The Independent Distribution of Amino Acid Near Neighbor Pairs into Polypeptides

Michael H. Klapper

Department of Chemistry, The Ohio State University, Columbus, Ohio 43210

Received August 30,1977

Summary

With the assumption that individual amino acids are independently distributed into naturally occurring polypeptide chains, it is shown that amino acid pairs with 0-2 arbitrary intervening residues are also independently distributed, with a few possible exceptions. This is not true of N- and C-terminal amino acids.

While statistical analyses of protein sequencies have been used extensively for discerning evolutionary relationships, and predicting three dimensional structure, more general statistical properties have not been widely uncovered. Initial attempts of correlations with the physical properties of proteins (1, 2, 3) have not been pursued, and composition analysis has been limited (4, 5, 6). Since an understanding of the distributions underlying protein compositions should prove invaluable for the more specific statistical analyses, I have begun a systematic study of protein compositions and sequences.

I have already shown (7, and report submitted elsewhere) that individual amino acids are independently distributed amongst a large sample of polypeptide chains; i.e., the number of times one amino acid is observed in a randomly chosen polypeptide is determined only by its frequency. (It is important to note that while this observation is general, it does not apply to a specific chain in which there may be a strong interaction between composition and biological activity.) This conclusion is important since it allows the question of whether the observed frequency of an amino acid depends somehow on its cohorts. This paper focuses on the frequencies of amino acid pairs.

Methods

Let us start with the following thought experiment. The sequences of randomly chosen proteins are laid out in a linear string with a blank between each chain. This string is composed of 21 elements - the 20 amino

ISSN 0006-291X

acids and the blank derived from a chain termination event. Irrespective of N- or C-terminal posttranslational processing each protein was terminated and requires a single blank. (Proteins with internal deletions will be avoided). Since the population of all polypeptide chains is very large, the construction of this string is equivalent to sampling the 21 elements individually and with replacement. Since amino acids are distributed independently, the variance of the frequency, f_i, for the i^{th} element is estimated by (8)

$$\sigma_i^2 = f_i\ (1-f_i)/n \qquad (1$$

where n is the total number of elements in the string.

I now wish to determine whether the probabilities associated with each of the elements remains unchanged as the sampling proceeds. There are 21^2-1, or 440 possible adjacent pairs R_iR_j (Since two blanks cannot be side by side, 1 must be subtracted). If R_i in the string does not change the probability of finding R_j then the frequency of the pair, f_{ij} is the product of the individual frequencies, and the ratio

$$B_{ij} = f_{ij}/f_if_j \qquad (2$$

is equal to 1. Random errors are inherent in any sampling procedure, so that B_{ij} need not be unity. The variance of B_{ij} can be estimated by the usual method of error propagation

$$\frac{\sigma^2 B}{B_{ij}} = \frac{\sigma_i^2}{f_i^2} + \frac{\sigma_j^2}{f_j^2} + \frac{\sigma_{ij}^2}{f_{ij}^2} \qquad (3$$

with the variances on the right computed using equation 1. Were variations from unity due only to sampling errors, then the normalized deviates

$$z_{ij} = (B_{ij} - 1)/\sigma_{ij} \qquad (4$$

would have a Gaussian distribution. Were element pairing nonrandom, then the normalized deviates would not have a Gaussian distribution. Hence, the observed distribution of the z_{ij}'s can be used to test for random pairing.

Compositional analysis was performed on a data set of 207 polypeptides, one each from the family list of Dayhoff, et.al. (9) and from subsequent literature sources using the same family strategy. Polypeptides with unassigned Glx or Asx were not included. Computations were done with an IBM 370/168 using algorithms written in SPITBOL a variant of SNOBOL4 (10). Frequencies were calculated for individual amino acids, and for amino acid pairs separated by 0-2 arbitrary residues; R_iR_j, R_ixR_j and R_ixyR. Pairs were counted by scanning a sequence in one residue steps from the N- to C-terminal end. Thus, the right hand member of one adjacent pair, becomes the left hand member of the next, and so on. A listing of the polypeptides is available upon request.

Results

The amino acid frequencies determined from the data set (Table I) are similar to the results of Jukes, et.al. (4), who used a smaller sample. Because each complete chain reflects one termination, the termination frequency may be estimated as the ratio of the total chain number to the sum of chains

Table I

Amino Acid Frequencies

Amino acid	Frequency	Amino acid	Frequency
	(x 10^{-2})		(x 10^{-2})
alanine	8.88(.171)	glycine	7.76(.161)
leucine	7.44(.158)	serine	7.08(.154)
lysine	6.95(.153)	valine	6.84(.152)
glutamic	6.11(.144)	threonine	5.98(.142)
aspartic	5.49(.137)	arginine	4.70(.127)
isoleucine	4.60(.126)	proline	4.56(.125)
asparagine	4.34(.122)	glutamine	3.90(.116)
phenylalanine	3.47(.110)	tyrosine	3.47(.110)
cysteine	2.81(.099)	histidine	2.04(.085)
methionine	1.69(.077)	tryptophan	1.12(.063)
termination	0.747(.0517)		

The frequency of each amino acid was computed from the ratio of its occurrence to the total number of amino acids counted 27506 plus 207, the number of sequences sampled. The termination frequency is the ratio of 207/(27506 + 207). Estimated standard deviations calculated with equation 1 are presented in parentheses.

plus amino acids. Because most chains will have been shortened by posttranslational processing, the estimated termination frequency will be too high. This error will not interfere with the subsequent analysis which only requires internal consistency.

