PROTEINS: Structure, Function, and Genetics 4:99–122 (1988)

Oligopeptide Biases in Protein Sequences and Their Use in Predicting Protein Coding Regions in Nucleotide Sequences

Peter McCaldon and Patrick Argos
Biocomputing Division, European Molecular Biology Laboratory (EMBL), 6900 Heidelberg, Federal Republic of Germany

ABSTRACT **We have examined oligopeptides with lengths ranging from 2 to 11 residues in protein sequences that show no obvious evolutionary relationship. All sequences in the Protein Identification Resource database were carefully classified by sensitive homology searches into superfamilies to obtain unbiased oligopeptide counts. The results, contrary to previous studies, show clear prejudices in protein sequences. The oligopeptide preferences were used to help decide the significance of sequence homologies and to improve the more general methods for detecting protein coding regions within nucleotide sequences.**

Key words: **protein structure, protein coding regions, sequence homology, reading frame**

INTRODUCTION

Many proteins show clear sequence homology and can be grouped into families—for example, the globins and cytochrome c's.[1–3] It remains unclear, however, whether there is any underlying sequence similarity, i.e., shared oligopeptides within sequences of proteins with no obvious overall sequence homology or evolutionary relationship. Such a bias in oligopeptides could arise from either evolutionary divergence from common ancestral proteins or convergence toward some favored sequences or local structures.

A number of studies have been made of such peptide prejudices, albeit with conflicting results. Saroff[4] and Vonderviszt et al.[5] found a bias in protein sequences but neglected to take into account the large number of homologous sequences in the database. On the other hand, Klapper[6] has shown random association of amino acids in near-neighbor pairs, and Wilson et al.[7] concluded that the postulation of prejudice in specific amino acid sequences does not appear warranted. The latter work contained an error in calculation and also used a crude grouping of homologous proteins.

In this paper we have resolved the question by an extensive analysis of independent oligopeptides (sometimes referred to as "nmers") with lengths ranging from 2 to 11 residues in the Protein Identification Resource (PIR) amino acid sequence database.[8] By independent oligopeptides we mean those that have arisen from proteins or protein fragments which show no overall detectable sequence homology. Identical oligopeptides in closely related proteins, or which occur as repeated sequences within proteins as a result of gene duplication, are unlikely to be independent. The term *analogy* will be used to indicate a sequence identity in independent oligopeptides.[7] A much larger database has been used than in any of the previous studies, and we have very carefully grouped the homologous proteins. The results have been used to define a new criterion for deciding the significance of possible protein sequence homologies, and also to improve general methods for predicting protein coding regions within nucleic acid sequences.

MATERIALS AND METHODS

Grouping Related Protein Sequences

This study was carried out using release 8.0 of the Protein Identification Resource (PIR) amino acid sequence database,[8] which contained 3,557 sequences and 809,285 residues. Within the PIR databank, related proteins are grouped into superfamilies. These superfamilies were taken as the starting point for our grouping of protein sequences. A modified version of the FASTP program[9] was used to compare every sequence in the database with every other sequence not in the same PIR superfamily. Those sequence pairs, which met the criteria for possible homology suggested by Lipman and Pearson[9] (an initial FASTP score of at least 50 and an optimized score of at least 100), were then compared using the sensitive comparison technique of Argos[10] to confirm the homology. Furthermore, all protein sequences that shared at least one oligopeptide of seven amino acids or longer were also compared by the latter method. The PIR superfamilies were then combined if at least one sequence in a superfamily was homologous to at least one sequence in another superfamily. If homology was possible, though somewhat uncertain, the superfamilies were also combined; cases of this type rarely occurred.

A computer program was written to identify all sequences in the PIR database which contain a per-

Received April 18, 1988; revision accepted June 23, 1988.

Address reprint requests to Patrick Argos, European Molecular Biology Laboratory (EMBL), Biocomputing Division, Postfach 10.2209, Meyerhofstrasse 1, 6900 Heidelberg, Federal Republic of Germany.

fect sequence repeat at least five residues long. UWGCG programs (WORDSEARCH, SEGMENTS, REPEATS),[11] based on residue identity searches and alignment procedures of Wilbur and Lipman,[12] were then used to produce the best nonidentical alignment of each of the sequences with themselves. Sequences were treated as being of repeating type if a) there was a known duplication documented in the PIR database or elsewhere; b) there was reasonable homology over at least 50 residues; or c) the repeat was at least eight residues long and the residues were of at least three different amino acids. In cases of doubt, sequences were treated as being repeating. A total of 126 repeating-type sequences were removed from the database.

Counting the Independent Oligopeptides

All oligopeptides of length 2 to 11 were extracted from the sequences in the regrouped PIR database. The nmers were generated by moving along the sequence in steps of one residue for each oligopeptide count. For example, if a sequence contained six amino acids (AGHIRD), then there are four trimers (AGH, GHI, HIR, and IRD). The format of the PIR database allows punctuation characters within sequences. These typically represent breaks or uncertainties in the sequence. All such characters, together with those representing ambiguous residues, were treated as breaks in the sequence.

The oligopeptides were sorted by sequence and superfamily. For each oligopeptide, the maximum number of occurrences in any one protein sequence of each superfamily was determined. These maximums were then summed over all superfamilies, so as to yield the total number of independent occurrences of each specific oligopeptide. A more complicated counting scheme could be imagined where an nmer count would be weighted according to the degree of homology of the contributing sequence with other members of the superfamily cluster. This approach, however, would suffer from the bias that sequences have been determined for only a limited number of all possible species.

Measuring the Oligopeptide Biases

For each superfamily, all the independent oligopeptides of a given length were concatenated and then randomly shuffled one residue at a time, so as to create a "pseudosequence" for the superfamily. For example, if the maximum number of occurrences of ATG in any given sequence of a superfamily cluster is 3 while that for the trimer HYM is 2, then the pseudosequence to be randomly shuffled would be ATGATGATGHYMHYM. The pseudosequences were then used to represent each superfamily, and the oligopeptide frequencies were again counted as described above. The procedure was repeated 30 times for each nmer length, and means and standard deviations for the counts of specific oligopeptides were calculated. Note that this approach takes into account the local distribution of amino acids within each superfamily. This would be difficult to do if the expected frequencies were estimated using only probability calculations.

For nmers of length 2 to 4, (observed/expected) ratios and z-values were calculated. For longer oligopeptides insufficient counts were available to make sensible measurements for individual oligopeptides. The z-value is defined as the difference between the observed and expected values, expressed in standard deviations of the expected value.

Structural Correlation

For those tri- and tetrapeptides which occur in the Brookhaven database of known tertiary protein structures,[13] the z-values were correlated with the secondary structures of the respective oligopeptides and with the deviations of their tertiary structures. The secondary structures within the database proteins were determined from the atomic coordinates by the method and program of Kabsch and Sander.[14] The structural deviation of a given nmer is taken as the mean of the rms deviations of the main chain atoms after superposition of all possible pairs of identical 3mers or 4mers. The main chain (C, O, N, Cα) superpositions were performed by the method and computer program of Kabsch.[15,16]

Correlation With Nucleotide Sequences

The UWGCG program PEPDATA[11] was used to create peptide translations of the EMBL nucleotide sequence library[17] in all six possible reading frames. These translations were then searched with all the analogous octa-peptides in order to find candidates for the corresponding nucleotide sequences. The candidate sequences were then confirmed by checking the ten residues at either end of the search oligopeptide and by comparing the descriptions of the protein and nucleotide sequences in the respective databases.

The fraction of identical codons was determined for the analogous nucleotide sequence pairs. The expected fractional codon identity was calculated as follows, using local codon frequency tables for each species[18] to take into account codon biases for particular biological species:

$$E = \frac{\sum_{r=1}^{n} (\sum_{i=1}^{c} f_{ria} \cdot f_{rib})}{n}$$

where E is the expected fraction of identical codons in the nucleotide sequence, n is the number of residues (codons) in the peptide sequence, f_{ria} is the codon frequency for codon i coding for amino acid r in nucleotide sequence a, f_{rib} is similarly defined for the identical residue r coded for by nucleotide sequence b, and c is the number of different codons for the amino acid.

Computing Facility

All computing was performed on the EMBL VAX cluster. Special purpose programs were written in the programming language "C".[19] The PSQ program[8] was used to access the PIR database, and the UWGCG sequence analysis package[11] used for some of the sequence manipulations.

RESULTS AND DISCUSSION

Grouping Related Protein Sequences

Our grouping of related protein sequences reduced the 1,109 PIR superfamilies to 1,021 unrelated superfamilies. Proteins such as the immunoglobulin variable region, the immunoglobulin constant region, and the histocompatibility antigens, which exist as distinct superfamilies in the PIR classification, were grouped together by us. Further examples include our cluster of the genome polyproteins of poliovirus, foot-and-mouth disease virus, cowpea mosaic virus, and encephalomyocarditis virus.[20] These examples alone have considerable impact on the oligopeptide counting statistics as the former contains many members and the latter represent very long sequences.

PIR also define a broader grouping of superfamilies, which was used by Wilson et al.[7] This more extensive grouping, which reduces the entire database to only 27 clusters with such categories as DNA-associated proteins or miscellaneous proteins, groups many proteins that are clearly unrelated. Furthermore, we detected relationships between some of the members of even these groups; for example, some growth factors contain sequences from actin of the muscle contractile system, or cytochrome b's were found amongst the hypothetical proteins.

Recalculating Overall Amino Acid Composition in Proteins

Given the careful clustering of sequences into superfamilies, it was possible to calculate proper values for the amino acid composition of proteins in general. One protein sequence was randomly selected from each superfamily, and the composition of the resulting collection of sequences was calculated by merely counting the total number of each amino acid type. Dayhoff et al.[21] also performed this task with a limited data set available about 20 years ago. We subsequently list our results with the often considerably different Dayhoff values given in parentheses:

Ala 0.083 (0.087); Cys 0.017 (0.033); Asp 0.053 (0.047); Glu 0.062 (0.050); Phe 0.039 (0.040); Gly 0.072 (0.089); His 0.022 (0.034); Ile 0.052 (0.037); Lys 0.057 (0.081); Leu 0.090 (0.085); Met 0.024 (0.015); Asn 0.044 (0.040); Pro 0.051 (0.051); Gln 0.040 (0.038); Arg 0.057 (0.041); Ser 0.069 (0.070); Thr 0.058 (0.058); Val 0.066 (0.065); Trp 0.013 (0.010); and Tyr 0.032 (0.030).

Are Protein Oligopeptide Sequences Biased?

