of hydrophobic residues in sliding window sums for bacteriorhodopsin and bovine rhodopsin. Solid circles show the observed distributions $N_0(n)$ and stippled bars the theoretical binomial distributions $N_t(n) = [L - W + 1] \cdot P(n,W)$. The sliding windows used for measuring the actual window occupancies overlap and are therefore not independent. However, one can show that $[N_0(n) - N_t(n)]/\sigma_n$ is normally distributed where σ_n is the square root of the variance which depends upon the degree of window overlap. One can calculate the expected standard deviations (heavy vertical lines) under these circumstances (Ibragimov and Linnik, 1971). (*A*) Bacteriorhodopsin. The observed distribution fits the expected binomial distribution very well because the distribution of hydrophobic residues for BR is random as determined by the run test (Table 1). (*B*) Bovine rhodopsin. The observed distribution deviates significantly and systematically from the binomial distribution because it is nonrandom. The actual distribution is flattened and broadened because there are more hydrophobes in certain regions than expected by chance. This leads to a decreased number of runs (121 rather than the 155 expected [Table 1]). If the hydrophobes were uniformly distributed along the chain, all windows would contain the mean number of hydrophobes ($\mu_b = f \cdot W$) and the number of runs would be increased. Thus, chains with the hydrophobes distributed more uniformly than chance would predict, will tend to have a narrower distribution.

tions for $P(n,W) \cdot N_w$ for the overlapping (nonindependent) windows and these are shown as heavy vertical lines in the figures (Ibragimov and Linnik, 1971). BR fits the random binomial distribution well. Bovine rhodopsin, however, shows a significantly increased number of windows with large (and small) numbers of hydrophobic residues as expected because it is nonrandom (Table 1). (This suggests caution in using the BR structure as a model for bovine rhodopsin.)

To see how this binomial-based distribution can determine the fundamental folding properties of a protein, it is interesting to consider the characteristics that distinguish membrane proteins from soluble ones. The principal distinction, of course, is the existence of stretches of hydrophobic sequences of sufficient length to form transmembrane helices. If the hydrophobes are distributed randomly, the binomial character of the sliding window sums leads to an estimate of the number n_h of transbilayer helices. Fig. 1 *C* shows that regions of the *L* subunit chain associated with the five transbilayer helices have a minimum of 15 hydrophobes in a 21AA window with the windows centered at the major peaks containing as many as 20AA. The binomial window occupancy statistics (Fig. 4) mean that the probability of windows with large numbers of hydrophobes is relatively small. But, it is precisely these windows (regions of the chain), originating from the right-hand wing of the distribution, that are important in protein insertion and folding. This observation forms the basis for the calculation of n_h. The total probability of finding 15 or more hydrophobes in a window is $P(15,W) + P(16,W) + \ldots + P(21,W) = \Sigma P(n,W)$. This is likely to be an *upper* limit because most chain regions associated with transbilayer helices have $n > 15$ (Fig. 1 *C*). The number of *independent* windows along the chain will be $n_w = L/W$. None of these windows need necessarily be centered on the longest runs of hydrophobes but this can easily be accomplished by shifting a window by no more than $\pm W/2$. A reasonable first estimate for the number of helices is thus $n_h = n_w \cdot (\Sigma P(n,W))$ $(n = 15 \ldots 21)$. Calculated values of n_h (W = 21) are shown in Table 1. The values of n_h appear to be upper limits because the *L* and *M* subunits are known to have five helices (n_h = 8–9) and the H subunit one (n_h = 2–3) (Allen et al., 1987); BR and, presumably, halorhodopsin (HR), have seven (n_h = 8) (Henderson and Unwin, 1975; Blanck and Oesterhelt, 1987). Further, some of the soluble proteins with no membrane-spanning segments have values of $n_h \geq 1$. It is thus interesting that the apparent upper limits for both the β-adrenergic receptors and the ACh α-subunit are three helices. The β-receptor is widely assumed to have six or seven helices (Kerlavage et al., 1986) and the ACh α-subunit four or five (McCrea et al., 1988) but these numbers are controversial (Kerlavage et al., 1986; McCrea et al., 1988; Lodish, 1988).

