85

J. Mol. Biol. (1993) 229, 1065–1082

Empirical and Structural Models for Insertions and Deletions in the Divergent Evolution of Proteins

Steven A. Benner, Mark A. Cohen

Institute for Organic Chemistry
Swiss Federal Institute of Technology, CH-8092 Zurich, Switzerland

and Gaston H. Gonnet

Institute for Scientific Computation
Swiss Federal Institute of Technology, CH-8092 Zurich, Switzerland

(Received 27 July 1992; accepted 27 October 1992)

The exhaustive matching of the protein sequence database makes possible a broadly based study of insertions and deletions (indels) during divergent evolution. In this study, the probability of a gap in an alignment of a pair of homologous protein sequences was found to increase with the evolutionary distance measured in PAM units (number of accepted point mutations per 100 amino acid residues). A relationship between the average number of amino acid residues between indels and evolutionary distance suggests that a unit 30 to 40 amino acid residues in length remains, on average, undisrupted by indels during divergent evolution. Further, the probability of a gap was found to be inversely proportional to gap length raised to the 1·7 power. This empirical law fits closely over the entire range of gap lengths examined. Gap length distribution is largely independent of evolutionary distance. These results rule out the widely used linear gap penalty as a satisfactory formula for scoring gaps when constructing alignments. Further, the observed gap length distribution can be explained by a simple model of selective pressures governing the acceptance of indels during divergent evolution. Finally, this model provides theoretical support for using indels as part of "parsing algorithms", important in the *de novo* prediction of the folded structure of proteins from the sequence data.

Keywords: protein structure; evolution; insertions/deletions; protein structure prediction

1. Introduction

Alignments of homologous protein sequences are among the most important tools in the analysis of protein structure (Edwards & Cavalli-Sforza, 1963: Zuckerkandl & Pauling, 1965; Fitch & Margoliash. 1967; Doolittle, 1990). Sequence alignments are the starting point for all successful methods for predicting *de novo* the folded structure of proteins (Crawford *et al.*, 1987; Benner, 1989; Niermann & Kirschner, 1990; Bazan, 1990; Benner & Gerloff. 1991); the remarkable accuracy of three predictions made using these methods (Hyde *et al.*, 1988: Knighton *et al.*, 1991; de Vos *et al.*, 1992) have opened a new generation of protein structure prediction efforts (Thornton *et al.*, 1991; Benner, 1992). Alignments are also the starting point for knowledge-based structural models of proteins (Blundell *et al.*, 1987), and are used to estimate the number of different types of protein folds (Taylor, 1990; Dorit *et al.*, 1990: Doolittle, 1991; Gonnet *et al.*, 1992), to interpret data from various genome sequencing projects (Sulston *et al.*, 1992; Oliver *et al.*, 1992), and to resolve evolutionary issues from the origin of man to the origin of life (Benner *et al.*, 1989).

Protein sequence alignments can be constructed using the dynamic programming algorithm of Needleman and Wunsch (Needleman & Wunsch, 1970; Sellars, 1974; Sankoff & Kruskal, 1983; Arratia *et al.*, 1986). When used with scoring matrices derived as proposed by Dayhoff *et al.* (1978), this algorithm provides the alignment of two sequences that maximizes the probability that the two have evolved from a common ancestor, as opposed to their having arisen independently (the null hypothesis). The algorithm is therefore a "maximum likelihood estimator". As such indicators are unbiased (Freund, 1971), estimates of probabilities can be assigned to alignments constructed with this algorithm and used to evalu-

0022-2836/93/041065-18 $08.00/0

ate the evolutionary relation between aligned protein sequences.

Alignment procedures were further advanced by the pioneering work of Dayhoff *et al.* (1978). This work produced empirical Dayhoff matrices indicating the relative probability of each of 210 possible matches and mismatches between the 20 standard amino acids. These matrices provide the empirical grounds for scoring matches and mismatches in a protein alignment. Feng *et al.* (1985) have found that the quality of a Needleman–Wunsch alignment can be improved with the help of a Dayhoff matrix, especially when aligning distantly homologous sequences.

When an alignment contains gaps, the parameters needed as inputs by a dynamic programming algorithm before constructing alignments have remained elusive, however (Fitch & Smith, 1983; Feng & Doolittle, 1987; Altschul, 1989; Demchuk *et al.*, 1989; Thorne *et al.*, 1992; Pascarella & Argos, 1992). Gaps result from insertion and deletion events (indels) during divergent evolution. They are traditionally scored using arbitrary numerical recipes. The most common of these involves assigning a gap penalty of the form $(ak+b)$, where k is the length of the gap, and a and b are arbitrarily chosen parameters. There has been no theory, however, to assist the selection of the parameters to be used in such numerical recipes, to estimate the confidence that should be placed in alignments derived using these recipes, or even to suggest that such recipes offer a valid approach for scoring alignments that contain gaps (Thorne *et al.*, 1992).

We recently reported the organization of the entire protein sequence database using a "patricia tree" data structure (Gonnet *et al.*, 1992). In addition to allowing rapid retrieval of the sequences of homologous proteins, the organization makes possible an exhaustive matching of the sequence database. Exhaustive matching is defined as the product of an attempted Needleman–Wunsch alignment of every subsequence in the database with every other subsequence. The 1·7 million pairs of matched sequences obtained in the exhausting matching provided a basis for systematic investigation of divergent evolution at the level of protein sequences. One conclusion of this investigation was that the Dayhoff matrices widely used to score alignments (Dayhoff *et al.*, 1978) are not optimal, especially for protein pairs separated by large evolutionary distance (Gonnet *et al.*, 1992). Another paper describing the preparation of Dayhoff mutation matrices was published recently (Jones *et al.*, 1992).