The total number of analogous matches in the database was determined for peptides of lengths 2 to 11. An analogous match or oligopeptide pair is defined by two identical sequence spans of length n residues found in protein sequences from different superfamilies or within the protein sequence itself. For example, suppose a 7mer was found twice (1A, 1B) in one protein sequence of a given superfamily and also found once (2A, 3A) in two other proteins, each from an unrelated superfamily. The number of analogous pairs or matches is six: i.e., (1A, 1B), (1A, 2A) (1A, 3A), (1B, 2A), (1B, 3A), and (2A, 3A). The number of independent 7mers in this example is four (1A, 1B, 2A, 3A). The expected number of matches was obtained by carrying out the same measurements on 30 randomized versions of the database. Means and standard deviations were calculated for the expected data for each possible nmer length (Tables I, II). All the results are at least 9 standard deviations above the expected values and indicate a clear bias in protein sequences for these oligopeptide lengths. It must be emphasized that all results given in this paper are based on analogous match counts. Since each occurrence of an nmer is an independent evolutionary event, then all possible matches or oligopeptide pairs must be counted.

For the longer nmers, a large number of the analogous matches resulted from internal repeats within a sequence that do not appear to be a result of overall gene duplication. The calculations described above were then repeated except that each oligopeptide was only counted once for each superfamily. Table I shows the results; the biases are still 6 or more standard deviations above expectation except for dimers where internal repeats are most unlikely to indicate evolutionary duplication events.

The number of analogous matches may be expressed as

$$N = T \cdot (P^n)$$

where T is the total number of pairwise comparisons, P is the probability of a single residue's being identical, and n is the oligopeptide length. The ratio of observed (o) to expected (e) matches then becomes

$$\frac{No}{Ne} = \frac{(Po)^n}{(Pe)^n}$$

where Po is the observed probability of a single residue's being identical and Pe is the expected probability of a single residue's being identical. The ratio (Po/Pe) is a measure of the bias observed in the oligopeptide distribution. Since the values of (No/Ne) for the different oligopeptide lengths are known (Table I),

TABLE I. Total Number of Analogous nmer Matches Including or Excluding Repeats Within Sequences

nmer length	Total nmers observed*	Different nmers observed**	Independent nmers observed***	Observed analogous matches†	Expected analogous matches‡	STD for expected matches¶	Observed/ expected	z-value§
Including repeats within sequences								
2	741,570	400	316,886	167,658,049	167,016,350	67,922.2	1.00	9.4
3	735,697	8,000	418,944	16,943,872	16,636,449	13,233.3	1.02	23.2
4	730,119	128,483	511,075	1,570,721	1,489,891	1,731.0	1.05	46.7
5	724,755	453,352	542,005	116,438	100,286	362.2	1.16	44.6
6	719,616	547,890	555,697	9,523	6,249	94.9	1.52	34.5
7	714,709	564,189	564,990	1,399	413	24.2	3.39	40.7
8	709,949	571,749	571,959	497	27	6.2	18.41	75.8
9	705,333	577,150	577,277	280	5	5.2	56.00	52.9
10	700,826	581,317	581,405	183	1	2.2	183.00	82.7
11	696,459	584,518	584,578	117	0	—	—	—
Excluding repeats within sequence								
2	741,570	400	164,847	38,469,806	38,526,874	68,922.6	1.00	−0.8
3	735,697	8,000	388,607	13,839,006	13,582,583	12,806.1	1.02	20.0
4	730,119	128,483	507,592	1,520,010	1,425,981	1,615.0	1.07	58.2
5	724,755	453,352	541,308	111,313	98,051	317.8	1.14	41.7
6	719,616	547,890	555,378	7,800	6,102	84.1	1.28	20.2
7	714,709	564,189	564,784	623	380	24.7	1.64	9.8
8	709,949	571,749	571,806	64	22	6.5	2.91	6.5
9	705,333	577,150	577,163	14	1	1.1	14.00	11.8
10	700,826	581,317	581,322	5	0	0.0	—	—
11	696,459	584,518	584,520	2	0	0.0	—	—

*The total No. of nmers or oligopeptides includes all the nmers (identical or not) in all sequences of all superfamilies.

**The value given refers to the total No. of oligopeptides with different sequences (e.g., AAA, AAC, AAD, etc.).

***The No. of independent nmers or oligopeptides is the sum of the counts for all nmers over all the superfamilies where the maximum occurrences of any given nmer in any one sequence of a superfamily is the nmer's superfamily count.

†The No. of observed matches is the No. of all possible pairwise combinations of identical independent oligopeptides or nmers. For example, if the maximum No. of AAA repeats in any one globin sequence is three while for cytochrome c's it is two, then there are six analogous matches when including repeats within sequences and only one match if repeats are excluded.

‡As for the observed analogous matches, the expected matches are mean values over 30 randomizations of the superfamily pseudosequences.

¶The standard deviation (STD) is calculated from 30 randomizations of the superfamily pseudosequences.

§(Observed–expected)/(standard deviation).

(Po/Pe) may be calculated. The results are presented in Table II. Some care needs to be taken with these figures since the standard deviations of the counts are greater for longer oligopeptides. Nevertheless, the results demonstrate that protein sequences show bias, especially in the longer nmers that they share.

Our analysis is similar to that carried out by Wilson et al.[7] They concluded that there is no oligopeptide bias in protein sequences. However, close inspection of their paper reveals an error. The expected number of analogous oligopeptide matches was not compared with the observed number of analogous matches but with the observed number of independent oligopeptides. For example, an oligopeptide which occurs independently four times, should be represented by six analogous matches. Furthermore, Wilson et al.[7] used the broad PIR sequence grouping to represent clusters of homologous proteins; this extensive grouping associates many nonhomologous proteins.

For nmers of length 2 to 4, the number of actual and expected occurrences of each different oligopeptide was counted. The distribution of the (observed/expected) ratios and z-values is presented in Table III. Di-peptide ratios range from 0.73 for avoided 2mers to 1.25 for preferred dipeptides. The range is larger for tripeptides and greater again for tetrapeptides. The data for 4mers is statistically rather poor, each tetrapeptide being represented on average by only about four independent occurrences in the database. Approximately 20% of the possible tetrapeptides do not occur at all in the database, and a further 20% only show a single independent occurrence.

About 32% of dipeptides may be classed as unusual, in that they occur with a z-value of more than 2.0 or less than −2.0. Using ±3.0 cutoffs, Saroff[4] observed that 30% of dipeptides were unusual, while we find 17%. However, his result was obtained without taking account of known homologies between sequences. The difference between the two figures (13%) probably represents those dipeptides which occur in known sequence homologies. For 3mers and 4mers, the respective frequencies of unusual peptides with ±2.0 as

TABLE II. Single Residue Bias

Peptide length	Bias (Po/Pe)
2	1.002
3	1.006
4	1.013
5	1.030
6	1.073
7	1.192
8	1.443
9	1.579
10	1.684

TABLE III. Distribution of nmer Data

nmer length	Mean (observed/ expected)	Standard deviation	Minimum value	Maximum value
2	1.00	0.078	0.73	1.25
3	1.02	0.213	0.13	2.09
4	1.05	1.030	0.00	66.67

Distribution of nmer z-values

nmer length	Mean z-value	Standard deviation	Minimum z-value	Maximum z-value
2	−0.13	2.25	−5.81	9.27
3	−0.03	1.34	−4.99	9.44
4	0.01	1.26	−4.13	13.53

z-value ⩽	2mer percent of total*	3mer percent of total*	4mer percent of total*
−5.00	0.8	0.0	0.0
−4.00	3.3	0.1	0.0
−3.00	9.8	1.0	0.0
−2.00	17.7	5.7	1.4
2.00	86.0	93.3	93.0
3.00	92.8	98.2	97.6
4.00	96.5	99.5	99.2
5.00	97.5	100.0	100.0

*The "percent of total" refers to the percentage of the total No. of observed nmers with specific length that have a z-value less than or equal to that shown.

TABLE IV. 2mers With z-Value >3.0 or < −3.0

			Standard		
			deviation		
	Observed	Expected	of	Observed/	
	count	count	expected	expected	z-value
AA	2,568	2,183	41.6	1.18	9.27
DI	1,018	901	19.9	1.13	5.86
EE	1,580	1,322	33.2	1.20	7.76
EN	949	870	25.7	1.09	3.07
EQ	901	824	25.4	1.09	3.04
FD	734	655	26.3	1.12	3.00
FS	951	838	26.8	1.14	4.22
GG	1,705	1,563	24.5	1.09	5.81
GK	1,391	1,235	43.6	1.13	3.57
HH	253	205	12.5	1.23	3.81
IN	935	784	27.3	1.19	5.53
IT	1,095	987	29.3	1.11	3.68
KK	1,348	1,239	32.2	1.09	3.38
LK	1,702	1,573	30.5	1.08	4.24
LS	2,029	1,874	33.0	1.08	4.70
LT	1,738	1,596	39.1	1.09	3.64
MA	722	634	27.1	1.14	3.26
NP	823	677	28.4	1.22	5.15
NY	553	499	17.6	1.11	3.09
PE	1,134	935	26.4	1.21	7.53
PP	926	868	18.4	1.07	3.17
PV	1,120	1,027	26.7	1.09	3.47
QQ	736	587	20.9	1.25	7.15
RC	383	333	16.3	1.15	3.08
RR	1,350	1,148	30.2	1.18	6.70
SG	1,701	14,54	30.3	1.17	8.17
SS	1,793	1,594	32.1	1.12	6.20
TP	1,027	923	29.4	1.11	3.53
TV	1,353	1,233	33.5	1.10	3.58
VT	1,360	1,227	30.2	1.11	4.39
AN	1,015	1,091	25.1	0.93	−3.03
AP	1,127	1,235	29.8	0.91	−3.63
AT	1,347	1,466	35.5	0.92	−3.36

(continued)

TABLE IV. 2mers With z-Value >3.0 or < −3.0 (Continued)

	Observed count	Expected count	Standard deviation of expected	Observed/ expected	z-value
DQ	560	697	31.0	0.80	−4.42
DR	812	941	28.0	0.86	−4.59
EF	690	756	18.3	0.91	−3.62
EP	758	930	32.9	0.82	−5.22
ES	1,096	1,291	33.5	0.85	−5.81
FA	810	956	34.6	0.85	−4.22
FM	270	335	15.6	0.81	−4.15
GA	1,602	1,760	41.5	0.91	−3.80
GP	898	1,061	36.1	0.85	−4.52
HE	400	465	16.7	0.86	−3.89
HM	143	197	14.5	0.73	−3.71
IL	1,350	1,476	35.3	0.91	−3.57
IM	358	446	17.1	0.80	−5.13
IV	1,017	1,110	27.7	0.92	−3.36
KM	411	475	21.1	0.86	−3.06
KS	1,089	1,245	34.4	0.87	−4.54
LG	1,718	1,820	29.8	0.94	−3.42
LI	1,318	1,466	36.5	0.90	−4.07
LV	1,695	1,806	36.9	0.94	−3.00
ND	671	762	23.8	0.88	−3.81
PI	718	793	24.6	0.90	−3.07
PL	1,238	1,345	30.9	0.92	−3.46
PM	318	394	25.1	0.81	−3.02
QD	627	703	25.1	0.89	−3.03
QS	811	887	21.5	0.91	−3.54
RM	405	470	20.5	0.86	−3.18
RT	904	1,034	25.0	0.87	−5.20
SE	1,172	1,290	28.1	0.91	−4.19
SK	1,107	1,225	31.8	0.90	−3.70
SN	886	981	28.8	0.90	−3.29
SP	1,023	1,097	24.8	0.93	−3.00
TE	1,005	1,109	27.5	0.91	−3.80
VG	1,275	1,444	38.9	0.88	−4.35
WG	271	324	16.9	0.84	−3.11
WP	191	235	12.6	0.81	−3.53
YE	574	644	19.4	0.89	−3.58