The formula for calculating n_h shows the statistical likelihood of transbilayer helices is determined, not surprisingly, by f and L. Bacteriorhodopsin and the *L* and *M* subunits of the photosynthetic reaction center have a fraction $f \approx 0.7$ of hydrophobic residues, chain lengths of ~250–300AA, and five to seven transbilayer helices. The H subunit is in the same size class but with $f \approx 0.6$ should, statistically, have fewer helices and does. The hydrophobicity of this subunit is close to hemoglobin's which has no long helices of sufficient hydrophobicity to be considered as transbilayer helices. The difference is in the chain length; the hemoglobin chain is comparatively short (146AA) and thus has a much smaller chance of having a long enough run of hydrophobes to form a transbilayer helix. The β-adrenergic receptors and nicotinic ACh receptor α-subunits also have $f \approx 0.6$ but are of much greater length than either hemoglobin or the H subunit and therefore have a greatly increased likelihood of transbilayer segments. However, if the hydrophobes are not distributed randomly so that there are longer runs or larger clumps, then the number of helices can possibly be increased. Thus, the nonrandom β-receptors may have more than the theoretical number of three helices, whereas the random ACh α-subunit is likely to have a number closer to the theoretical value. Regarding nonrandomness, it is interesting that *any* nonrandom binary sequence of fixed f must give simultaneously an increased frequency of both over- and underfilled binomial windows (Fig. 4 *B*) because the mean number μ_b of residues per window averaged over all the windows is contstrained to be $\mu_b = f \cdot W$. That is, a fixed f implies a fixed number of hydrophobes which must be conserved.

As a final aspect of the role of statistics in the determination of structural features, consider the number of residues comprising secondary structure elements. The window length $W = 21$ used above for membrane proteins was not chosen entirely arbitrarily nor because an $\approx$ 20AA window is commonly used due to the expectation that this number of residues is required to span the 30-Å-thick hydrocarbon core of the bilayer. Rather, we wished to have the window wide enough to assure the accommodation of chain regions with very long runs of hydrophobes should they occur. The standard deviation of a binomial distribution is $\sigma_b = \sqrt{[\mu_b(1-f)]}$ and we arbitrarily chose $W = \mu_b + 3\sigma_b$. This choice defines the "natural" window as $W = 9f/(1-f)$. For membrane proteins with $f = 0.7$, $W = 21$AA which is a good estimate of the lengths of helices in known membrane proteins which tend to have about 25 residues per helix. If one uses $f = 0.55$, which is typical of soluble proteins (Table 1), then $W = 11$AA. Interestingly, $W = 11$ is a reasonable

estimate of the average secondary structure lengths of soluble proteins: α-helices average about 12 residues (Srinivasan, 1976), whereas β-strands average six to seven residues and generally occur in sheets of two to six strands or 12 to 14 residues per pair of strands (Sternberg and Thornton, 1977). These results clearly reveal a fundamental difference between membrane and soluble proteins and suggest that knowledge of the statistical distribution of hydrophobic residues may be useful for the prediction of the size of secondary structural elements in proteins.

EXTENSION TO OTHER CLASSES OF AMINO ACIDS

The basic statistical approach described above is not limited to hydrophobicity. Any single amino acid or class of amino acids of interest can be examined by means of an appropriate binary scale. The run test is simple and quickly tests the hypothesis that a distribution of residues along a chain is random. Of particular interest is the distribution of charged residues and reverse-turn formers. An examination of the PIR data base shows the observed r_0 density functions for these classes of amino acids to be similar to that of the hydrophobes shown in Fig. 3. The run test can also be applied to conserved *positions* along the chain. In this regard, we considered the distribution of the amino acid positions which are highly preserved in 67 globins (Ptitsyn, 1974). They fall into two classes: the reverse-turn formers which are highly conserved at helix ends and those such as His which associate with the heme pocket. There are 25 such positions in the 146AA chain of horse hemoglobin β-subunit. The run test shows that the distribution of these positions cannot be distinguished from the random case ($P[r > r_0] = 0.87$). This is entirely consistent with the idea that the primogenitor was a sequence with random characteristics that happened to work well and was subsequently preserved by evolutionary processes.

EPILOGUE

We have shown that the majority of proteins in the PIR data base have their hydrophobic residues distributed along the chains in a random fashion. The distribution function is peaked very close to $\langle r_0 \rangle = 0$ and has an approximately Gaussian shape. Charged residues and reverse-turn forming residues behave similarly. It is difficult to interpret the density distributions precisely because the data base is highly heterogeneous and its relation to the extant proteins is uncertain. In broad terms, however, there can be little doubt that randomness is a fundamental feature of proteins. The most profound implication of this observation is that the ancestral proteins of the biosphere may have originated from random sequences. This has a direct bearing on the question of how life could have arisen in a chaotic primordial environment. Given the existence of primitive chemical systems for the coding, translation, and synthesis of peptide sequences at some time in the primordial past, one can imagine that the systems produced astronomically large numbers of random copolymers (primitive proteins). Many of these would have been able to fold into compact structures and some of these might have had useful catalytic activities which in turn bestowed "survival advantages" on one or more of the systems. In effect, the survival process would act as a biological "Maxwell's demon" which selects useful proteins from a vast array of random sequences.

