J. Mol. Evol. 13, 245–252 (1979)

Journal of
Molecular Evolution

Degeneracy of the Information Contained in Amino Acid Sequences: Evidence from Overlaid Genes

Christian Sander and Georg E. Schulz

Institut für theoretische Physik der Universität Heidelberg, FRG
Max-Planck-Institut für medizinische Forschung, Heidelberg, FRG

Summary. The observed gene overlays in the viruses ΦX174 and SV40 show a surprising economy of information storage; two different amino acid sequences are read in different frames from the same stretch of DNA. This phenomenon appears contradictory in that the information in the two overlaid amino acid sequences is strongly interdependent, yet each of the two proteins has evolved to its own well-defined function. The contradiction can be resolved by assuming sufficiently large degeneracy of the information contents of amino acid sequences with respect to function. Such a degeneracy is familiar from homologous proteins where a given biological function is implemented by many different amino acid sequences. It is shown that the very existence of viral overlays allows to derive a lower limit for the magnitude of this degeneracy: The degeneracy is equal to, or greater than fourfold; on the average, at each position of the chain a choice of 1 out of 5 or less amino acids, and not a choice of 1 out of 20 is necessary for constructing a protein with a specified function. In addition, the strong dependence of overlay probabilities on chain length allows the definition of a maximal length of overlays; in bacterial viruses overlay regions should be shorter than about 150 residues.

Key words: Protein structure – Overlaid genes – Amino acid exchangeability – Informational degeneracy

Introduction

When correlating the DNA sequence of bacteriophage ΦX174 with the amino acid sequences of viral proteins, a surprising discovery was made (Sanger et al. 1977, Shaw et al. 1978): three genes are overlaid onto others, that is, three stretches of nucleotides are

*Present address: Chemical Physics Department, Weizmann Institute of Sciences, Rehovot, Israel

0022–2844/79/0013/0245/$ 01.60

translated twice into proteins. A similar situation has been encountered in the oncogenic virus SV40 (Durham 1978, Fiers et al., 1978, Reddy et al., 1978). The respective streches of DNA are read in two different phases, or frames, giving rise to two quite distinct amino acid sequences.

Because of omnipresent evolutionary pressure it is safe to assume that the overlaid viral proteins perform useful functions. In order to fulfill particular functions proteins have to consist of particular amino acid sequences. But overlaid sequences are strongly interdependent, neither one can be chosen without affecting the other. The probability, for instance, of achieving an overlay of two given sequences of only 30 amino acid residues each is as low as $(0.14)^{30} \cong 10^{-26}$ (Eq. 1 and Table 2). Therefore, it is virtually impossible to overlay two amino acid sequences that are specified exactly. In this light, the discovery of overlay seems puzzling.

The key to the overlay puzzle lies in the fact that the virus need not obey the constraints in such a stringent form. It is known from families of homologous proteins (Dayhoff 1972) that a protein with a given function may be constructed from a number of different amino acid sequences. Therefore, the information contained in the sequence is degenerate with respect to function. To simplify matters we consider this degeneracy as evenly distributed along polypeptide chains, that is, we do not distinguish between highly conserved residue positions as found at active centers and highly variable positions as found at many protein surfaces. This simplification seems justified as long as we are interested only in qualitative conclusions based on probability averages.

Method

The degeneracy can be quantitated on the basis of the observed viral overlays. For this purpose we divide the 20 standard amino acids into n groups. A protein is described by a sequence of group names (or 'group sequence', for short), implying that any member of a particular group can equally well act as group representative in the polypeptide chain. The procedure is illustrated in Fig 1. For group sequences the probability of overlays assumes reasonable values because not a particular amino acid residue but merely any member of a particular group of amino acid residues is required at each chain position. The overlay probability depends critically on the number of groups; at n = 20 it is as low as shown above, at n = 1 overlays are trivial. Using such a description our aim is to determine the partitioning into groups for which the assumption of a one-to-one correspondence between protein function and group sequence becomes reasonable. This partitioning is given by the maximum number n which still allows the observed viral overlays with reasonable probability. The maximal n is a measure for the degeneracy; at n = 20 there is no degeneracy at all, at n = 1 the degeneracy is complete.