TABLE V. 3mers With z-Value >4.0 or <-4.0

	Observed count	Expected count	Standard deviation of expected	Observed/ expected	z-value
AAA	436	281	17.3	1.55	8.93
DDW	35	17	4.4	2.09	4.11
DED	109	81	6.9	1.34	4.07
DGK	123	88	8.0	1.40	4.37
DPE	98	67	7.5	1.46	4.12
DPN	80	49	6.5	1.63	4.71
DPR	82	57	6.2	1.44	4.06
EEE	209	129	11.8	1.62	6.75
EEL	230	150	10.5	1.53	7.58
EEM	69	43	6.3	1.62	4.18
EQL	139	93	7.9	1.50	5.87
FGG	117	80	8.5	1.46	4.30
FKD	78	52	5.8	1.51	4.55
GKT	132	96	8.2	1.37	4.35
HQQ	39	20	4.7	1.98	4.12
INN	84	54	7.4	1.55	4.00
INP	73	45	6.1	1.61	4.53
IPE	106	65	10.2	1.63	4.02
KDI	112	71	9.0	1.57	4.54
LKD	157	111	9.9	1.42	4.64
LKE	189	135	10.2	1.40	5.30
LRR	165	118	11.6	1.40	4.04
MAE	82	53	5.7	1.55	5.07
NGK	107	74	6.0	1.45	5.47
NPE	95	53	8.1	1.80	5.25
NYI	59	36	5.0	1.65	4.61
PAP	127	88	7.7	1.44	5.11
PEE	120	83	6.4	1.45	5.81
PPP	134	73	9.5	1.83	6.41
QAA	155	121	8.4	1.28	4.05
QQL	106	68	5.6	1.56	6.77
QQQ	77	37	7.6	2.06	5.23
RRC	43	26	4.2	1.65	4.01
RRR	183	95	9.4	1.93	9.44
RWL	50	28	4.8	1.76	4.48
SSS	319	194	14.2	1.65	8.81

(continued)

TABLE V. 3mers With z-Value >4.0 or < −4.0 (Continued)

	Observed count	Expected count	Standard deviation of expected	Observed/ expected	z-value
TIP	89	64	6.0	1.39	4.18
TLT	167	127	9.1	1.31	4.39
TPV	120	80	8.4	1.51	4.81
TSG	148	111	9.0	1.33	4.07
TWD	36	19	4.1	1.92	4.20
YNP	55	32	4.4	1.75	5.40
YQQ	47	25	4.3	1.87	5.10
DTQ	31	54	5.1	0.58	−4.47
EKS	62	108	9.2	0.57	−4.99
EPA	63	97	7.6	0.65	−4.47
ESK	68	103	8.0	0.66	−4.38
ESP	59	86	6.6	0.69	−4.00
EVG	75	112	8.9	0.67	−4.16
FEL	54	88	7.7	0.61	−4.43
IKI	45	79	7.2	0.57	−4.74
LEV	103	146	10.4	0.71	−4.13
LLI	126	179	13.3	0.70	−4.01
RSE	66	96	6.9	0.69	−4.33

z-value thresholds is about 12% and 8%. While the percentages are smaller, it must be emphasized that the possible number of tripeptides and tetrapeptides is 20 and 400 times that of dipeptides, respectively.

Table III shows that the strongly preferred nmers have much higher z-values than those most avoided; for example, the smallest 4mer z-value is −4.13 while the largest is 13.53. The examples of Tables IV–VI also illustrate the observation. Apparently protein structural preferences are much more finely graded and exaggerated, whereas those not preferred are simply avoided.

Unusual Oligopeptides and Their Properties

We have identified two groups of unusual oligopeptide sequences.

1. Oligopeptides with significantly high or low z-values (Tables IV–VI). We can only calculate z-values for nmers of length 2 to 4, and the data for 4mers is statistically somewhat weak. Particularly unusual are tripeptides such as CCH, DDW, and QQQ, which occur more than twice as often as they would if protein sequences were simply random, and DHE, RFM, and YPP, which occur less than half as often as expected. Even more striking are the tetrapeptides AAAA, which occur 122 times compared with an expected 50 times; PTGC, which is found 5 times as often as expected; and EPAL, which is not found though expected to occur nearly 12 times.
2. Oligopeptides which occur in more than one superfamily (Table VII). Only the nmers that are seven residues long and fulfill this condition are listed in Table VII; the shorter oligopeptides are simply too numerous to show. When designing a protein or linker, it is of use to know that a string of a repeated amino acid type confines itself to Ala and Gly for small residues, Gly and Pro for turn-prone amino acids, Ser only for a polar but uncharged residue, Leu for hydrophobic, and only Arg and Glu amongst charged amino acids. Table VIII lists nonredundantly the oligopeptides, protein names, PIR cryptic designation, and species of the sequences where analogous 8mers to 11mers were found.

The mean z-values of the nmers which are constituents of longer favored oligopeptides are all positive and nonzero, indicating that unusual oligopeptides tend to contain smaller unusual oligopeptides (Table IX). This is confirmed by the cross-correlation coefficients calculated between the z-values of oligopeptides and their constituent oligopeptides, though the correlation coefficients are not all large (Table X).

The z-values for trimers were correlated with a set of structural parameters derived from the Brookhaven database of known protein tertiary structures. Only the three-dimensional structures with amino acid identity less than 20% were used in any given tertiary structural family. The secondary structure and atomic coordinates for the 12 main-chain atoms

TABLE VI. 4mers With z-Values >7.0 or < −3.0

	Observed count	Expected count	Standard deviation of expected	Observed/ expected	z-value
AAAA	122	50	7.3	2.44	9.85
AICH	5	1	0.6	7.94	7.12
CATW	4	0	0.4	9.30	9.42
CCNP	5	1	0.6	6.25	7.49
CDGF	6	1	0.6	5.45	8.25
CHVY	4	0	0.4	9.30	9.42
CIFC	5	0	0.5	15.15	8.78
CQSW	4	1	0.5	7.55	7.11
CTYD	4	1	0.5	5.71	7.01
DEQM	8	2	0.9	5.23	7.32
ECCH	4	0	0.4	12.12	9.39
ECCN	4	1	0.4	8.00	8.05
EEEE	49	14	3.6	3.58	9.91
ENWH	4	0	0.5	9.30	7.86
FMPN	4	1	0.5	5.71	7.01
FNPE	12	3	1.2	4.18	7.50
GPWM	4	1	0.5	8.00	7.00
HFYT	4	1	0.4	5.19	7.43
HHHC	2	0	0.2	28.57	7.88
HHHH	8	0	0.6	17.02	12.16
HHPA	7	1	0.9	7.00	7.05
HHWM	2	0	0.2	28.57	7.88
HICR	5	0	0.5	13.51	9.96
HLWD	5	1	0.6	5.56	7.06
HPYF	5	1	0.5	9.43	8.19
HRHI	6	1	0.5	11.32	10.02
HYLN	9	1	0.6	6.77	13.53
IPWI	8	1	1.0	6.67	7.05
IYHC	3	0	0.3	17.65	8.18
KCNQ	5	1	0.5	6.25	8.94
KPKP	18	5	1.7	3.75	7.72
KQCH	4	1	0.5	8.00	7.00
KTWT	7	1	0.7	5.83	7.85
KYFQ	7	1	0.7	5.69	8.23
LCKQ	8	2	0.8	4.91	8.17
LTPE	23	8	2.0	2.99	7.77
MFFS	9	2	1.1	5.88	7.08
MGNI	8	1	0.9	5.44	7.43
MKWV	4	1	0.4	6.35	8.92
MMKR	5	1	0.6	5.75	7.05
MRTF	5	1	0.5	5.38	8.34
NEKW	6	1	0.7	5.61	7.52
NYHL	7	2	0.8	4.67	7.29
PEPE	22	5	2.0	4.58	8.64
PHPY	6	1	0.7	6.00	7.46
PKPK	17	5	1.6	3.23	7.25
PPPP	28	7	2.1	4.26	10.39
PTGC	12	2	1.3	5.53	7.86
QCCH	2	0	0.2	28.57	7.88
QQFE	7	2	0.7	4.58	7.86
QQQQ	17	2	1.2	7.17	12.25
QQWM	4	0	0.4	23.53	8.85
QWYQ	3	0	0.4	11.11	7.09
RHIY	6	1	0.7	5.31	7.42
RHPD	6	1	0.6	5.13	7.51
RRRR	48	9	3.6	5.20	10.89
RYTQ	8	2	0.8	4.28	7.67

(continued)

TABLE VI. 4mers With z-Values >7.0 or < −3.0 (Continued)

	Observed count	Expected count	Standard deviation of expected	Observed/ expected	z-value
SSSS	104	34	6.7	3.07	10.54
SWWS	5	1	0.5	10.00	9.00
TPEQ	18	4	1.8	4.36	7.57
TWDP	5	1	0.5	5.56	7.52
TWRH	4	0	0.4	13.33	9.49
TYHC	4	0	0.5	10.81	7.81
VCGH	4	1	0.4	5.19	7.43
VCMD	6	1	0.5	11.32	11.21
WCFK	3	0	0.4	15.00	7.69
WFHW	2	0	0.2	66.67	10.94
WLNG	5	1	0.5	4.07	7.52
WQQL	6	1	0.7	6.00	7.46
WYNR	5	0	0.6	10.64	7.99
YRQL	11	3	1.2	4.12	7.13
YSHP	6	1	0.6	5.61	7.69
AKEN	1	7	1.9	0.15	−3.11
AQLV	1	10	2.8	0.10	−3.37
AQRT	0	6	1.6	0.00	−3.71
ASAQ	2	10	2.5	0.20	−3.28
DGGD	2	8	1.9	0.25	−3.07
DYLG	0	5	1.6	0.00	−3.30
EPAL	0	12	3.4	0.00	−3.51
EVGL	2	12	3.1	0.16	−3.36
FAIA	1	7	1.9	0.14	−3.07
GAKN	2	8	2.0	0.25	−3.02
GLKI	1	11	3.2	0.09	−3.05
GSND	0	6	1.5	0.00	−3.60
IAIK	1	7	1.6	0.14	−3.99
IDDQ	0	4	1.1	0.00	−3.14
KGAE	2	10	2.4	0.19	−3.41
KLIL	4	15	3.3	0.27	−3.35
LDYG	0	5	1.7	0.00	−3.18
PALA	4	14	2.4	0.29	−4.13
PIES	0	6	1.8	0.00	−3.02
SKTT	2	8	1.6	0.26	−3.70
TEGA	3	10	2.4	0.29	−3.11
TKSA	1	11	2.7	0.09	−3.65
TQRD	0	4	1.2	0.00	−3.09
VAVV	5	14	2.8	0.35	−3.30
VIAT	2	9	2.4	0.22	−3.05

were then collected for all possible 3mers in the nonhomologous database. The trimer z-values as observed from the protein sequence database were then associated with identical trimers in the structural database. The resulting list was subdivided into z-value ranges as $-2.0 \leqslant$ z-value< -1.0. The main-chain atoms for all pairwise combinations of identical trimers were superposed, and the mean rms deviation for all pairs in a given z-range was calculated. The mean main-chain rms deviations decrease with increasing z-value (Fig. 1), suggesting that over-represented oligopeptides are structurally conservative. This is supported by the observation that the percentage of helical residues observed in the trimers increases with increasing z-value range (Fig. 2).