An important aspect of this scenario is that even for a randomly produced set of proteins, a wide range of r_0 values will exist; those with large values would clearly have what we consider to be non-random characteristics. For example, in the case of membrane proteins such as the β-adrenergic receptor which have values of f typical of soluble proteins, the runs of hydrophobes are concentrated in limited regions of the chain. This confers nonrandomness on the chain which could permit the protein to have more transbilayer domains than one might expect for the mathematically perfectly random chain. It is unlikely, of course, that the *direct* primogenitors of all the modern proteins were present in the initial set of primitive proteins produced during the hypothetical protein synthetic "big bang." One must suppose that f, L, and r_0 of the primitive ancestors have changed over the eons by the various evolutionary mechanisms so that membrane proteins could evolve from soluble proteins and vice versa. These ideas emphasize that randomness as defined by the run test is a convenient mathematical concept. In a random set, $r_0 = 0$ is the most likely value but certainly not the only value. Just as in a game of bridge, if enough hands are dealt eventually all 13 spades will be dealt to a single hand.

The most important questions which remain to be answered concern the heteronomous nature of the data base (and possibly the extant proteins). For example, can families of proteins be distinguished according to their extent of nonrandomness as defined by r_0? Table 1 leaves the impression, for example, that membrane proteins associated with the central nervous system may have a strong tendency toward nonrandomness. One might also expect structural proteins such as collagen with strongly repeating sequences to be nonrandom. Another important question concerns the distribution functions for families

of proteins and the positions of proteins in the distribution relative to evolutionary distance. A particularly interesting question concerns the possibility of uncovering nonrandom processes which are fundamental to evolution and to protein structure prediction.

In terms of predicting secondary and tertiary structure from primary structure, our results suggest that the most fruitful course of future investigation is likely to be in the area of physical studies of folding rather than in predictive sequence patterns. Indeed, the apparently random character of many sequences may explain why secondary structure prediction schemes based upon statistical analyses of data bases have only been marginally successful (Rooman and Wodak, 1988). On the other hand, a detailed knowledge of the randomness of amino acid class distributions and a study of the deviations from randomness may provide new insights to such predictive schemes. As shown in this paper, the knowledge that the distribution of certain residues is strictly statistical provides a useful mathematical starting point for structure prediction.

Discussions with Prof. H. G. Tucker concerning the statistics of Bernoulli sequences were invaluable. We thank Profs. Larry Vickery and Ralph Bradshaw for critically reading early versions of the manuscript. We particularly thank Prof. Ken Dill for his comments and for providing us with advance copies of his publications.

This work was supported by a grant from the National Science Foundation (DMB-8807431). R. E. Jacobs is an Established Investigator of the American Heart Association.

Received for publication 11 September 1989 and in final form 10 January 1990.

REFERENCES

Allen, J. P., G. Feher, T. O. Yeates, H. Komiya, and D. C. Rees. 1987. The structure of the reaction center from *Rhodobacter sphaeroides* R-26: the protein subunits. *Proc. Natl. Acad. Sci. USA.* 84:6162–6166.

Blanck, A., and D. Oesterhelt. 1987. The halo-opsin gene. II. Sequence, primary structure of halorhodopsin and comparison with bacteriorhodopsin. *EMBO (Eur. Mol. Biol. Organ.) J.* 6:265–273.

Chothia, C. 1984. Principles that determine the structure of proteins. *Annu. Rev. Biochem.* 53:537–572.

Creighton, T. E. 1984. Proteins. W. H. Freeman, New York. 515 pp.

Dickerson, R. E. 1980. The cytochromes *c*: an exercise in scientific serendipity. *In* The Evolution of Protein Structure and Function. D. S. Sigman and M. A B. Brazier, editors. Academic Press, Inc., New York. 173–202.

Engelman, D. A., T. A. Steitz, and A. Goldman. 1986. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu. Rev. Biophys. Biophys. Chem.* 15:321–353.

Finkelstein, A. V., and O. B. Ptitsyn. 1987. Why do globular proteins fit the limited set of folding patterns? *Prog. Biophys. Mol. Biol.*50:171–190.

Fitch, W. M., and E. Margoliash. 1967*a*. A method for estimating the number of invariant amino acid coding positions in a gene using cytochrome *c* as a model case. *Biochem. Genet.* 1:65–71.

Fitch, W. M., and E. Margoliash. 1967*b*. Construction of phylogenetic trees. *Science (Wash. DC).* 155:279–284.

George, D. G., W. C. Barker, and L. T. Hunt. 1986. The protein identification resource (PIR). *Nucleic Acids Res.* 14:11–15.

Henderson, R., and P. N. T. Unwin. 1975. Three-dimensional model of purple membrane obtained by electron microscopy. *Nature (Lond.). 257:28–32.*

Hopp, T. P. 1985. Prediction of protein surfaces and interaction sites from amino acid sequences. *In* Synthetic Peptides in Biology and Medicine. K. Alitalo, P. Partanen, and A. Vaheri, editors. Elsevier Science Publishers, Amsterdam. 3–12.