Such an estimate should take into account the presently available structural knowledge. This is achieved by partitioning the 20 standard amino acids into groups which reflect their mutual exchangeabilities as found in families of homologous proteins. For this purpose we took the exchangeabilities m_{ij}, i,j = 1,2,.....20 given in Fig. 9-10 of Dayhoff's Atlas of Protein Sequences (1972), and rearranged them in such a way that $\Sigma_{i,j} m_{ij} \cdot (i-j)^2$ is minimized. The result is given in Table 1. This procedure yields a uniquely ordered list

G,P,D,E,A,N,Q,S,T,K,R,H,V,I,M,C,L,F,Y,W,

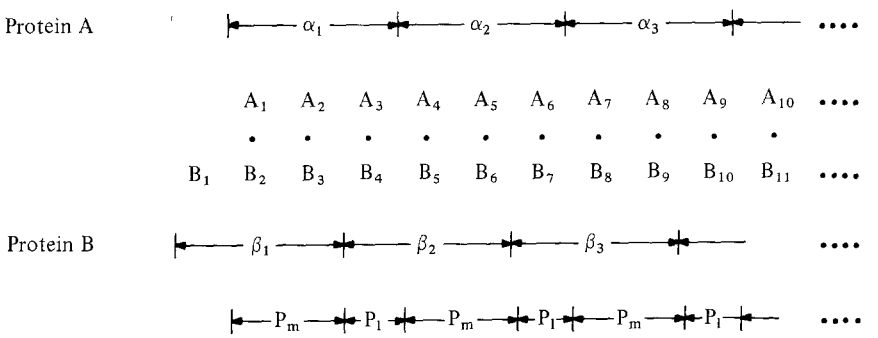

Fig. 1. Proteins A and B with amino acid group sequences $\alpha_1\alpha_2\alpha_3$... and $\beta_1\beta_2\beta_3$..., overlaid on the same nucleotide sequence. The two group sequences are given independently of each other. Relative to protein A the reading frame of protein B is shifted to the left, here α_i and β_i do not denote a specific amino acid, but a group name representative of any member of a specific group of amino acids. For example, if partitioning III of Table 2 is assumed, α_3 is, say, a member of the fourth group (either Lys or Arg or His) which corresponds to any of the codons of any of the members of the groups, i.e. the codon $A_7A_8A_9$ can be AAA, AAG, (for Lys); or CGU, CGC, CGA, CGG, AGA, AGG (for Arg); or CAU, CAC (for His). For a successful overlay the nucleotide sequences for protein A and protein B have to agree at all points. The probability of achieving this can be broken down into match (e.g. $A_1 = B_2$ and $A_2 = B_3$) probabilities P_m and link (e.g. $A_3 = B_4$) probabilities P_l. The values for P_m and P_l are characteristic of the particular partitioning of amino acids (see Table 2). A procedure for calculating these probabilities is given in the text

where on the average amino acid residues exchange with each other the better the closer they are in the list. Structural knowledge is then included by taking the list as the basis for partitioning the amino acids into groups. Groups are defined by splitting the list into several parts so that all members of a group are contiguous in the list (Table 2). Using this method, each group approximately corresponds to a certain type of amino acids, as for instance those with large aliphatic side chains. One has to keep in mind, however, that groups are introduced merely as a means to simplify the calculations. Only the list is based on observed data, and not the way of splitting it.

Now we derive expressions for overlay probabilities for two given group sequences, first in general, then for particular partitionings of the standard amino acids into groups. We first consider the overlay problem at each position in the sequence, define, 'match' and 'link' probabilities (P_m and P_l) characteristic of the particular partitioning and then compound these probabilities, yielding an overall probability for overlaying two given group sequences. The reader not interested in details may go on to Eq. 2 which is the main result.

We assume randomly chosen sequences of groups α_i and β_i, overlaid as shown in Fig. 1. Each group can be represented by any codon of any of its amino acids. Matching the pair $\alpha_1\beta_1$ means finding codons $A_1A_2A_3$ for α_1 and $B_1B_2B_3$ for β_1 which agree where they overlap: $A_1 = B_2$ and $A_2 = B_3$. The next match in the sequence, the pair $\alpha_2\beta_2$, requires $A_4 = B_5$ and $A_5 = B_6$. These two matched pairs are then linked by the condition $A_3 = B_4$.