Calculated pair frequencies are too numerous to tabulate here. Using equations 1 through 4 normalized deviates, z_{ij}, were calculated, and plotted in a linearized form of the cumulative Gaussian distribution function. If the z_{ij}'s are normally distributed then the plot should be straight. The results

Table II

Normalized Pair Deviates of Absolute Magnitude Greater than 3

Pair	Normalized deviate	
···Tyr	-6.17	
Leu-Cys	-5.24	
···Asn	-4.92	
···Gln	-4.26	
His-Asp	-4.06	
Pro···	-3.69	
Asp-Cys	-3.47	
Met-Tyr	-3.47	
Glu-Pro	-3.40	(Gln-Pro 2.15)
Val···	-3.28	
···Thr	-3.25	
Gln-Glu	-3.22	(Glu-Glu 2.94)
Pro-Arg	-3.19	
Glu-Ser	-3.02	
Cys-His	3.40	
···Met	3.88	
···Ala	4.14	
Arg-Arg	4.18	
Cys-Cys	4.51	

Normalized deviates were calculated with equations 1 through 4
···R indicates the residue is N-terminal; R···, C-terminal.

for adjacent pairs are presented in Figure 1; the plot obtained with 440 R_iR_j pairs is almost straight, but shows distinct nonlinearity at both ends. To examine this curvature which indicates some nonrandom behavior, element pairs with an arbitrarily chosen $|z_{ij}|>3$ were singled out (Table II). C- or N-terminal residues constitute 42% (8 of 19) of this set, but only 9% of the total number of possible pairs. A plot of the 40 terminal residues (R_i···, and ···R_j) was also not linear - results not presented here. Hence, the terminal amino acids are not distributed independently.

Replotting the z_{ij}'s without the 40 terminal residues yields a straighter line with some remaining curvature (Figure 1). Asp/Asn or Glu/Gln are contained in 5 of the 11 amino acid pairs of Table II. Since the assignment of amide

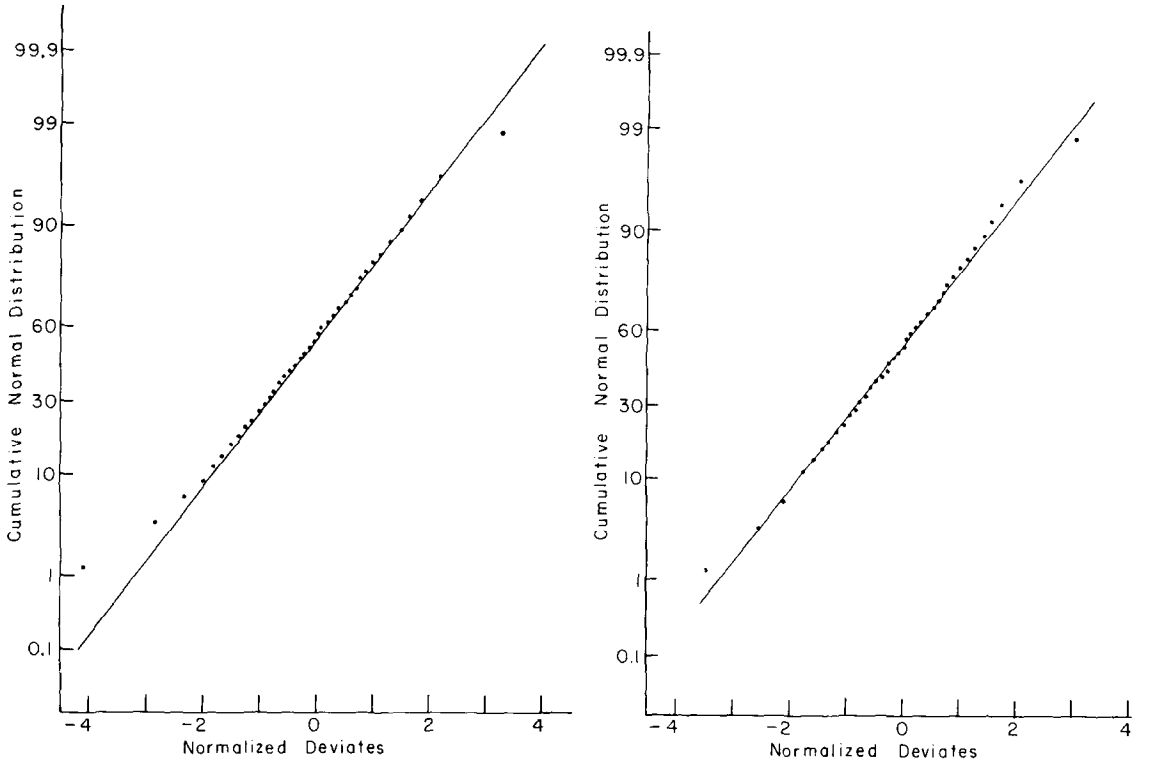

Figure 1. Normalized deviates of adjacent pair frequencies

Normalized deviates calculated from pair frequencies with equations 1-4 are plotted in a linearized form of the cumulative Gaussian distribution function. Each point is the average of 11 (left figure) or 10 (right figure) values. The right plot includes 400 amino acid pair frequencies; the left 400 amino acid pair, plus 40 N- and C-terminal frequencies.