Ibragimov, I. A., and Yu. V. Linnik. 1971. Independent and stationary sequences of Random variables. Edited by J.F.C. Kingman. Wolters-Noordhoff, Groningen, FRG. 443 pp.

Jacobs, R. E. and S. H. White. 1989. The nature of the hydrophobic binding of small peptides at the bilayer interface: implications for the insertion of transbilayer helices. *Biochemistry.* 28:3421–3437.

Jukes, T. H. 1969. Evolutionary pattern of specificity regions in light chains of immunoglobulins. *Biochem. Genet.* 3:109–117.

Kauzmann, W. 1959. Some factors in the interpretation of protein denaturation. *Adv. Protein Chem.* 14:1–63.

Kerlavage, A. R., C. M. Fraser, F.-Z. Chung, and C. Venter. 1986. Molecular structure and evolution of adrenergic and cholinergic receptors. *Proteins.* 1:287–301.

Kolinski, A., and J. Skolnick. 1989. Monte Carlo simulation of equilibrium globular protein folding: α-helical bundles with long loops. *Proc. Natl. Acad. Sci. USA.* 86:2668–2672.

Kyte, J., and R. F. Doolittle. 1982. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 157:105–132.

Lau, K. F., and K. A. Dill. 1990. Theory for protein mutability and biogenesis. *Proc. Natl. Acad. Sci. USA.* 87:638–642.

Levitt, M. 1977. Conformational preferences of amino acids in globular proteins. *Biochemistry.* 17:4277–4285.

Lodish, H. F. 1988. Multi-spanning membrane proteins: how accurate are the models? *Trends Biochem. Sci.* 13:332–334.

McCrea, P. D., D. M. Engelman, and J.-L. Popot. 1988. Topography of integral membrane proteins: hydrophobicity vs. immunolocalization. *Trends Biochem. Sci.* 13:289–290.

Nolan, C., and E. Margoliash. 1968. Comparative aspects of primary structures of proteins. *Annu. Rev. Biochem.* 37:727–791.

Orcutt, B. C., D. G. George, and M. O. Dayhoff. 1983. Protein and nucleic acid sequence data base systems. *Annu. Rev. Biophys. Bioeng.*12:419–441.

Ptitsyn, O. B. 1974. Invariant features of globuin primary structure and coding of their secondary structure. *J. Mol. Biol.* 88:287–300.

Ptitsyn, O. B., and A. V. Finkelstein. 1980. Similarities of protein topologies: evolutionary divergence, functional convergence or principles of folding? *Q. Rev. Biophys.* 13:339–386.

Rao, S. P., D. E. Carlstrom, and W. G. Miller. 1974. Collapsed structure polymers. A scattergun approach to amino acid copolymers. *Biochemistry.* 13:943–951.

Richardson, J. S. 1981. The anatomy and taxonomy of protein structure. *Adv. Protein Chem.* 34:167–339.

Roach, S. A. 1968. The Theory of Random Clumping. Methuen & Co., London. 94 pp.

Rooman, M. J., and S. J. Wodak. 1988. Identification of predictive

sequence motifs limited by protein structure data base size. *Nature (Lond.)*. 335:45–49.

Rose, G. D. 1978. Prediction of chain turns in globular proteins on a hydrophobic basis. *Nature (Lond.)*. 272:586–590.

Rose, G. D., and S. Roy. 1980. Hydrophobic basis of packing in globular proteins. *Proc. Natl. Acad. Sci. USA*. 77:4643–4647.

Skolnick, J., A. Kolinski, and R. Yaris. 1989. Dynamic Monte Carlo study of the folding of a six-stranded Greek key globular protein. *Proc. Natl. Acad. Sci. USA*. 86:1229–1233.

Srinivasan, R. 1976. Helical length distribution from protein crystallographic data. *Indian J. Biochem.* 13:192–193.

Sternberg, M. J. E., and M. Thornton 1977. On the conformation of proteins: an analysis of β-pleated sheets. *J. Mol. Biol.* 110:285–296.

von Heijne, G. 1988. Transcending the impenetrable: how proteins come to terms with membranes. *Biochim. Biophys. Acta*. 947:307–333.

Wani, J. K. 1971. Probability and Statistical Inference. Appleton-Century-Crofts, New York. 315 pp.

White, S. H., and R. E. Jacobs. 1990. Observations concerning topology and locations of helix ends of membrane proteins of known structure. *J. Membr. Biol.* In press.

Zielenkiewicz, P., D. Plochocka, and A. Rabczenko. 1988. The formation of protein secondary structure. Its connection with amino acid sequence. *Biophys. Chem.* 31:139–142.