Exhaustive matching also provides the resources needed to construct an empirically grounded model of indels during divergent evolution. This is the topic of this paper. The primary data reported here concern the frequency of indels as a function of evolutionary distance, the relationship between the length of a gap and its frequency of occurrence, and the types of amino acid residue in the regions flanking the gap and within the insert, drawn from data that include the entire protein sequence database.

From these data, mathematical models describing the probability of an indel as a function of various parameters are constructed, together with estimates of these parameters and approximate values for the length distribution of gaps in an alignment. The most successful model restores an accurate notion to the similarity score describing an alignment that includes gaps. These scores are normally expressed as logarithms of a conditional probability (multiplied by 10). Further, a structural model based on an underlying view of protein folding is developed to account for the successful mathematical model.

The most significant aspects of this model are:

(1) The probability of a gap in an alignment of two protein sequences is a function of the evolutionary distance between the sequences. A linear relationship is observed when the average number of amino acid residues per indel is plotted against the average number of amino acid residues per mutation, proportional to the reciprocal of evolutionary distance measured in PAM† units (accepted point mutations per 100 amino acid residues). An extrapolation to infinite evolutionary distance suggests that a polypeptide segment averaging 30 to 40 amino acid residues in length remains undisrupted by indels. This length corresponds to an often presumed size of a folding unit in peptide chemistry (Wetlaufer, 1981; Thomas & Luisi, 1986; Patthy, 1991).

(2) The distribution of gap size is essentially independent of the evolutionary distance between two sequences, with only a modest decrease in average gap length at increasing PAM distance.

(3) With remarkable precision, the distribution of gap length follows a generalized Zipfian distribution (Gonnet & Baeza-Yates, 1991), where the probability distribution of gap length is inversely proportional to gap length raised to the 1·7 power

(4) These results exclude an exponential distribution of gap lengths, and the corresponding $(ak+b)$ recipe used in most sequence alignment programs for scoring gaps.

(5) The Zipfian gap length distribution is consistent with the hypotheses that an indel is accepted by natural selection when its ends lie near in space, that the insert adopts a random coil conformation, and that the behaviour of a randomly coiled component of a folded polypeptide is governed by laws governing the statistical mechanics of isolated randomly coiled polymers (Flory, 1953).

(6) This analysis adds theoretical support to the intuitive rule, widely suspected for some time, that indels occur between standard secondary structural elements (alpha helices and beta strands) of folded proteins. The use of gaps as parsing elements in the *de novo* prediction of the conformation of proteins

† Abbreviations used: PAM distance: accepted point mutations per 100 amino acid residues; r.m.s., root-mean-square.

from sequence data has been described in length elsewhere (Crawford *et al.*, 1987; Benner, 1989; Benner & Gerloff, 1991).

2. Methods

(a) *The exhaustive matching*

An exhaustive matching of the protein sequence database was recently completed in these laboratories (Gonnet & Benner. 1991; Gonnet *et al.*. 1992). This corresponds in result (but not in method) to attempting a dynamic programming (Needleman & Wunsch, 1970) matching between every subsequence and every other subsequence within the database. Results of the matching, obtained using the MIPS Version 64 database. have been confirmed by a second exhaustive matching of Version 19 of the Swiss-Prot database (Bairoch & Boeckmann. 1991, 1992). As the first database contains approx. 8·4 million subsequences (n), an exhaustive matching is equivalent to approx. 35 trillion ($n^2/2$) attempted alignments. These, of course. could not be performed by even the fastest supercomputer given several millennia. Therefore, a search algorithm based on a "patricia tree" data structure (Gonnet & Baeza-Yates. 1991) was applied; this data structure allows a search that yields a final result identical to that which would be obtained by direct cross-matching of the entire database in far less than n^2 operations. The exhaustive matching thus implemented required only 404·5 days of central processor unit time obtained over 19 weeks in the background obtained from up to 6 work stations running in parallel (Gonnet *et al.*. 1992).

The procedure detects all significant matches within the database. regardless of where in an entry the matched sequence might lie. Significant matches of sequences within the same entry (as would be produced. for example. by internal repeats within the same sequence). significant alignments of partial sequences. and significant alignments between different parts of a single protein and segments of 2 or more different proteins (as might be produced. for example. by domain shuffling) are all found by this approach.

The matrix from Dayhoff *et al.* (1978) and standard gap penalties were used in the first phase of the exhaustive matching. A liberal target score ensured that every match with potentially significant sequence similarity was examined. The initial matching yelded 6·4 million matches with an aligned similarity score of 80 or better. These were then refined by running the dynamic programming algorithm from the point where each match began in one direction along the sequence alignment to the point where the alignment was optimized (or the sequences exhausted). running in the reverse direction to achieve the same goal. and repeating the process until the alignment score was no longer improved. After refining, 1·7 million matches remained. each optimally aligned. These matches were then used to calculate new Dayhoff matrices (Gonnet *et al.*. 1992). which then provided new scoring parameters used to refine further the matches to self-consistency.