side chains has been subject to some difficulties, the inclusion of these pairs may be due to experimental errors. The relatively large positive deviates associated with Gln-Pro and Glu-Glu support this suggestion. At least two more pairs, Cys-Cys and Cys-His, may be due to nonrandomness in the original data set. Approximately 39% of all Cys-Cys pairs were found in only three proteins, bovine serum albumin and two keratins. Approximately 63% of all Cys-His pairs were obtained from c type cytochromes which contain this pair as an invariant feature of the heme site. Thus, the curvature observed when terminal residues are omitted may be due primarily to sequencing errors, and nonrandom selection of the polypeptide sample. Nonetheless, four amino acid

pairs with $|z_{ij}| > 3$ are not accounted for, when only one would be statistically expected. Whether this is experimentally significant is unclear.

The same analysis for the noncontiguous pairs $R_i x R_j$, and $R_i xy R_j$ yielded closely similar results. I propose, therefore, that with a few possible exceptions amino acid pairs with 0-2 intervening, arbitrary residues are constructed independently.

Discussion

Independent distribution of amino acids into near neighbor pairs means that given the amino acid frequencies (which are not random, Table I and Jukes, et.al., 4), the construction of near neighbor pairs is random. Little or no chemical and biological pressures appear to "bias" local protein sequences. This result, which is not intuitively obvious, suggests that attempted correlations of three dimensional structure with local sequence for predictive purposes may have only slight chances of success. A sequence-structure correlation requires individual amino acids to show nonrandom preferences amongst various structural units, such as the α-helix, random coil, etc. But in turn this requires that amino acids are grouped nonrandomly in polypeptide sequences, which is not indicated by the results presented here.

There is, however, evidence for some nonrandom distribution, especially with the N- and C-terminal amino acids. This suggests that posttranslational processing follows some rules yet to be uncovered. Particularly noteworthy is the extremely low frequency of N-terminal tyrosine, occurring once in a sample of 235 chains. Apparent conversion on the left of a tyrosine is a "forbidden transition". N-terminal methionine occurs 7x more frequently than would be predicted. The high frequency of N-terminal methionine in bacterial proteins is well known (11), and is explained by its role of chain initiator. Of the 26 observed proteins with N-terminal methionine 21 are of bacterial or viral origin, prokaryotic proteins constitute approximately 30% of the sample set. Apparently trimming on the N-terminal side is more important in eukaryotes. There is one strong candidate for nonrandom pairing, Leu-Cys. I do not have a reasonable interpretaion of this observation.

The primary structures of biological polypeptides apparently do not exhibit a unique signature at a local level. Is this also true of higher structural levels? Unfortunately, there is insufficient data for the investigation of triples, and longer residue runs. However, other higher order order structures can be considered; e.g., the distributions of amino acid classes-apolar, polar, and charged, and pair separation distributions. These are currently under investigation.

Acknowledgements

I am indebted to the Ohio State University IRCC for providing free computer time. This work was supported in part, by a Career Development Award from the National Institutes of Health.

References

1. Bull, H.B., and Breese, K (1973) Arch. Biochem. Biophys. 158, 681-686.
2. Fisher, H.F. (1964) Proc. Nat. Acad. Sci. (USA) 51, 1285-1291.
3. Capaldi, R.A. and Vanderkooi, G. (1972) Proc. Nat. Acad. Sci. (USA) 73, 1964-1968.
4. Jukes, T.H., Holmquist, R., Moise, H. (1975) Science 189, 50-51.
5. Holmquist, R. (1975) J. Mol. Evol. 4, 277-306.
6. Holmquist, R., Moise, H. (1975) J. Mol. Evol. 6, 1-14.
7. Klapper, M.H. (1977) Fed. Proc. Abstracts 36, 837.
8. Dwass, M. (1970) Probability and Statistics (Benjamin and Co., New York) pp. 498-514.
9. Dayhoff, M.O., Barker, W.C., and Hunt, L.T. (1976) in Atlas of Protein Sequence and Structure vol. 5, suppl. 2, ed. Dayhoff, M.O. (National Biomedical Research Foundation, Washington, D.C.) pp. 11-18.
10. Griswold, R.G., Poage, J.F., and Polansky, I.P. (1971) The SNOBOL4 Programming Language, 2nd ed. (Prentice-Hall, Englewood Clifts, N.J.).
11. Waller, J.P. (1967) J. Mol. Biol. 7, 483-496.