For 2, 3, and 4mers, cross-correlation coefficients have been calculated between the oligopeptide z-values and the z-values of the oligopeptide with mirror sequence. For example, the mirror sequences of AV, AVG, and AVGT are, respectively, VA, GVA, and TGVA. The results are shown in Table XI. The maximum correlation coefficient observed is less than 0.2, which demonstrates that sequence preferences are dependent on amino acid order. Table XII lists some examples where the sequence mirrors display radically different z-values.

Oligopeptides were considered to be repetitive if i) for lengths less than or equal to 4, one amino acid type accounts for at least 75% of the oligopeptide sequence, and ii) for lengths greater than 4, one or two amino acid types represent at least 80% of the oligopeptide sequence.

For oligopeptides of length 2 to 4, the mean z-values of repetitive nmers were determined (Table XIII). For longer oligopeptides (length 5 to 11), the fraction of nmers which are repetitive was determined for the nmer sets which appear in analogous matches (Table XIV). The results indicate a striking preference for

TABLE VII. Oligopeptides of Length 7 That Occur in More Than One Protein Sequence Family

AAAAAAA	AAAAAAL	AAAAAAS	AAAAAAT	AAAAATA	AAAATAA	AAAGGAA	AAALAAA
AAATAAA	AAEEAGL	AALALEK	AALKQFD	AAPAPAE	AASRAAA	AEAAAEA	AEEAGVD
AEGDVAA	AEIAQRL	AEITSQT	AERVRQL	AERVVAD	AFLLLLS	AGDGAAA	AGFLEGG
AGGAAAA	AGGEGLV	AGGKAGK	AGTRPAA	AIAAADL	ALAKELN	ALGSPFD	ALLALLA
ALLALVL	ALLDTGA	ALLPRLL	ALNNGTL	AMKILDK	APAGVTT	APAPAPA	APRAAPP
APVAAPA	AQTHLKG	AQVVIET	ARLEALK	ASDLVTL	ASEASRL	ATSSSSS	AVAVAAG
AVIGAVV	AVVAGLL	AVVPITA	AYEEDRE	AYLVGLF	AYSNANK	CDERVSS	CGIVLNF
CKTHDCC	CVDTSGS	CVGSPTT	DAAAAAA	DAIAAAD	DALKAAG	DAQKDAD	DDDSADD
DDLIIGL	DEDEDED	DEDEEEE	DEEEEEE	DEEGGGL	DEVVDVY	DFIDTYL	DGSGNPV
DLAEVAP	DLIAYLE	DLKDKRV	DLLLLSE	DLLPFLS	DLSAKEA	DLVASVS	DLVKAIE
DNEIFLT	DPDPAVT	DPKAKSK	DPKTGKQ	DSGARGS	DSGGPLI	DSSSSSS	DYSVSAG
EAAASTA	EAARAGE	EAGAKKV	EAKALAE	EALELAE	EALKDAQ	EAQAKKE	EAVAKAD
EDEDEDD	EDEDEDE	EDEEEEE	EDEEGGG	EDEVPSG	EDLAALE	EEAEEEA	EEAGVDL
EEARKKA	EEDEEGG	EEEEEEE	EEEGAQE	EEGAQEE	EEGGGLF	EELLTTO	EESGGGL
EGEAEEE	EGLHLLA	EGVPSTA	EKEARKK	EKEESEE	EKKAADA	EKTSPYE	EKVDFDD
ELAGNAA	ELANKVD	ELATKAG	ELETAKS	ELLDFLH	EMTKKQV	EPEPEPE	EPSTAKT
EPVPGDP	ERARSER	ERIRRER	ESDTAQQ	ESEDEED	ETLRIYL	EVPEVTV	FALSVVS
FDTKAIE	FEKINEA	FETLDDL	FFEQESS	FIBLFDS	FISRHNS	FLGFLPK	FLLLLAD
FLLLLSL	FLQEAQV	FLRVADI	FLVAMSL	FLVGILF	FSDGLES	FSLLLLL	FSQLRAA
FVILTVL	FVLLLSL	GAGKSTS	GAKTFAE	GASASAA	GASIVHR	GDFGAPQ	GDTDSLF
GDVVRAR	GEAEEEG	GEALARM	GELFDSL	GERRAVE	GERVTLT	GGAAAAA	GGDLVKP
GGGKGGS	GGGSGGG	GGKGGKG	GHNVTVI	GIILLLA	GKATLTV	GKKLVLS	GKKVIHA
GKLQHLE	GKVGGHA	GKWSPEL	GLARVTR	GLKTEDE	GLSERGN	GLSVGLV	GLTFQQN
GLVLAAG	GMANPSP	GPASRSV	GPEVEAA	GPGATNA	GPPPSGP	GQLLASA	GRGEISA
GSAPAAA	GTFAESR	GTRRRRR	GTVDFDE	GVANLDN	GVASALA	GVETTTP	GYLSALR
GYSGPTV	HEGGNVS	HLRELLT	HPDQDIS	HVAGAAA	IGIILLL	IGRKYDD	IKSKAIG
ILALVGA	ILCLSLS	ILDEVIR	IPEGEKV	IPKDIQL	IPRTPAR	IPSGVDA	ISFLLSD
ISLEDLP	ITEDDIE	IVADDLT	IVAVIGA	IVEPTEK	IWYNNNV	KAAKKAG	KAAVTTI
KAGKRRR	KAIDVNG	KARKEAE	KAVAEAY	KDEILAL	KDLIAYL	KECQKLL	KEDLVVC
KEEAEEL	KEGTLDF	KERSGVS	KGDKVKA	KGEDITL	KGEKPYD	KGLEWVS	KGRTWTL
KGYEKAL	KITPSLA	KKRLQAF	KKVLAAF	KLKERMD	KLLIEME	KMDEALA	KMIGGIG
KNMDNIK	KPKPGKR	KSAVTAL	KSCVGCH	KTELQAI	KTNQQFE	KTPQNSA	KTVTSLD
KVIMGAV	KVLGADG	LAASLAN	LAGLAAA	LAGLLLL	LAKEVQA	LAKVEKE	LALLVSI
LALVGAA	LANEGKV	LANENFE	LARAAAR	LAREKFA	LARRLRG	LARVTRA	LDGSRLI
LDKYLEN	LDVNNPR	LEEAEKA	LEEPLRK	LEHTINN	LELALEA	LELLGQT	LFALSLD
LFLALLS	LFLLTLL	LGLVLAA	LHGSEDQ	LHLSVLR	LHVLIQF	LIIKRKP	LISLVDG

(continued)

TABLE VII. Oligopeptides of Length 7 That Occur in More Than One Protein Sequence Family (Continued)

LITPVLQ	LKDATSK	LKEQLEK	LKLPLSV	LKREDLL	LLADLVR	LLAIGGA	LLASAAS
LLAVTVF	LLAYFLP	LLDTGAD	LLEAIDA	LLEPGDT	LLGGLAS	LLGVFML	LLKEAEK
LLKESLL	LLLDLAL	LLLGGLT	LLLIIIL	LLLLLAG	LLLLLLC	LLLLLLL	LLLLLVV
LLLQLLG	LLLSLIG	LLLVAVL	LLMKYLG	LLQLTSG	LLSGALA	LLTLGLI	LLVTFLA
LMRIALA	LNLKRKV	LPEFSLT	LPLLGLI	LPLLLPL	LPNKKPN	LPPPEEE	LQGVLAN
LQRALEI	LQRLLQG	LRELLTT	LRQLEVA	LRRIGRF	LSGITGA	LSKMVSE	LSLSSLT
LSSLPEI	LSSSTQA	LSTSGTT	LSVGLVG	LSVPREE	LTCLLAV	LTGDTEP	LTGGLPE
LTLALSL	LTSANRR	LVEAVNH	LVLAAGA	LVLGLVA	LVVKGKV	LYLACGI	MAETAVI
MQLIAEA	MVLVVLS	NAIGVLI	NDSGETV	NEDGAVY	NEIFLTK	NGNNQIF	NIPVVSG
NLVFSPG	NNSIILP	NPEFGPA	NPKTKTY	NSRVLRS	NSTTLTY	NYLLPII	PAAAAPA
PAGTSST	PALEAGV	PAPAKPK	PAPAPAP	PAVIPLQ	PDNSAPY	PELPGEY	PEPEPEP
PGSYRLV	PIGRLLV	PKDIQLA	PKKTGGP	PLLLLDL	PLPVVSV	PNGVLRT	PPPPPPP
PRAPEAL	PRGPPPA	PRRRRQA	PSGVDAG	PSLLLLL	PSLPITV	PSPSPPP	PSSDSLL
PVPGDPD	PVRRRRR	PVSELIT	QAAAAAA	QAASGQL	QAAVTSN	QDLLQYL	QIQEMKE
QKSLNTL	QLEAIPA	QLEENLG	QLGARVG	QLKKSAD	QLYDPEK	QMLESMI	QPTAPPA
QQAAAAA	QVVSVGA	RAAAAVA	RAFAPKL	RAGALAG	RALKEQS	RAPEALS	REILLAL
RFLGFLL	RFLHMKV	RGAARRP	RGAFLVR	RIRRRRR	RKEYLER	RKSDELL	RLSSLKP
RLTASLR	RLYSGNL	RMIGEET	RPPLREQ	RQRSRKG	RRPDGSV	RRRRRRR	RRRRRRV
RRRRSRR	RSGVAEK	RSRARRA	RSSVPGV	RTFGGGT	RTLRRLL	RVLQGVL	RVPPPPP
RVRAYTY	RVVVIGA	SAAAAAA	SAFVPTN	SALDPEL	SASGLTS	SDDEDEE	SDEKLRD
SFLLLLA	SFLLSDL	SGALSRV	SGFTLDD	SGSGSDT	SGVAIAL	SGVKAIR	SHLPDDY
SILNSFV	SKAAAGR	SKLAMTI	SKLVGPS	SLVSLLF	SLLASLL	SLLLLLA	SLLLLLL
SLLLLLV	SLQSANG	SLTVWLL	SLVKRKT	SPSPPPP	SPSPSPP	SRLKYTE	SRLLLLL
SRRASGG	SSAASKI	SSAGGSF	SSCTIKV	SSDEANA	SSSSSNS	SSSSSRS	SSSSSSS
SSSTPPS	SSSTQAS	SSVPGVR	SSVSAAV	STDEPSE	STLTTPG	STNVTGD	STSRMRV
STTAKEF	SVDQSDQ	SVVRKAI	TAAAAAA	TAEGGEI	TERRRQQ	TFISRHN	TGALAAF
TGDVIGD	TGGLPEA	TIKDALG	TILAEQL	TITSAAT	TKAVAEA	TKGVVLD	TLAAALL
TLLCEAS	TLLSVLF	TLLTLGL	TLVSAVA	TLVSVGK	TLYAEPE	TPEIATR	TPTGWGL
TQAIVKN	TQPPPTS	TRRRRRR	TRVGVGV	TRVQQAT	TSLLLVL	TSNASTI	TSSSSSS
TTTYTAS	TVLPQGF	TVLSLFF	TVSGAAL	TVTAEGK	TYSVTLS	VAAACGL	VAAALAA
VAARLGE	VADVLAA	VANLDNL	VAVIGAV	VCVGSPT	VDSVSLG	VDTSGSM	VETIGVI
VETTTPS	VFCLVLL	VGAGVTR	VGDAVSK	VGDVRNG	VGPEVEA	VILLLTV	VKEAVAK
VKKPAAA	VLAVFGL	VLGLLFL	VLLLSLI	VLLVVAL	VNEALAA	VNHEAYD	VNKMTSD
VQESAAA	VSKEEAE	VSRRSRG	VSRSLTK	VTAEDVL	VTEKNVL	VTLTESG	VVAALMA
VVESTGV	VVTLIGV	VVVIGAG	WEAVSVK	WYNNNVI	YAESVKG	YEDGPNK	YESFRLT
YFSRPSS	YGDTDSL	YGLERLA	YKPGTVA	YLVGLFE	YMRNLLD	YNGTSMA	YPGSIEV
YRQMSLL							