The link probability (P_l) is essentially independent of the match probability (P_m), mainly because of the degeneracy of the genetic code in the third base of codons. For

Table 1. Observed amino acid exchangeabilities m_{ij} in families of related proteins. The numbers are taken from Dayhoff (1972). They are probability ratios, multiplied by 10 for convenience. Each number is the ratio of the observed probability that a particular amino acid will be replaced by another after an evolutionary distance of 256 accepted point mutations per 100 residues, divided by the probability that the exchange could occur by chance. For example, in two amino acid sequences at this evolutionary distance, Val and Ile would be found at equivalent positions 2.1 times more frequently than by chance. Dayhoff's numbers were reordered by minimizing $\Sigma_{ij} m_{ij} \cdot (i-j)^2$. This puts larger numbers closer to the diagonal such that amino acids which exchange preferentially with each other tend to be near each other in the list. More or less distinct groups of easily exchangeable amino acids emerge, such as Phe, Tyr, Trp. The ordered list was used as the basis for partitioning the standard amino acids into groups of roughly equivalent ones

	G	P	D	E	A	N	Q	S	T	K	R	H	V	I	M	C	L	F	Y	W
G	29																			
P	12	14																		
D	11	10	23																	
E	11	11	19	20																
A	13	14	12	12	14															
N	11	10	14	12	12	17														
Q	10	11	14	14	11	12	21													
S	12	11	12	11	13	13	11	13												
T	10	10	10	10	12	12	11	13	18											
K	7	8	10	10	9	12	11	10	9	24										
R	4	4	7	7	5	10	13	7	6	22	75									
H	6	6	9	8	8	13	12	9	9	11	20	59								
V	7	8	7	8	10	8	9	9	11	8	5	6	22							
I	6	6	6	7	8	7	8	8	10	7	5	7	21	25						
M	5	6	6	6	8	7	8	8	9	8	10	6	16	16	26					
C	6	4	5	5	7	6	5	11	9	4	2	3	11	9	12	166				
L	4	4	5	6	6	5	6	6	7	6	4	6	15	18	21	5	40			
F	2	2	2	3	4	3	3	4	5	3	4	7	7	11	11	2	12	70		
Y	1	1	1	1	2	2	2	3	2	3	3	7	3	6	6	1	6	66	137	
W	1	1	1	1	1	2	2	2	2	2	3	15	2	4	4	1	3	41	46	414

the same reason, the link probabilities from one position to the next are essentially independent of each other. In fact, in a number of overlay fits of random sequences done by computer we have found only a negligible number of cases where subsequent links were not consistent. This allows compounding P_m and P_l sequentially. The overall probability of overlaying two sequences of length N in a fixed relative phase is then

$$P_{ml}^N = P_m^N \cdot P_l^N \tag{1}$$

where $P_{ml} = P_m \cdot P_l$ is the probability for a single match and link. To obtain the match probability P_m for a particular partioning into n groups (Table 2) one can test all possible n^2 pairs of amino acid groups; counting all k_m cases in which the matching condition is fulfilled, one gets the match probability

$P_m = k_m/n^2$.

To calculate the link probability P_l one tests all $n^2 \cdot n^2$, pairs of pairs (such as $\alpha_1\beta_1$, $\alpha_2\beta_2$) for a successful match and link followed by another match ($A_1 = B_2$, $A_2 = B_3$, $A_3 = B_4$, $A_4 = B_5$, $A_5 = B_6$). The k_{mlm} successful cases yield the probability for a match-link-match

$P_m P_l P_m = k_{mlm}/n^4$.

So, the link probability is

$P_l = P_m P_l P_m / P_m^2 = k_{mlm}/k_m^2$.

The values of P_m and P_l are characteristic of a particular partitioning of amino acids.