The most probable evolutionary distance between each pair of matched sequences was then computed. Evolutionary distance was measured in PAM units. indicating the number of accepted point mutations per 100 amino acid residues separating the 2 sequences. Thus, 2 protein sequences 1 PAM unit distant differ by 1 accepted point mutation per 100 amino acid residues. The matrix describing the probability of pairwise matches between the 20 amino acid residues in this alignment is referred to as the "1% mutation matrix"; the sum of the off-diagonal terms of this matrix is 1%. For 2 sequences separated by a PAM distance of x, the highest probability of obtaining the second sequence from the first occurs after x transformations of the first by the 1% matrix.

(b) *A database for building an empirical model for deletions*

To be useful in modelling deletions and insertions during divergent evolution, matches must meet the following criteria.

(1) The gaps being analysed should reflect *bona fide* indels during divergent evolution, not recording or experimental errors. Further, identical sequences represented more than once in the database should not be compared. Statistics were therefore compiled from protein pairs at least 4·7 PAM units distant. This avoids counting duplicates within the database, as well as most of the cases where gaps arise from errors in the entry of closely related sequences. A sample of the data was examined by hand to ascertain that recording errors were not likely to influence significantly the empirical model derived from the remaining matches†.

(2) The sequences being compared must be indisputably homologous and the alignment relating them of high quality. This is necessary so that the gaps being counted can be reliably attributed to real insertion and deletion events during the evolutionary history of the 2 proteins and not to artifacts created by poor or fictitious alignments. To this end, the alignments used in this study were between sequence pairs less than 100 PAM units distant that achieved a similarity score greater than 150 and extended for more than 80 residues. These criteria reduced the total number of matches to 411,000. These criteria maintain a useful sample size, but are more than adequately conservative to guarantee that gaps scored lie within significant alignments.

(3) It is interesting to learn how the frequency and length of gaps between protein pairs depends on the evolutionary distance between those pairs. Therefore, data must be collected separately for protein pairs at different PAM distances, PAM windows (illustrated schematically in Fig. 1) were therefore defined by an upper PAM bound (p in Fig. 1, defining a "connected component") and a lower PAM bound (q in Fig. 1). Specific values for these PAM windows are collected in Table 1.

(4) Within each PAM window, insertion and deletion events occurring during divergent evolution should be counted only once. Therefore, in comparing 2 connected components joined by a bridge between PAM limit q and PAM limit p (Fig. 1), in cases where the tree has subbranches below q, only a single pair of sequences from

† Inspection of the data shows that some long deletions are due to events that are not properly modelled by the assumptions used here. Most trivially. this includes recording features of the database. e.g. single entries that contain multiple fragments that. when paired against complete sequences, yield gaps. Also. at low PAM distances, a significant number of long deletions appear to result formally from simultaneous deletion replacement events. These require treatment by a more sophisticated model not discussed here. To prevent these from having an impact on the interpretation presented here, eqn (11) was derived from sequence pairs at distances greater than 10 PAM units and gap lengths shorter than 60 amino acid residues.

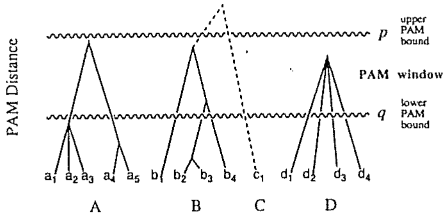
PAM Distance
upper PAM bound
p
PAM window
lower PAM bound
q
a_1 a_2 a_3 a_4 a_5 b_1 b_2 b_3 b_4 c_1 d_1 d_2 d_3 d_4
A
B
C
D

Figure 1. Sampling matches to avoid redundant counting of indels. The 1·7 million matched sequence alignments allow the division of the sequences in the database into connected components defined by an upper PAM bound, indicated by the upper line (p) in this diagram. Data are tabulated in sets drawn from pairwise matches in different PAM windows defined by an upper bound and a lower bound (the lower line q). A, B, C (a single sequence) and D. These sequences may, of course, be connected by matches at higher PAM distance (the broken line connecting B and C). However, because all possible matches within the database have been recorded during the exhaustive matching, it is guaranteed that any match connecting any of the 4 components will occur at a PAM distance higher than p. For the PAM window defined by p and q, only one pairwise alignment within the sequences defined by A (e.g. between sequence a_1 and sequence a_4) will be tabulated; tabulating an alignment between sequence a_1 and sequence a_5 as well would be redundant.

each sub-branch is compared. The set of protein pairs for each PAM window (Table 1) contained a reasonable sample size of approx. 250,000 aligned positions. The sets are available in electronic form to interested individuals.

3. Results

Gap length and frequency data for the entire protein sequence database were collected in different PAM sets. A summary of the data collected for each group is compiled in Table 1. These data were used first to develop descriptive mathematical functions and then structural models that best account for insertion and deletion events during divergent evolution.

(a) *Gap length distribution*

Table 2 collects data concerning the number of gaps of specified lengths in aligned pairs of protein sequences obtained at different PAM values. These data were examined to identify descriptive mathematical functions, beginning with exponential functions. Exponential functions are widely presumed to describe gap length distribution in aligned protein sequences. In particular, the linear penalty function $(ak+b)$, used as a standard default in most alignment programs, presumes an exponential distribution of gap lengths.