TABLE VIII. Analogous Oligopeptides of Length 8 or More

Sequence	PIR designation	Protein and species
AAAAAAATAAA	FDFL4W	Antifreeze peptide 4 precursor—winter flounder
	P9AD37	Hexon-associated protein (IX)—adenovirus
DAAAAAAATAA	FDFL4W	Antifreeze peptide 4 precursor—winter flounder
	P9AD37	Hexon-associated protein (IX)—adenovirus
AAAAAAAAAA	WJFFEN	Specific body pattern development protein—fruit fly
	OPBYC	Cytochrome c peroxidase precursor—Baker's yeast
EEDEEGGGLF	W6WLRB	Probable E6 protein—rabbit papillomavirus
	QQBE3R	Hypothetical BVRF 2 protein—Epstein-Barr virus
AAAAAATAA	FDFL4W	Antifreeze peptide 4 precursor—winter flounder
	WJFFEN	Specific body pattern development protein—fruit fly
	P9AD37	Hexon-assocxiated protein (IX)—adenovirus
DEDEEEEEE	RDBYUC	Ubiquinol cytochrome c reductase—Baker's yeast
	HXAD2	Hexon protein—adenovirus
EDEDEDEDD	RDBYUC	Ubiquinol cytochrome c reductase—Baker's yeast
	HIBPT4	Hexon protein—adenovirus
IVAVIGAVV	PWBOB	Mitochondrial ATPase, β-chain—bovine
	TVFVUR	Kinase-related transforming protein (ras)—avian sarcoma virus
LGLVLAAGA	QRRBG	Poly-Ig receptor protein—rabbit
	QXXL6M	Mitochondrial protein (SGC1)—toad
SPSPSPPPP	QQBE13	Hypothetical BMRF1 protein—Epstein-Barr virus
	W7AD25	72K DNA-binding protein—adenovirus
AAAAAAAT	QPBYC	Cytochrome c peroxidase precursor—Baker's yeast
	FDFL4W	Antifreeze peptide 4 precursor—winter flounder
	WMAD15	Late 100K protein—adenovirus
	P9AD37	Hexon-associated protein (IX)—adenovirus
AEEAGVDL	FEDH1	Ferredoxin—*Dunaliella salina*
	FIEC3	Initiation factor IF-3—*E. coli*
AFLLLLSL	ODBY1	Cytochrome c oxidase polypeptide I—Baker's yeast
	QXBY34	Gene E protein—bacteriophage
ALLDTGAD	PBLJH2	Probable protease—T-cell leukemia virus
	GNVWH3	Pol polyprotein—AIDS virus
APAPAPAP	MORTA1	Myosin catalytic light chain—rat
	MMECA	Outer membrane protein A precursor—*E. coli*
ATSSSSSS	WJFFEN	Specific body pattern development protein—fruit fly
	PBCH	Phosvitin—chicken
AYLVGLFE	HSBO3	Histone H3—bovine
	QXASBI	Mitochondrial cob A intron protein (SGC3)—*aspergilius nidulans*
CVDTSGSM	G2GP	Ig gamma-2 chain, C region—guinea pig
	QQECO3	Hypothetical protein F-300—*E. coli*
DAAAAAAA	FDFL4W	Antifreeze peptide 4 precursor—winter flounder
	WJFFEN	Specific body pattern development protein—fruit fly
	P9AD37	Hexon-associated protein (IX)—adenovirus
DNEIFLTK	QQBE1L	Probable glycoprotein—Epstein-Barr virus
	TLPB74	Tail fiber protein—bacteriophage
EEEGAQEE	QFPGL	Neurofilament triplet protein—pig
	EDBEIC	Immediate early protein—cytomegalovirus
EGEAEEEG	QFPGL	Neurofilament triplet protein—pig
	UBURAL	Tubulin α-chain—sea urchin
EPEPEPEP	BVEC	Gene ton B protein—*E. coli*
	Q2AD2	Early protein—adenovirus
EPVPGDPD	RKZML	Ribulose bisphosphate carboxylase—maize
	VVVP1	Coat protein VP1—mouse polyomavirus
GLSVGLVG	HLHUDX	HLA class II histocompatability α-chain—human
	BDEC	Melibiose carrier protein—*E. coli*
GTRRRRRR	FFOADM2	Minor core protein—adenovirus
	QQAD82	Protein c-168—adenovirus
GVANLDNL	HBTSM	Hemoglobin β-chain—musk shrew
	TVVPT	Large T antigen—mouse polyinavirus
GVETTTPS	L2HU	Ig lambda C region—human
	FOVMLV	gag polyprotein—AIDS virus

(continued)

TABLE VIII. Analogous Oligopeptides of Length 8 or More (Continued)

Sequence	PIR designation	Protein and species
HLRELLTT	MOCHG1	Myosin G1 regulatory light chain—chicken
	QQBE32	Hypothetical BKRF2 protein—Epstein-Barr virus
IGIILLLA	CBHU	Mitochondrial cytochrome b—human
	GFHOE	Glycophorin HA—horse
IPKDIQLA	YCEC	Acetolacetate synthase—*E. coli*
	HSMS34	Histone H3(4)—mouse
IPSGVDAG	CUVM	Plastocyanin—vegetable marrow
	CYBOA	Alpha crystallin A-chain—bovine
ISFLLSDL	TVHUB	Transforming protein (Blym-1)—human
	VVVP14	Coat protein VP1—simian virus 40
IWYNNNVI	LIPG	Triacylglycerol lipase—pig
	ZEBP4L	Ea47 gene protein—bacteriophage
LSSSTQAS	HKFF7S	Heat-shock protein—fruit fly
	Z4BPFD	Gene IV protien—bacteriophage
LTGGLPEA	SYECCS	Carbamoyl phosphate synthetase—*E. coli*
	TVMSF	Transforming protein (fos)—mouse
PRAPEALS	QQBE1	Probable membrane antigen p140—Epstein-Barr virus
	QQBE34	Hypothetical BBLF4 protein—Epstein-Barr virus
RSSVPGVR	VEPG	Vimentin—pig
	MFIV1	Matrix (M1) protein—influenza B virus
TFISRHNS	TVVPT4	Large T antigen—simian virus 40
	ZGBPF4	Gene G protein—bacteriophage
TKAVAEAY	SYECGA	Glycyl-tRNA synthetase, α-chain—*E. coli*
	BYEBT	Sulfate-binding protein—*Salmonella typhimurium*
VGPEVEAA	KIBYG	Phosphoglycerate kinase—baker's yeast
	GNNYF	Genome polyprotein—foot-and-mouth disease virus
VLLLSLIG	ALMSP	Alpha amylase—mouse
	YTECTO	Tetracycline resistance protein—*E. coli*

TABLE IX. Mean z-Values of Component Oligopeptides of Unusual nmers

Length of nmers	No. of nmers	Length of components within nmer	Mean z-value of components*	Comments on nmers used
9	12	3	3.78	ALL
9	12	2	3.87	ALL
8	50	3	2.01	ALL
8	50	2	1.74	ALL
7	577	3	1.00	ALL
7	577	2	0.89	ALL
6	7,246	3	0.63	ALL
6	7,246	2	0.61	ALL
5	70,675	3	0.32	ALL
5	70,675	2	0.35	ALL
4	24	2	−1.10	$-4.0 \leqslant$ z-value $\leqslant -3.0$
4	2,563	2	0.36	$3.0 \leqslant$ z-value $\leqslant 4.0$
4	844	2	0.47	$4.0 \leqslant$ z-value $\leqslant 5.0$
4	446	2	0.62	z-value $\geqslant 5.0$
4	24	3	−1.26	$-4.0 \leqslant$ z-value $\leqslant -3.0$
4	2,563	3	0.80	$3.0 \leqslant$ z-value $\leqslant 4.0$
4	844	3	1.03	$4.0 \leqslant$ z-value $\leqslant 5.0$
4	446	3	1.17	z-value $\leqslant 5.0$
3	148	2	2.38	z-value $\geqslant 3.0$
3	82	2	−1.96	z-value $\leqslant -3.0$

*Component refers to a sequence fragment of an nmer. For example, if a 5mer were AVGCT, then it would consist of 4 dimer components, each two amino acids in length (AV, VG, GC, CT).

TABLE X. Correlation Between the z-Values of nmers and Their Oligopeptide Components

Length of nmer	Length of nmer components	Correlation coefficient*
4	2	0.17
4	3	0.35
3	2	0.50

*The correlation coefficient was calculated between the z-value of the nmer and the mean z-values of the sequence components comprising the nmer. For example, the components with length 3 of the 4mer AVGC are AVG and VGC, while the components of length 2 are AV, VG, and GC.