Overlays of two given proteins A and B can occur in two relative reading frames; protein B can be shifted to the left as in Fig. 1, or to the right. Also, in principle, an overlaid protein can choose between many possible starting points on a genome; for two sequences of length M and N, with $M > N$, there are M-N possible overlay positions. These open choices increase the probability P(M,N) for a successful overlay. The probability of finding a possible overlay somewhere in one or the other reading frame is

$$P(M,N) = 1 - (1 - P_{ml}^N)^{2(M-N)} \qquad (2)$$

for two random group sequences of lengths M and N and for a partitioning with a match-link probability P_{ml}. Note, that the derivation of the probabilities implicitly assumes mutation rates large enough for the virus to test all possible codons compatible with the two specified sequences.

Results and Discussion

Eq. 2 can be used to calculate overlay probabilities for different partitionings in a given situation, say in ΦX174 or SV40. As an example, we show in Fig. 2a the probability for protein B of ΦX174, specified as a group sequence, to overlay somewhere on the viral genome without upsetting the existing sequence of groups. For a representative partitioning into six groups the overlay probability is smaller than 10^{-8} (Table 2). Considering the observed overlay structures it seems most likely that during evolution one of the overlaid proteins (probably protein A) preceded the other. With this assumption this value of 10^{-8} corresponds to the probability of evolving the later protein B with a given function without impairing the function of the already existing protein A. Even if one makes allowance for the fact that the virus would be happy to get any protein with a useful function, this probability is still unrealistically small because there are presumably much fewer than 10^8 functions useful for a virus. Consequently, for six groups, the assumption of a one-to-one correspondence between group sequence and protein function contradicts the existence of overlays. For seven or more groups the contradiction would be even worse.

In contrast to a partitioning into six groups, a partitioning into five (or less) groups yields a reasonably high value of 0.14 (or 1.00) for the overlay probability (Table 2). Thus, n = 5 is the largest number of groups for which the assumption is tenable

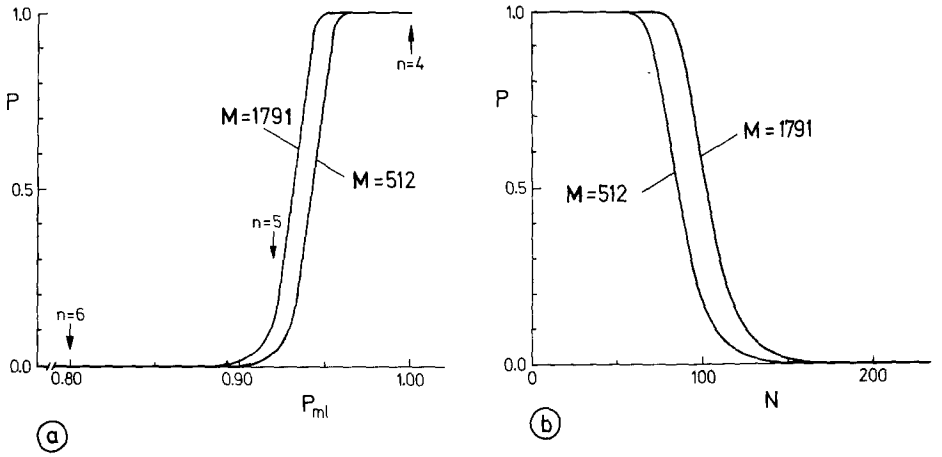

Fig. 2. Overlay probabilities P(M,N) for a given sequence of N amino acid residues on a given, longer sequence of M residues as calculated from Eq. 2. The lengths chosen are taken from ΦX174 where protein B of length 120 is overlaid on protein A of length 512, which is a part of the total genome with a length of 1791 nucleotide triplets. **a** The overlay probability as a function of the match-link probability P_{ml}, which depends on the particular partitioning chosen. The length of the shorter sequence is N = 120. The partitionings of Table 2 with n = 4, 5, or 6 groups are indicated by arrows. The sharp dependence on P_{ml} virtually rules out overlays for partitionings with more than five groups. **b** The overlay probability as a function of the length N of the overlaid protein. P_{ml} is taken as 0.92 corresponding to partitioning II of Table 2. The sharp cut-off as a function of N allows the definition of a 'maximal length', for example the length at P(M,N) = 0.1, which is here about 130 residues

that there exists one group sequence per protein function. Note that the probabilities for n = 5 and 6 differ by a factor as large as 10^7. Consequently, the simplifications we introduced, for example by partitioning the amino acids into groups, are not likely to change the main result.