According to an exponential distribution, the probability of a gap of length k is:

$$p_k = \alpha \frac{(1-\beta)\beta^k}{\beta}, \qquad (1)$$

where α is the probability of an insertion or deletion event, and β is the probability of removing an additional amino acid residue once a gap is formed. The logarithm of the probability, the value used in a dynamic programming alignment, is given by the equation:

$$\log p_k = \log(\alpha(1-\beta)) + (k-1)\log\beta. \qquad (2)$$

For example, a gap penalty parameter of 12 and the increment cost parameter of 4 in a typical alignment program correspond to a value of 0·1048 for α, and a value of 0·3981 for β.

Inspection of the data in Table 2 shows that an exponential description of the gap length distribution does not provide an adequate fit to the data. Three separate observations illustrate this. First, for an exponential distribution, the ratio of successive probabilities is a constant:

$$p_1/p_2 = p_2/p_3 = p_3/p_4 = \ldots p_i/p_{i+1} = 1/\beta. \quad (3)$$

Similarly:

$$\begin{aligned} p_1/p_{1+x} = p_2/p_{2+x} &= p_3/p_{3+x} \\ &= \ldots p_i/p_{i+x} = 1/\beta^x. \end{aligned} \quad (4)$$

If the gap length probabilities follow an exponential distribution, then the observed frequencies (N_k) should also follow these relationships. They clearly do not. For example, for the set of data collected at a PAM window of 29·5 to 40, we can compute from Table 2:

$$\begin{aligned} &N_1/N_2 = 2{\cdot}60 \quad N_3/N_4 = 1{\cdot}47 \quad N_5/N_6 = 1{\cdot}12 \\ &N_1/N_5 = 2{\cdot}60 \quad N_6/N_{10} = 2{\cdot}66 \quad N_{11}/N_{15} = 1{\cdot}43. \end{aligned} \quad (5)$$

The values in neither row are constant as required by an exponential distribution.

A second more powerful demonstration of the inadequacy of an exponential fit to account for the gap length distribution is provided by an analysis of the tail of the distribution. A significant number of gaps of length 61 and longer is observed in Table 2 in all PAM windows. For example, in the window defined by PAM 29·5 and PAM 40 as lower and upper bounds (using data from the MIPS database), the conditional probability of having a gap of length 61 or longer is:

$$\frac{1}{\alpha}\sum_{k=1}^{\infty} p_k = \beta^{60} \approx \frac{11}{2063}. \qquad (6)$$

β must be approximately unity (or, more precisely, 0·916) to account for these data. This is inconsistent with the β required to account for the length distribution for shorter gaps. For example, with gaps of length one and two, which account for more than half of the total number of gaps, $N_2/N_1 = \beta = 0{\cdot}393$, again using data from the window defined by PAM 29·5 and PAM 40 as lower and upper bounds (Table 2). In other words, the tail of the distribution is too long and its initial decline too steep for both to be accounted for by a single exponential distribution.

A third illustration comes from estimates for the parameters in a hypothetical exponential distribu-

Table 1

Indels in aligned pairs of homologous sequences at various PAM distances

PAM range start	PAM range end	Mid PAM range	Matches read	Connected components used	Number of aligned positions	Number of of indels	Number of deleted amino acids	Average gap length	Positions per indel	Indels per position	Positions between mutation at mid PAM
A. *MIPS version 64*											
4·7	6·4	5·55	25,295	725	206,210	423	3242	7·7 ± 2·3	487·5	0·00205	18·02
6·4	8·7	7·55	32,756	759	236,350	605	4237	7·0 ± 1·1	390·7	0·00256	13·25
8·7	11·8	10·25	42,120	734	228.323	730	4682	6·4 ± 0·8	312·8	0·00320	9·76
11·8	16·0	13·9	55,730	695	215,104	749	3954	5·9 ± 2·0	287·2	0·00348	7·19
16·0	21·7	18·85	77,408	753	252,761	1224	6752	5·6 ± 0·5	206·5	0·00484	5·31
21·7	29·5	25·6	105,933	723	254,187	1522	7715	5·4 ± 0·7	167·0	0·00599	3·91
29·5	40·0	34·75	146,609	755	261,476	2063	10,263	5·4 ± 0·5	126·7	0·00789	2·88
40·0	54·3	47·15	219,384	749	264,888	2704	11,776	4·6 ± 0·3	98·0	0·01020	2·12
54·3	73·7	64·0	333,060	740	261,463	3165	12,305	4·2 ± 0·3	82·6	0·01211	1·56
73.7	100·0	86·85	524,639	708	241,097	3165	12,936	4·2 ± 0·3	69·3	0·01443	1·15
B. *Swiss-Prot version 19*											
4·7	6·4	5·55	5042	549	176,311	298	2491	8·4 ± 2·8	591·6	0·00169	18·02
6·4	8·7	7·55	7767	645	230,017	493	4270	8·7 ± 3·1	466·6	0·00214	13·25
8·7	11·8	10·25	11,969	723	248,841	698	5082	7·3 ± 1·4	356·5	0·00281	9·76
11·8	16·0	13·9	18,862	717	235,541	825	4897	5·9 ± 2·0	285·5	0·00350	7·19
16·0	21·7	18·85	30,229	835	279,416	1308	7256	5·5 ± 0·5	213·6	0·00468	5·31
21·7	29·5	25·6	42,748	822	294,881	1733	9339	5·4 ± 0·7	170·2	0·00588	3·91
29·5	40·0	34·75	60,966	845	303,030	2254	12,173	5·4 ± 0·5	134·4	0·00744	2·88
40·0	54·3	47·15	84,326	883	322,428	3076	14,226	4·6 ± 0·2	104·7	0·00955	2·12
54·3	73·7	64·0	114,716	896	324,424	3835	16,465	4·3 ± 0·2	84·6	0·01182	1·56
73·7	100·0	86·85	180,187	790	284,997	4108	16,987	4·1 ± 0·3	69·4	0·01441	1·15