TABLE XI. Correlation of z-Values for Mirror Sequences

Oligomer length	No. of observations	Mean z-value*	Correlation coefficient
2	380	−0.30	0.176
3	7,600	−0.04	0.197
4	159,600	0.01	0.047

*The mean z-value is given for the oligomers and their counterparts with mirrored sequence.

repetitive oligopeptides. The average z-value of repetitive di- and tripeptides is about 3.0, and is as high as 9.27 for AA and 9.44 for RRR. Similarly, the proportion of repetitive analogous oligopeptides increases dramatically with peptide length. It would seem likely that these repetitive sequences are the product of some favored evolutionary mechanism.

Are the Corresponding Nucleotide Sequences Biased?

The nucleotide sequences were found for 16 of the 50 distinct analogous matches of octa-peptide sequences. The percentage codon identity was determined for each of these pairs and compared with that expected, taking into account local codon usage.[18] The results are shown in Table XV. The average observed fraction of identical codons is 0.36 while the expected value is 0.29. In only 7 of the 16 cases did the observed codon identity count exceed the expected by one or more. Although the sample size is small, the mean observed similarity in the nucleotide sequences is not much greater than expected. There is insufficient support for any divergence from common ancestors. The sample size was small due to the relative difficulty in gathering the data but was deemed sufficiently large to justify no further collection of data.

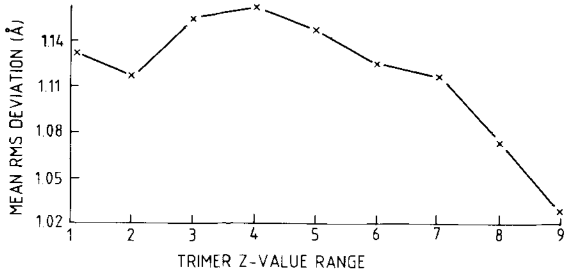

Fig. 1. Plot of the mean rms deviation (→Å) of trimer main-chain atoms versus the z-value range of the respective 3mers. The main-chain atoms for all pairwise combinations of trimers within a given z-value range were superposed by the method of Kabsch[15,16] and the mean rms deviation in →Å was taken. The ranges 1 through 9 refer, respectively, to trimer z-values (z) that are $z<-3.0$, $-3.0\leqslant z<-2.0$, $-2.0\leqslant z<-1.0$, $-1.0\leqslant z<0.0$, $0.0\leqslant z<1.0$. $1.0<z<2.0$, $2.0\leqslant z<3.0$, $3.0\leqslant z<4.0$, and $z\geqslant 4.0$.

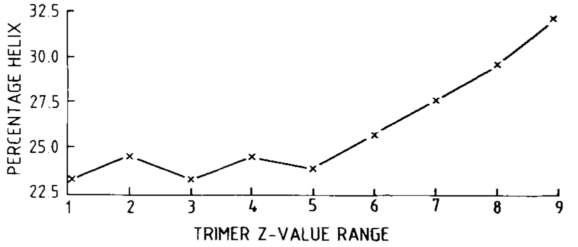

Fig. 2. Same as Figure 1 except that the vertical axis refers to the percentage of residues found in helical secondary structure.

An Aid to Determine Sequence Homology Significance

One of the difficulties in protein sequence analysis is estimating the significance of more marginal sequence homologies. This situation is almost immediately faced when a new protein sequence is compared with all those in a sequence databank by some fast algorithm such as that of Lipman and Pearson.[9] Invariably, aligned spans of 20–50 residues in length will provide possible homology candidates; the alignments often contain two or three segments three to four amino acids long that are identical. Should these putative homologies be pursued for more extensive homology using more sensitive search procedures? Even if expansion of the alignment region fails or is not pursued, are the short homologies significant? Our results suggest an approach, based upon the experimentally determined frequencies of analogous oligopeptide matches, that may provide an answer.

We can calculate the probability (P) of obtaining m analogous matches of length n when comparing two unrelated protein sequences, one with X total number

TABLE XII. Examples of Mirror Sequences

nmer	z-value	Mirror nmer	z-value
PE	7.53	EP	−5.22
SG	8.17	GS	−0.86
GK	3.57	KG	−2.52
IP	2.80	PI	−3.07
AG	1.75	GA	−3.80
EN	3.07	NE	−2.42
LK	4.24	KL	−1.05
NP	5.15	PN	−0.13
PG	0.62	GP	−4.52
MA	3.26	AM	−1.86
GI	2.05	IG	−2.85
LP	1.42	PL	−3.46
NPE	5.25	EPN	−3.75
PEE	5.81	EEP	−2.34
IPE	4.02	EPI	−3.19
DPE	4.12	EPD	−2.73
APE	1.68	EPA	−4.47
PEQ	2.40	QEP	−3.65
TPE	3.96	EPT	−2.00
PEL	3.99	LEP	−1.96
ASG	3.72	GSA	−2.23
NGK	5.47	KGN	−0.34
NPQ	2.75	QPN	−3.02
GVD	1.81	DVG	−3.93
HYLN	13.53	NLYH	0.58
WFHW	10.94	WHFW	−0.35
VCMD	11.21	DMCV	0.28
CATW	9.42	WTAC	−1.05
ECCH	9.39	HCCE	−1.01
TPEQ	7.57	QEPT	−2.33
HRHI	10.02	IHRH	0.15
MKWV	8.92	VWKM	−0.89
MRTF	8.34	FTRM	−1.36
HPYF	8.19	FYPH	−1.36
WYNR	7.99	RNYW	−1.33
CCNP	7.49	PNCC	−1.80

TABLE XIII. Mean z-Values of Repetitive nmers

Repetitive nmer length	Mean z-value	Minimum z-value	Maximum z-value
2	3.07	−1.70	9.27
3	2.92	−2.99	9.44
4	0.55	−3.30	12.25

TABLE XIV. Percent of nmers That Are Repetitive

nmer length	Percent repetitive
5	15
6	13
7	18
8	36
9	67
10	80
11	100

of amino acids and the other with Y. Assuming a binomial distribution, then

$$P_n(m) = f_n^{\,m}(1 - f_n)^{X \cdot Y - m} \cdot \frac{(X \cdot Y)!}{m! \cdot (X \cdot Y - m)!}$$

where f_n is the probability that an analogous match will occur after one sequence comparison involving two segments of length n. For the two sequences compared, there are (X · Y) such comparisons possible. The frequency f_n can be calculated from the results of Tables I and II by

$$f_n = \frac{O_n}{N_n(N_n - 1)/2}$$

where O_n is the number of observed analogous pairs with peptide length n, and N_n is the number of independent oligopeptides observed. The term in the denominator of f_n is simply the total number of oligopeptide comparisons possible. Table XVI lists the f_n values as calculated from the results of Table II.

Table XVII lists the probabilities that two sequences of equal length are related when a certain number of nmer matches are observed. The probabilities are based on observations among unrelated protein sequences. As an example of the calculation for this probability, suppose two sequences, each 400 residues long, are compared and four tetramer matches are found. The probability that at least four matches will occur by chance is simply 1 minus the sum $[P_4(0)+P_4(1)+P_4(2)+P_4(3)]$ with X,Y taken as 16×10^4 and f_4 taken from Table XVI. The probability that the two sequences are homologous is simply the sum of the four *P*-values just mentioned. Table XVII shows this probability to be 0.88. The table can also be used when comparing proteins of unequal sequence length if the sequence length product is the same as that for the two equal sequence lengths.

Table XVIII gives results for comparison of the human hemoglobin α-chain with closely and distantly related sequences and with unrelated sequences. Bacterial hemoglobin is the most distantly known relative of human hemoglobin[22]; the tetramer matches give this comparison a 79% chance for significant homology while the chicken myoglobin relationship is given a 97% chance.

A control was performed to ascertain the error rate at given probabilities for homology. A globin, cytochrome c, and immunoglobulin sequence were selected at random from the respective families; they were each compared to 1,020 unique and unrelated sequences (one representative sequence taken from

TABLE XV. Comparison of the Nucleotide Sequences for Some Analogous Octapeptides

Peptide sequence	Nucleotide sequence	Genetic locus
aaaaataa	gcagccgccgccgctactgctgcc gccgccgcagccgctactgcggcc observed identity = 0.63 expected identity = 0.29	Adenovirus 7 genome *Drosophila* engrailed locus
aaaaataa	gcagccgccgccgctactgctgcc gccgcagcagccgccaccgcagcc observed identity = 0.25 expected identity = 0.29	Adenovirus 7 genome Fish (winter flounder) antifreeze protein
aaaaataa	gccgccgcagccgctacctgcggcc gccgcagcagccgccaccgcagcc observed identity = 0.50 expected identity = 0.35	*Drosophila* engrailed locus Fish (Winter flounder) antifreeze protein
apapapap	gctccggctccagctccggcaccg gctcctgctcctgccccagccccg observed identity = 0.38 expected identity = 0.19	*E. coli* ompA gene outer membrane protein Rat fast myosin alkali light chain
atssssss	gccaccagctcgagctcctcctcg gccacctcttcttcctcatcatct observed identity = 0.25 expected identity = 0.25	*Drosophila* engrailed locus Chicken vitellogenin gene coding for phosvitin
aylvglfe	gcctacttggtagggctctttgag gcctacttagtaggattgtttgaa observed identity = 0.50 expected identity = 0.25	Human histone H3 gene *A. nidulans* mitochondrion apocytochrome b gene
epepepep	gagcccgagccagaaccggagcct gagccagaaccggaacctgagccg observed identity = 0.38 expected identity = 0.36	Adenovirus5 genome *E. coli* tonb gene coding for a membrane protein
epvpgdpd	gagcccgttcctggggacccagat gaacctgtaccgggggaccctgat observed identity = 0.38 expected identity = 0.36	*Zea mays* chloroplast large subunit of RUBP polyoma A2 virus genome
glsvglvg	gggctgtctgtgggtttggtcggt ggattgtctgtgggcctcgtgggc observed identity = 0.25 expected identity = 0.27	*E. coli* melb gene coding for melibiose carrier Human HLA-DC class II histocompatibility antigen
glvlaaga	gggttggtattagcggccggggct gggctggtgctggcagcgggggcc observed identity = 0.25 expected identity = 0.17	*Xenopus laevis* complete mitochondrial genome Rabbit mRNA for poly-immunoglobulin (IG) receptor
ipkdiqla	atcccaaaagatatccagttagcc attccaaaagatatccagttagca observed identity = 0.75 expected identity = 0.39	*E. coli* genes *ilvL*, *ilvG*, and *ilvE'* Mouse gene coding for embryonic H3 histone

(continued)

TABLE XV. Comparison of the Nucleotide Sequences for Some Analogous Octapeptides (Continued)