The borderline between unrealistic (n = 6) and reasonable (n = 5) probabilities can also be stated in terms of P_{ml} values. On the basis of the observed length of overlays of 56, 91 and 120 residues in ΦX174 (Sanger et al., 1977, Shaw et al., 1978) and 38 residues in SV40 (Fiers et al., 1978, Reddy et al., 1978) we estimate that the P_{ml} value should be around 0.90 as obtained for partitioning II of Table 2. Note, that P_{ml} is a universal value intrinsically connected to protein architecture and independent of any of the simplifications and assumptions made here.

A further conclusion can be drawn about the length of overlaid polypeptide chains. In Fig. 2b the overlay probability is shown as a function of chain length. There occurs a sharp cut-off within about 40 residues, which allows the definition of a 'maximal length'. In the case of ΦX174 (M = 1791) for P_{ml} = 0.92 (partition II of Table 2) the 'maximal length' is about 125 residues. This length depends only logarithmically, that is weakly, on the length M of the region allowed for overlay on the genome (see Fig. 2b). Considering genome sizes typical for bacterial viruses the maximal length depends only on P_{ml} and thus should be universal. From the present data one expects a maximal length of about 150 residues. This prediction is subject to experimental verification.

Table 2. Four schemes of partitioning amino acids into groups and the corresponding matching and linking probabilities P_m and P_l. In all cases the amino acid ordering defined in Table 1 is retained. Here, partitioning is only a convenient basis for simple combinatorial mathematics. In reality the transitions between groups are smooth

Partitioning	No. of groups	P_m	P_l	$P_{ml}=P_m \cdot P_l$	P(1791,120)	P(512,120)	Amino acids
I	4	1.00	1.00	1.00	1.00	1.00	GPDE / ANQS / TKRH / VIMCLFYW
II	5	0.92	1.00	0.92	0.14	0.03	GPDE / ANQS / TKRH / VIMC / LFYW
III	6	0.86	0.93	0.80	$8 \cdot 10^{-9}$	$2 \cdot 10^{-9}$	GPD / EANQ / ST / KRH / VIMC / LFYW
IV	20	0.20	0.69	0.14	10^{-65}	$3 \cdot 10^{-67}$	all single

In conclusion, the viral overlays tell us that the degeneracy of the information contained in amino acid sequences is equal to or greater than fourfold; on the average, at each position of the chain a choice of 1 out of 5 or less amino acids, and not a choice of 1 out of 20 is necessary for constructing a protein with a specified function. Expressing this result in a simplified manner one can state that not all 20 standard amino acids are necessary to build a protein with a desired function but, on the average, only about five amino acids (or less) representing five (or less) groups of standard amino acids. Nothing in the argument presented here, however, allows choosing particular amino acids as these group representatives. Since our main results are based on the very strong dependence of the overlay probability on n, they are not likely to be affected by changes of details in the simplifications, such as the exact partitioning for a given n. The results are a first quantitative estimate of the degeneracy of the information contents of amino acid sequences, the existence of which is known from homologous proteins. This may constitute a first step in the process of understanding this degeneracy, which is relevant both in the context of early evolution and in the relationship between amino acid sequence and protein tertiary structure.

References

Dayhoff, M.O. (1972). Atlas of Protein Sequence and Structure, Vol. 5, Silver Spring: Natl. Biomed. Res. Found.

Durham, A. (1978). New Scientist **77**, 785–787

Fiers, W., Contreras, R., Haegeman, G., Rogiers, R., Van der Voorde, A., Van Heuverswyn, H., Van Herreweghe, J., Volckaert, G., Ysebaert, M. (1978). Nature **273**, 113–120

Reddy, V B., Thimmappaya, B., Dhar, R., Subramanian, K.N., Zain, B.S., Pan, J., Ghosh, P.K., Celma, M.L., Weissman, S.M. (1978). Science **200**, 494–502

Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, J.C., Hutchinson, II, C.A., Slocombe, P.M., Smith, M. (1977). Nature **265**, 687–695

Shaw, D.C., Walker, J.E., Northrop, F.D., Barrell, B.G., Godson, G.N., Fiddes, J.C. (1978). Nature **272**, 510–515

Received September 1, 1978; Revised June 2, 1979