Table 2

Lengths of gaps in alignments of matched sequences at various PAM distances

	PAM window									
Low bound	4·7	6·4	8·7	11·8	16·0	21·7	29·5	40·0	54·3	73·7
Upper bound	6·4	8·7	11·8	16·0	21·7	29·5	40·0	54·3	73·7	100
Gap length										
1	182	211	269	331	450	610	796	1039	1276	1406
2	61	85	93	89	196	236	313	451	580	672
3	29	45	70	86	148	159	231	298	350	379
4	15	45	53	53	87	97	162	202	233	263
5	22	32	34	38	55	56	83	135	161	174
6	20	20	28	30	42	57	85	109	108	122
7	10	20	23	25	29	44	53	77	84	91
8	11	20	20	18	16	39	54	80	65	76
9	4	21	17	8	34	30	31	51	57	52
10	8	11	12	9	21	28	40	45	40	40
11	7	10	11	9	11	19	29	39	32	34
12	5	9	12	7	19	18	13	21	22	15
13	1	8	3	3	6	13	19	16	21	20
14	3	3	8	5	7	14	18	14	23	19
15	4	7	6	4	9	8	20	11	18	16
16	3	4	4	2	3	4	12	13	6	13
17	4	5	4	2	10	9	13	7	10	10
18	3	1	4	3	11	9	14	5	10	13
19	1	3	8	5	4	8	9	7	8	7
20	3	3	11	2	6	6	4	6	9	2
21	1	3	2	2	3	7	5	8	6	3
22	0	3	3	0	3	3	6	1	5	5
23	0	1	0	1	0	3	7	8	6	4
24	0	1	0	0	3	4	2	6	2	4
25	1	2	0	0	1	5	3	1	1	4
26	2	4	2	2	4	3	3	8	3	1
27	1	1	2	0	4	2	3	2	2	1
28	0	1	0	2	3	0	0	4	1	2
29	1	2	2	0	1	1	0	4	0	2
30	1	4	3	2	1	3	1	3	1	4
31	3	2	2	1	1	0	2	2	3	3
32	0	0	2	1	1	4	3	2	1	4
33	1	1	0	1	1	4	5	1	1	0
34	0	0	2	1	1	3	0	3	1	0
35	1	0	0	1	3	1	2	0	0	1
36	0	1	2	0	0	0	5	1	0	1
37	1	0	0	1	2	0	0	3	0	1
38	1	1	1	1	2	0	0	2	1	1
39	1	0	0	0	0	3	2	0	2	0
40	0	0	1	0	5	1	1	1	0	0
41	0	2	2	0	3	0	0	0	1	0
42	0	1	0	0	0	0	0	0	0	1
43	0	0	0	0	1	1	0	0	0	1
44	1	0	1	0	1	0	0	1	0	0
45	0	0	1	0	0	2	0	0	2	0
46	0	0	0	0	0	0	0	1	1	0
47	0	0	0	0	1	0	0	0	0	0
48	1	0	0	0	1	0	1	0	0	0
49	0	0	0	0	1	0	1	3	1	0
50	0	1	0	0	1	0	1	1	0	0
51	0	0	0	0	1	0	0	1	0	0
52	0	0	0	0	0	0	0	0	0	1
53	0	0	0	0	0	0	0	0	0	0
54	0	0	0	0	0	0	0	0	0	0
55	1	1	0	0	0	0	0	0	0	0
56	1	1	0	0	0	0	0	0	0	0
57	1	0	0	0	0	0	0	1	0	0
58	0	1	0	0	0	0	0	0	0	0
59	0	0	2	0	1	0	0	0	0	2
60	0	0	1	0	0	0	0	1	0	1
61–70	2	0	4	0	3	2	2	3	3	3
71–80	0	2	1	0	1	3	0	4	4	1
81–90	0	3	0	0	0	2	2	0	2	1
91–100	0	0	2	1	0	0	1	1	0	1
>100	3	4	2	1	0	1	6	0	2	2

Table 3

Estimates of the parameters for a hypothetical exponential fit to the gap length distribution

Gap range	$-10 \log \beta$	Q
1–20	1·313	−103·9
1–40	1·076	−159·8
1–60	1·036	−178·2
1–∞	0·896	−294·5

β is the parameter from the hypothetical linear gap penalty expressed by eqn (1) describing the ratio of probabilities of a gap of length (k) and a gap of length ($k+1$). Q is a maximum likelihood estimator of the quality of the fit of the exponential function with the specified β to the data in the specified gap range. More negative values of Q indicate a worse fit.