Peptide sequence	Nucleotide sequence	Genetic locus
isfllsdl	atttctttccttctcagtgacctg atttcctttttgttaagtgaccta observed identity = 0.38 expected identity = 0.24	Human blym-1 transforming gene Simian virus 40 (SV40) genome
lssstqas	ctctcctcgtccacgcaggccagc ttgagttcttctactcaggcaagt observed identity = 0.13 expected identity = 0.20	*Drosophila* heat shock cognate hsc70 gene Bacteriophage F1 complete genome
tfisrhns	acttttatttctcgccataattca acctttataagtaggcataacagt observed identity = 0.25 expected identity = 0.38	phiX174 complete genome SV40 genome
vgpeveaa	gtcggacccgaagttgaggctgcc gtcggtccagaagttgaagccgct observed identity = 0.38 expected identity = 0.29	Foot and mouse disease virus polyprotein *S. cerevisiae* 3-phosphoglycerate kinase gene
vlllslig	gtgctgttgttgtcattaataggc gttctgctgctttccctcattggg observed identity = 0.13 expected identity = 0.30	*E. coli* Transposon Tn 10 (tetracycline resistance) Mouse amy-2 gene fragment for alpha-amylase

TABLE XVI. Frequency of Occurrence of Analogous Matches

nmer length (n)	Frequency (fn)
2	0.0033392556479576
3	0.0001930775344508
4	0.0000120270916645
5	0.0000007927184172
6	0.0000000616777407
7	0.0000000087653035
8	0.0000000030384868
9	0.0000000016804294
10	0.0000000010827403
11	0.0000000006847494

all other families). The 4mer counts and sequence lengths were used to estimate a probability of homology, as previously discussed. For example, only 2 sequences in 1,020 were given probabilities greater than 99.9% for homology with the cytochrome c sequence. By combining similar results for all three test sequences, a list of homology probability versus error rate at that probability was determined (Table XIX).

These results are useful in comparing a new sequence with those of an entire database. After running a Lipman-Pearson search, almost invariably aligned regions of 20–50 residues are found with two or three segments of four or five amino acids that are consecutively identical. Looking up the appropriate probabilities and error rates in Tables XVII and XIX should indicate if the possible homology should be pursued by use of more refined and sensitive procedures for possible extension. It is suggested that, when probabilities of homology are greater than 0.75, further study of the two sequences is warranted; at 0.75, the error rate is about 1 in 10 (Table XIX).

Waterman and Karlin and their colleagues[29,30] have used more sophisticated statistics to examine the expected distributions of short repeats in sequences where neighboring and overlapping runs are not independent. However, for the quick look-up judgments discussed here, the assumption of the simpler binominal distribution should be adequate.

Identifying Protein Coding Regions in Nucleic Acid Sequences

A number of methods exist for predicting protein coding regions within nucleic acid sequences.[23–27] These methods are all based upon the unequal use of codons in protein coding regions. Staden[23] has demonstrated that successful predictions may be made using codon frequencies calculated from the Dayhoff et al.[21] average amino acid composition. We would like to suggest a refinement of this method which uses the observed frequencies of oligopeptides in protein sequences to calculate the codon frequencies.

TABLE XVII. Probability (×10000) That Oligopeptide Matches Between Two Protein Sequences Are Homologous

Peptide length	No. of matches	Length of both protein sequences									
		100	200	300	400	500	600	700	800	900	1000
4	1	8930	6270	3461	1502	0513	0138	0029	0005	0001	0000
	2	9941	9197	7134	4350	2036	0727	0198	0042	0007	0001
	3	9998	9880	9082	7049	4298	1990	0694	0182	0036	0005
	4	10000	9986	9770	8755	6538	3796	1658	0539	0131	0024
	5		9999	9953	9563	8201	5730	3066	1222	0360	0078
	6		10000	9992	9869	9190	7389	4712	2265	0803	0209
	7			9999	9966	9679	8574	6315	3593	1518	0470
	8			10000	9992	9887	9299	7653	5043	2507	0915
	9				9998	9964	9688	8630	6427	3703	1580
	10				10000	9990	9873	9265	7602	4989	2463
	11					9997	9953	9635	8500	6234	3520
	12					9999	9984	9832	9123	7328	4667
	13					10000	9995	9928	9520	8211	5811
	14						9998	9971	9753	8869	6863
	15						10000	9989	9881	9323	7761
	16							9996	9945	9616	8477
	17							9999	9976	9793	9011
	18							10000	9990	9894	9387
	19								9996	9948	9637
	20								9999	9976	9794
	21								10000	9989	9888
	22									9995	9942
	23									9998	9971
	24									9999	9986
	25									10000	9993
	26										9997
	27										9999
	28										9999
	29										10000
5	1	9927	9700	9329	8831	8228	7546	6811	6051	5292	4555
	2	10000	9995	9977	9929	9833	9671	9427	9091	8660	8137
	3			9999	9997	9989	9970	9929	9854	9731	9545
	4				10000	9999	9998	9993	9982	9959	9914
	5						10000	9999	9998	9995	9987
	6								10000	9999	9998
	7										10000
6	1	9994	9977	9946	9904	9850	9784	9706	9618	9518	9408
	2			10000	10000	9999	9998	9996	9993	9988	9982
	3								10000	10000	10000
7	1	9999	9997	9992	9986	9979	9969	9958	9945	9930	9914
	2							10000	10000	10000	10000
8	1	10000	9999	9997	9995	9993	9989	9985	9981	9976	9970

Staden[23] represents the probability of a particular nucleic acid sequence, a1b1c1a2b2c2a3b3c3...anbncn where anbncn represents three nucleotides of a codon, coding for protein in each of the three forward reading frames as

$$p1 = F(a1b1c1)\cdot F(a2b2c2)\cdot \ldots \cdot F(anbncn)$$
$$p2 = F(b1c1a2)\cdot F(b2c2a3)\cdot \ldots \cdot F(bncnan+1)$$
$$p3 = F(c1a2b2)\cdot F(c2a3b3)\cdot \ldots \cdot F(cnan+1bn+1)$$

where F(anbncn) is the frequency of codon (anbncn). These codon frequencies may either be taken from known tables for the organism and gene in question, or may be calculated using known amino acid distributions. In the latter case

$$F(a1b1c1) = \frac{f(a1b1c1)}{n(a1b1c1)}$$

TABLE XVIII. Probabilities of Sequence Pairs' Being Homologous Based on the Number of Matching Tetrapeptides Between the Sequences

No. of matches	Probability to be homologous	Sequence pairs*	Sequence length
12	1.00	HAHU–human alpha-hemoglobin	141
		HANE–newt alpha-hemoglobin	142
2	0.97	HAHU–human alpha-hemoglobin	141
		MYCH–chicken myoglobin	153
1	0.79	HAHU–human alpha-hemoglobin	141
		GGZLB–bacterial hemoglobin	145
1	0.46	HAHU–human alpha-hemoglobin	141
		IQECDA–*E. coli* DNA A protein	467
1	0.23	HAHU–human alpha-hemoglobin	141
		RNBP17–RNA polymerase	883

*The PIR database protein sequence identifier is given first.

TABLE XIX. Rate of Error in Predicting Homology Between Unlike Sequences*

Cutoff score for calculated probability	Percent of false (over) prediction
0.250	25.603
0.500	21.837
0.600	18.897
0.700	14.899
0.800	10.737
0.850	8.325
0.875	6.905
0.900	5.319
0.910	4.823
0.920	4.394
0.930	4.063
0.940	3.931
0.950	3.271
0.960	3.006
0.970	2.180
0.980	1.454
0.990	0.826
0.995	0.562
0.997	0.429
0.998	0.330
0.999	0.231
1.000	0.033

*These values were obtained by comparing a cytochrome c, a hemoglobin, and an immunoglobulin sequence with 1,020 different unrelated sequences. For each sequence pair a probability score was calculated based upon the no. of identical tetrapeptides in the sequence pair.

where f(a1b1c1) is the frequency of occurrence of the amino acid coded for by codon (a1b1c1) and n(a1b1c1) is the number of codons which translate to the amino acid.

We have used a more accurate calculation for the codon frequencies by taking account of the observed biases in protein oligopeptides; i.e., instead of simply using the Dayhoff value for the frequency of occurrence of an amino acid, we use the frequency with which the amino acid is observed to follow the preceding dipeptide. Thus,

$$\mathrm{F(anbncn)} = \frac{\mathrm{f(an-2bn-2cn-2,\ an-1bn-1cn-1,\ anbncn)}}{\mathrm{f(an-2bn-2cn-2,\ an-1bn-1cn-1)}} \cdot \frac{1}{\mathrm{n(anbncn)}}$$