Table 4

Fit of a Zipfian distribution to data from aligned homologous sequences

Gap length	Number of occurrences	Approximation	Cumulative number of occurrences	Approximation
1	796	767·7	796	767·7
2	313	389·6	1109	1157·3
3	231	218·4	1340	1375·7
4	162	136·2	1502	1511·9
5	83	92·2	1585	1604·1
6	85	66·2	1670	1670·2
7	53	49·7	1723	1720·0
8	54	38·7	1777	1758·6
9	31	30·9	1808	1789·5
10	40	25·2	1848	1814·7
11	29	21·0	1877	1835·7
12	13	17·7	1890	1853·5
13	19	15·2	1909	1868·6
14	18	13·1	1927	1881·8
15	20	11·5	1947	1893·2
16	12	10·1	1959	1903·4
17	13	9·0	1972	1912·3
18	14	8·0	1986	1920·4
19	9	7·2	1995	1927·6
20	4	6·5	1999	1934·1
21	5	5·9	2004	1940·0
22	6	5·4	2010	1945·4
23	7	4·9	2017	1950·4
24	2	4·5	2019	1954·9
25	3	4·2	2022	1959·1
26	3	3·9	2025	1963·0
27	3	3·6	2028	1966·6
30	1	2·9	2029	1976·0
31	2	2·7	2031	1978·8
32	3	2·6	2034	1981·3
33	5	2·4	2039	1983·8
35	2	2·2	2041	1988·2
36	5	2·0	2046	1990·2
39	2	1·7	2048	1995·7
40	1	1·7	2049	1997·4
48	1	1·1	2050	2008·1
49	1	1·1	2051	2009·3
50	1	1·1	2052	2010·3
61	1	0·7	2053	2019·7
63	1	0·7	2054	2021·1
86	1	0·4	2055	2032·2
90	1	0·3	2056	2033·5
91	1	0·3	2057	2033·9
103	1	0·3	2058	2037·2
104	1	0·2	2059	2037·5
135	1	0·1	2060	2043·3
138	1	0·1	2061	2043·7
146	1	0·1	2062	2044·8
251	1	0·0	2063	2052·3

Data drawn from the MIPS version 64 database and tabulated for 60,966 matches lying between PAM 29·5 and PAM 40 found in 845 connected components. Identification of single matches between connected components yielded 2679 suitable matches. Columns 3 and 5 show the expected values for a Zipfian distribution (eqn (7), see the text) and the cumulative counts when $\theta = 1·7$.

tion. This is done for four sets of data drawn again from PAM window 29·5 to 40 in Table 2, first for all gaps shorter than 21. the second for all gaps shorter than 41, the third for all gaps shorter than 61, and finally for the entire sample. These sets maintain significant sample sizes, include overlapping data, yet yield remarkably different values for β (Table 3).

It is, of course, possible to approximate the data arbitrarily well by a larger number of exponential functions (Demchuk *et al.*, 1989). However, it proved to be more productive to search for an alternative mathematical description to account for the gap length distribution. The data in Table 2 turned out to fit remarkably well a generalized Zipf law (Gonnet & Baeze-Yates. 1991), where the frequency of a gap of length k is proportional to $k^{-\theta}$ (eqn (7)):

$$\text{frequency of gap} = mk^{-\theta}. \qquad (7)$$

Columns 3 and 5 of Table 4 show the expected values for the distribution and the cumulative counts when $\theta = 1·7$. applied to the data from PAM window 29·5 to 40 (Table 2). A Zipfian distribution with an exponent of 1·7 approximates quite closely the observed gap length distribution. In other words. the probability of a gap is inversely proportional to the length of the gap raised to the 1·7 power.

Zipfian distributions with an exponent of 2 or less have infinite first and second moments (Gonnet & Baeza-Yates. 1991). This means that for any finite sample, the estimated mean and estimated standard deviation can be arbitrarily large. This is consistent with the empirical observation that the mean gap length has an abnormally large standard deviation in various samples of the data examined.

The parameters of the Zipfian distribution were found to be largely independent of the PAM distance of the pairs of proteins being examined, as summarized in Table 1. The only trend is to moderately shorter average gap lengths at longer PAM distances, although the sample size at low PAM distance is sufficiently small to make this trend less statistically significant than Table 1 might make it appear. In any case, the data eliminate the conjecture that, as a rule, gaps enlarge over evolutionary distance. Further, they imply that insertions and deletion events occur with a particular probability distribution and, once the event has occurred, the probability of subsequent insertion and deletion events in the sample position is not greatly different than in the protein generally.