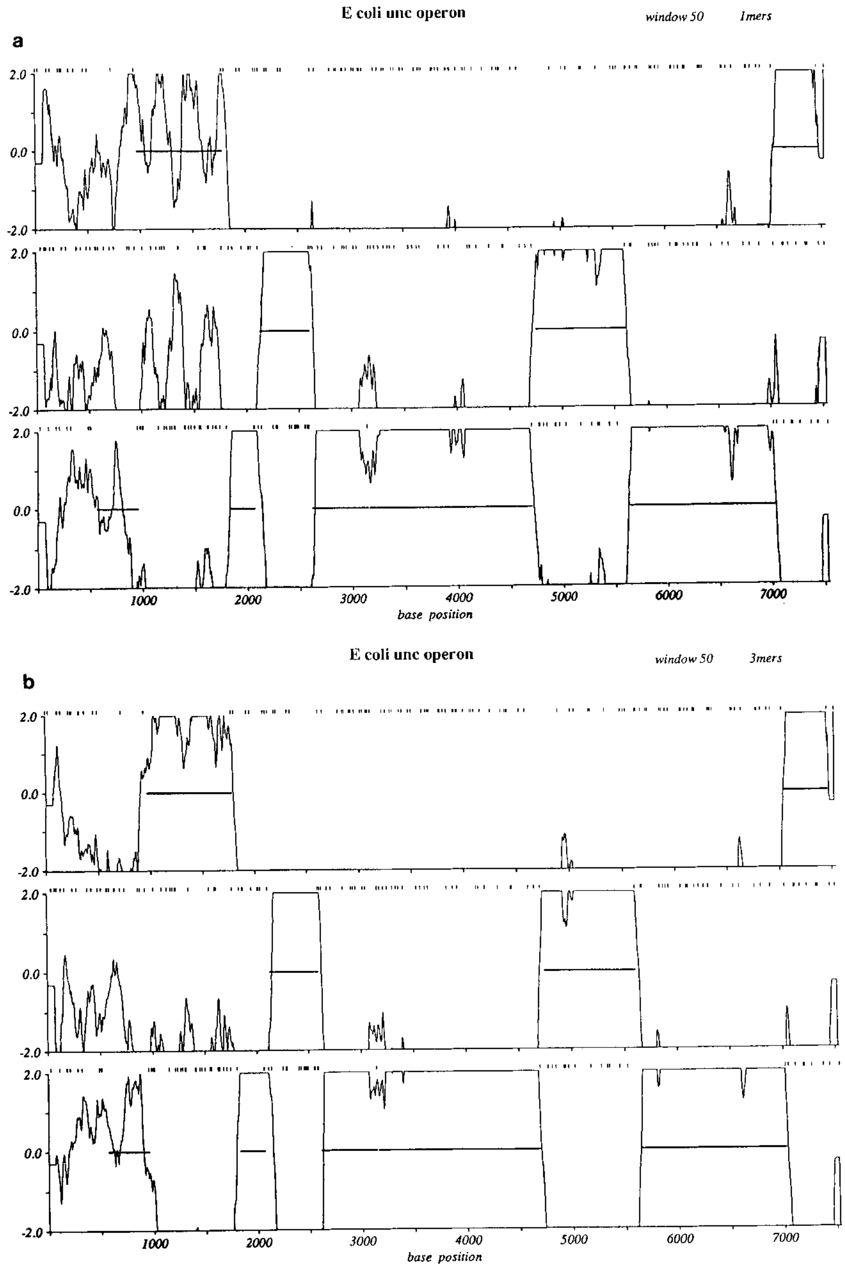

Fig. 3. **a:** Plot of the nucleotide base position versus ln(P/1 − P), indicating a possible coding region in the *E. coli unc* operon.[28] The expected codon frequencies used depends only on the amino acid composition of proteins generally. The black bars indicate known protein regions. The three plots given correspond to the three possible reading frames. STOP codons are indicated as small vertical bars above the plot. The plots are slightly smoothed. **b:** Same as panel a, except the expected codon frequency is based on the amino acid trimer preferences. It is clear that the coding frame predictions for the two leftmost coding regions have improved. The background noise has also been reduced.

where the numerator is the observed frequency of occurrence of a particular trimer coded for by the nucleotides (an-2 to cn), while the denominator is the product of the number of codons for the amino acid coded by (anbncn) and the observed frequency of occurrence of the specific dimer coded for by the (an-2 to cn-1) nucleotides. The frequency is simply the probability that the third amino acid in the trimer will follow the first two. Then, following Staden,[23] we calculate a fractional probability for coding in each of the three frames:

$$P(1) = \mathrm{p1}/(\mathrm{p1} + \mathrm{p2} + \mathrm{p3})$$
$$P(2) = \mathrm{p2}/(\mathrm{p1} + \mathrm{p2} + \mathrm{p3})$$
$$P(3) = \mathrm{p3}/(\mathrm{p1} + \mathrm{p2} + \mathrm{p3})$$

The probability scores were calculated for a window of a given length, and the window moved along the sequence in increments of three bases, maintaining the reading frame. ln(P/(1 − P)) is then plotted against the central nucleotide sequence position where the ln function exaggerates the probability extremes for easy visualization. We found that window lengths of 25 to 100 yield reasonable results, with 50 being generally preferred.

The procedure for finding reading frames can best be understood by considering the following example. In a nucleotide sequence under consideration, the amino acids coded in each of the three reading frames are listed; STOP codons are also annotated. For a window length of 50, the frequencies F(anbncn) associated with the first 50 trimers (moving one amino acid forward at a time) are multiplied to yield pl; similarly for the other reading frames, p2 and p3 are determined. If any STOP codon is encountered in a trimer, the frequency for the codon is taken as 0.0178 which is the mean codon frequency based on the Dayhoff et al.[21] composition of amino acids in proteins generally. Then ln(P/1 − P) is determined for each reading frame, and the values are plotted at the central nucleotide position. The window is then moved one amino acid forward and the process repeated for the 2nd to 51st trimers in each reading frame.

Eleven trials were run using genes from mammalian, plant, viral, and bacterial sources; the examples included whole viral genomes and overlapping reading frames. Calculations were made based on the trimer preferences and the single amino acid frequencies determined by Dayhoff et al.[21] and used by Staden.[23] The results for the *Escherichia coli unc* operon are shown in Figure 3. It is clear that the trimer-based predictions are better, especially for the two early coded proteins in reading frames 1 and 3 (leftmost position of the Fig. 3 plot); the noise has also been reduced with the trimer data. Of the 11 trials, six showed some improvement with the 3mer preferences while five others were about the same for the 3mer and 1mer data. A further control was performed. Each protein sequence and any homologous ones associated with the coding region to be predicted were removed from the database for three of the above examples. The trimer preferences were recalculated and used to predict the coding frame. These probability plots were almost indistinguishable from those with trimer frequencies determined over the entire database. When the amino acid sequence database is sufficiently large to provide good statistics for tetramers, it is expected that their use will further improve the reading frame results.

We also calculated protein coding probabilities based on the single amino acid frequencies taken from our superfamily grouping. The results for the 11 trials showed improvements over those based on the Dayhoff amino acid frequencies and approached the probability profiles generated from the 3mer preferences.

CONCLUSIONS

The results presented here demonstrate that similarities exist between protein sequences with no obvious sequence or evolutionary relationship. These similarities take the form of preferences in the occurrence of oligopeptide sequences. In particular, repetitive sequences are highly favored. The strongly preferred oligopeptides tend to represent conserved structures. The longer preferred peptides are generally composed from the shorter preferred peptides. The order in which the amino acids appear in the oligopeptide is also important. These sequence prejudices are useful aids in determining the significance of sequence homology and the protein coding regions of nucleotide sequences. The careful grouping of related sequences in the database allowed accurate calculation of the overall amino acid composition of proteins.

ACKNOWLEDGMENTS

The authors wish to thank Chris Sander and Toby Gibson for many useful discussions. Christine Raulfs was essential in the preparation of this manuscript.

REFERENCES

1. Dayhoff, M.O., Barker, W.C., Hunt, L.T., Schwartz, R.M. Protein superfamilies. In: "Atlas of Protein Sequence and Structure." Vol. 5, Suppl. 3. Washington DC: National Biomedical Research Foundation 1978: 9–24.
2. Zuckerkandl, E. On the molecular evolutionary clock. J. Mol. Evol. 26:34–46, 1987.
3. Doolittle, R.F. Similar amino acid sequences: chance or common ancestry? Science 214:149–159, 1981.
4. Saroff, H.A. The uniqueness of protein sequences: Uniqueness diagrams for the Dayhoff file. Bull. Math. Biol. 46:661–671, 1984.
5. Vonderviszt, F., Matral, G., and Simon, T. Characteristic sequential residue environment of amino acids in proteins. Int. J. Pept. Protein. Res. 27:483–492, 1986.
6. Klapper, M.H. The independent distribution of amino acids near neighbor pairs into polypeptides. Biochem. Biophys. Res. Commun. 78:1018–1024, 1977.
7. Wilson, I.A., Haft, D.H., Getzoff, E.D., Teiner, J.A., Lerner, R.A., Brenner, S. Identical short peptide sequences in unrelated proteins can have different conformations: a testing ground for theories of immune recognition. Proc. Natl. Acad. Sci. U.S.A. 82:5255–5259, 1985.
8. Barker, W.C., Hunt, L.T., George, D.G., Yeh, L.S., Chen, H.R., Blomquist M.C., Seibel-Ross, E.T., Elzanowski, A., Hong, M.K., Ferrick, D..A., Blair, J.K., Chen, S.L., Ledley, R.S. "Protein Sequence Database," National Biomedical Research Foundation. Washington DC: Georgetown University Medical Center.
9. Lipman, D.J., Pearson, W.R. Rapid and sensitive protein similarity searches. Science 227:1435–1441, 1985.
10. Argos, P. A sensitive procedure to compare amino acid sequences. J. Mol. Biol. 193:385–396, 1987.
11. Devereux J., Haeberli P., Marquess, P. A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids. Res. 12:387–395, 1984.
12. Wilbur, W.J., Lipman, D.J., Rapid similarity searches of nucleic acid and protein data banks. Proc. Natl. Acad. Sci. U.S.A. 80:726–730, 1983.
13. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T.,

Tasumi, M. The Protein Data Bank: A computer-based archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.

14. Kabsch, W., Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637, 1983.
15. Kabsch, W. A solution for the best rotation to relate two sets of vectors. Acta Cryst A32:922–923, 1976.
16. Kabsch, W. A discussion of the solution for the best rotation to relate two sets of vectors. Acta Cryst A34:827–828, 1978.
17. Hamm, G.H., Cameron, G.N. The EMBL data library. Nucleic Acids Res. 14:5–9, 1986.
18. Maruyama, T., Gojobori, T., Aota, S.-I., Ikemura, T. Codon usage tabulated from the GenBank genetic sequences data. Nucleic Acids Res. 14:r151–197, 1986.
19. Kernighan, B.W., Ritchie, D.M. "The C Programming Language." Englewood Cliffs, NJ: Prentice Hall Inc. 1978.
20. Argos, P., Kamer, G., Nicklin, M., Wimmer, E. Similarity in gene organization and homology between proteins of animal picornaviruses and a plant comovirus suggest common ancestry of these virus families. Nucleic Acids Res. 12:7251–7268, 1984.
21. Dayhoff, M.O., Schwartz, R.M., Oscutt, B.C. A model of evolutionary change in proteins. In: "Atlas of Protein Sequence and Structure," Vol. 5, Suppl. 3. Washington, DC: National Biomedical Research Foundation, 1978:345–358.
22. Bashford, D., Chothia, C., Lesk, A.M. Determination of a protein fold: Unique features of the globin amino acid sequences. J. Mol. Biol. 196:199–216, 1987.
23. Staden, R. Measurements of the effects that coding for a protein has on a DNA sequence and their use for finding genes. Nucleic Acids Res. 12:551–567, 1984.
24. Shepherd, J.C.W. Method to determine the reading frame of a protein from the purine pyrimidine genome sequence and its possible evolutionary justification. Proc. Natl. Acad. Sci. U.S.A. 78:1596–1600, 1981.
25. Staden, R., McLachlan, A.D. Codon preference and its use in identifying protein coding regions in long DNA sequences. Nucleic Acids Res. 10:141–156, 1982.
26. Fickett, J.W. Recognition of protein coding regions in DNA sequences. Nucleic Acids Res. 10:5303–5318, 1982.
27. Gribskov, M., Devereux, J., Burgess, R. The codon preference plot: Graphic analysis of protein coding sequences and prediction of gene expression. Nucleic Acids Res. 12:539–549, 1984.
28. Saraste, M., Gay, N.J., Eberle, A., Runswick, M.J., Walker, J.E. The atp operon: Nucleotide sequence of the genes for the γ, β, and ϵ subunits of Escherichia coli ATP synthase. Nucleic Acids Res. 9:5287–5296, 1981.
29. Karlin, S., Morris, M., Ghandour, G., Leung, M.Y. Algorithms for identifying local molecular sequence features. Comput. Applications Biosci. (CABIOS) 4:41–51, 1988.
30. Perlwitz, M.D., Burks, C., Waterman, M.S. Pattern-analysis of the genetic code. Adv. Appl. Math. 9:7–21, 1988.