Table 5

Distribution of amino acids in and around gap

	1	1/*f*	2	2/*f*	3	3/*f*	4	4/*f*	5	5/*f*	6	6/*f*	Z	Z/*f*
A. *PAM bounded between 11·8 and 16* 55,730 matches read of which 55,730 were within given bounds; 695 connected components with a suitable match out of 3816														
Ala	10·01	1·33	9·08	1·20	10·75	1·42	8·74	1·16	7·74	1·03	9·08	1·20	8·42	1·12
Arg	5·34	1·05	4·14	0·81	3·63	0·71	3·63	0·71	4·81	0·95	4·41	0·87	4·98	0·98
Asn	5·61	1·30	4·54	1·05	5·51	1·28	4·84	1·12	4·81	1·11	4·01	0·93	4·02	0·93
Asp	5·74	1·11	4·94	0·96	5·51	1·07	4·70	0·91	3·60	0·70	4·01	0·78	5·18	1·00
Cys	1·47	0·79	1·07	0·58	1·34	0·72	1·34	0·72	0·67	0·36	0·67	0·36	0·86	0·46
Gln	6·28	1·51	5·07	1·22	6·72	1·62	5·11	1·23	6·68	1·61	5·61	1·35	5·69	1·37
Glu	5·87	0·94	6·01	0·96	6·18	0·99	6·72	1·07	5·87	0·94	5·21	0·83	5·31	0·85
Gly	9·35	1·26	10·68	1·44	8·47	1·14	11·29	1·07	12·28	1·66	12·15	1·64	10·98	1·48
His	1·87	0·83	2·80	1·25	2·02	0·90	2·42	1·08	2·80	1·25	3·34	1·49	2·35	1·05
Ile	2·40	0·44	2·00	0·37	3·09	0·56	2·28	0·42	4·14	0·76	3·34	0·61	4·20	0·77
Leu	4·81	0·52	5·07	0·55	4·44	0·48	5·91	0·64	7·61	0·83	6·68	0·73	6·50	0·71
Lys	4·67	0·79	4·67	0·79	5·24	0·89	5·11	0·87	3·74	0·64	4·41	0·75	4·70	0·80
Met	1·34	0·60	2·80	1·25	0·94	0·42	2·28	1·02	1·20	0·54	2·27	1·01	1·85	0·83
Phe	2·14	0·53	3·34	0·83	1·61	0·40	2·96	0·74	2·40	0·60	2·80	0·70	3·39	0·84
Pro	8·54	1·63	8·95	1·71	8·60	1·64	6·99	1·34	8·68	1·66	7·08	1·35	7·71	1·47
Ser	10·81	1·56	10·28	1·49	12·50	1·81	11·29	1·63	8·41	1·22	8·14	1·18	8·04	1·16
Thr	6·28	1·07	5·87	1·00	6·85	1·17	5·78	0·98	6·14	1·05	6·81	1·16	6·17	1·05
Trp	0·67	0·50	0·80	0·59	0·67	0·50	0·40	0·30	0·53	0·39	0·53	0·39	0·83	0·61
Tyr	2·94	0·92	3·34	1·03	2·55	0·78	3·63	1·12	3·34	1·03	3·47	1·07	3·06	0·94
Val	3·87	0·59	4·54	0·69	3·23	0·49	4·30	0·66	4·41	0·67	6·01	0·92	5·72	0·87
Unk	0·00		0·00		0·13		0·27		0·13		0·00		0·03	
r.m.s.	2·07		1·96		2·39		1·89		1·83		1·64		1·36	
B. *PAM bounded between 16 and 21·7* 77,408 matches read of which 77,408 were within given bounds; 753 connected components with a suitable match out of 3707														
Ala	10·05	1·33	8·91	1·18	10·16	1·35	9·17	1·21	9·15	1·21	9·23	1·22	9·15	1·21
Arg	5·23	1·03	4·66	0·92	4·59	0·90	4·59	0·90	4·82	0·95	4·08	0·80	4·35	0·86
Asn	4·00	0·93	4·74	1·10	4·10	0·95	4·91	1·14	3·59	0·83	3·51	0·81	3·64	0·84
Asp	6·94	1·34	4·58	0·89	6·47	1·25	5·16	1·00	4·17	0·81	4·66	0·90	4·28	0·83
Cys	1·23	0·66	1·31	0·71	1·23	0·66	1·06	0·57	1·14	0·62	1·39	0·75	1·41	0·76
Gln	4·49	1·08	6·29	1·52	4·42	1·07	6·39	1·54	4·98	1·20	5·88	1·42	7·39	1·78
Glu	7·76	1·24	8·82	1·41	7·70	1·23	9·01	1·44	7·68	1·23	7·35	1·17	7·06	1·13
Gly	8·91	1·20	10·87	1·47	9·58	1·29	10·81	1·46	12·50	1·69	11·68	1·58	9·91	1·34
His	1·96	0·88	1·31	0·58	1·39	0·62	1·39	0·62	2·04	0·91	2·12	0·95	2·55	1·14
Ile	3·02	0·55	2·61	0·48	2·70	0·49	2·70	0·49	3·92	0·72	3·43	0·63	3·47	0·63
Leu	4·90	0·53	4·90	0·53	5·49	0·60	4·75	0·52	5·88	0·64	6·37	0·69	6·72	0·73
Lys	5·72	0·97	6·62	1·13	6·14	1·04	6·47	1·10	4·49	0·76	4·82	0·82	4·98	0·85
Met	1·14	0·51	1·47	0·66	1·23	0·55	1·64	0·73	1·55	0·69	1·80	0·80	1·61	0·72
Phe	2·37	0·59	2·21	0·55	2·05	0·51	1·88	0·47	1·80	0·45	1·72	0·43	2·27	0·56
Pro	8·74	1·67	7·03	1·34	7·86	1·50	5·90	1·13	8·25	1·58	8·17	1·56	8·01	1·53
Ser	8·09	1·17	10·46	1·51	9·42	1·36	11·88	1·72	8·50	1·23	7·19	1·04	8·18	1·18
Thr	7·11	1·21	5·39	0·92	7·70	1·31	5·49	0·94	8·01	1·36	7·60	1·29	6·56	1·12
Trp	0·57	0·42	0·49	0·36	0·82	0·61	0·41	0·30	0·57	0·42	1·06	0·79	1·04	0·77
Tyr	2·53	0·78	2·37	0·73	2·62	0·81	2·38	0·73	1·88	0·58	2·53	0·78	2·31	0·71
Val	5·15	0·79	4·74	0·72	4·34	0·66	3·93	0·60	5·07	0·77	5·39	0·82	5·11	0·78
Unk	0·08		0·25		0·00		0·08		0·00		0·00		0·00	
r.m.s.	1.75		1.98		1.81		2.18		1·94		1·73		1·57	
C. *PAM bounded between 21·7 and 29·8* 105,933 matches read of which 105,933 were within given bounds; 723 connected components with a suitable match out of 3579														
Ala	9·33	1·24	10·78	1·43	9·99	1·46	11·05	1·46	10·32	1·37	9·53	1·26	8·04	1·06
Arg	4·47	0·88	5·26	1·04	4·60	0·91	4·08	0·80	4·47	0·88	4·73	0·93	4·51	0·89
Asn	4·99	1·16	3·42	0·79	4·93	1·14	4·40	1·02	3·81	0·88	4·40	1·02	4·51	1·04
Asp	5·65	1·09	5·06	0·98	6·25	1·21	5·13	0·99	4·93	0·95	5·39	1·04	5·31	1·03
Cys	1·12	0·61	1·38	0·75	0·85	0·46	1·38	0·75	0·72	0·39	1·38	0·75	1·08	0·58
Gln	5·58	1·34	5·52	1·33	5·92	1·43	5·85	1·41	5·98	1·44	5·85	1·41	6·16	1·48
Glu	7·42	1·19	7·62	1·22	7·10	1·13	7·69	1·23	5·85	0·93	5·39	0·86	6·71	1·07
Gly	10·97	1·48	10·71	1·45	10·19	1·38	10·32	1·39	11·70	1·58	11·30	1·53	9·81	1·33
His	3·42	1·53	2·10	0·94	2·63	1·17	2·37	1·06	1·97	0·88	1·71	0·76	2·22	0·99
Ile	2·69	0·49	2·83	0·52	2·70	0·49	2·10	0·38	2·69	0·49	2·56	0·47	3·36	0·61
Leu	6·31	0·69	5·65	0·61	5·39	0·59	6·18	0·67	6·96	0·76	7·16	0·78	6·52	0·71
Lys	5·72	0·97	5·65	0·96	5·65	0·96	5·98	1·02	6·04	1·03	4·99	0·85	5·96	1·01
Met	1·18	0·53	1·12	0·50	1·05	0·47	1·18	0·53	1·64	0·73	1·77	0·79	1·49	0·67
Phe	2·23	0·55	3·15	0·78	2·56	0·64	2·37	0·59	2·50	0·62	2·10	0·52	2·75	0·68
Pro	8·02	1·53	6·64	1·27	7·96	1·52	6·57	1·26	8·15	1·56	8·34	1·59	8·08	1·54
Ser	8·08	1·17	8·67	1·25	9·07	1·31	10·59	1·53	6·96	1·01	7·42	1·07	8·41	1·22
Thr	5·65	0·96	5·98	1·02	6·18	1·05	5·13	0·87	6·57	1·12	6·96	1·19	6·34	1·08
Trp	0·59	0·44	0·53	0·39	0·46	0·34	0·39	0·29	1·05	0·78	1·31	0·97	1·11	0·82

Table 5 *(continued)*

	1	1/f	2	2/f	3	3/f	4	4/f	5	5/f	6	6/f	Z	Z/f
Tyr	1·58	0·49	2·69	0·83	2·10	0·65	2·37	0·73	1·97	0·61	2·17	0·67	2·50	0·77
Val	4·86	0·74	4·86	0·74	4·21	0·64	4·87	0·74	5·72	0·87	5·52	0·84	5·13	0·78
Unk	0·13		0·39		0·20		0·00		0·00		0·00		0·00	
r.m.s.	1·70		1·69		1·82		1·89		1·71		1·64		1·38	
D. *PAM bounded between 29·5 and 40* 146,609 matches read of which 146,609 were within given bounds: 755 connected components with a suitable match out of 3446														
Ala	8·29	1·10	7·42	0·98	8·24	1·09	8·58	1·14	8·53	1·13	7·42	0·98	8·28	1·10
Arg	5·28	1·04	4·90	0·96	5·09	1·00	4·12	0·81	4·51	0·89	4·94	0·97	4·79	0·94
Asn	4·99	1·16	5·33	1·23	5·48	1·27	5·19	1·20	3·93	0·91	4·36	1·01	4·41	1·02
Asp	5·57	1·08	5·77	1·12	6·20	1·20	6·06	1·17	4·07	0·79	4·46	0·86	4·57	0·88
Cys	1·65	0·89	1·65	0·89	1·60	0·86	1·41	0·76	0·82	0·44	1·41	0·76	1·37	0·74
Gln	4·75	1·14	4·85	1·17	4·85	1·17	5·62	1·35	5·53	1·33	5·19	1·25	5·77	1·39
Glu	7·46	1·19	7·80	1·25	7·61	1·22	7·95	1·27	6·01	0·96	5·77	0·92	6·50	1·04
Gly	10·42	1·41	10·37	1·40	9·69	1·31	10·71	1·45	12·60	1·70	12·55	1·70	11·91	1·61
His	2·04	0·91	2·13	0·95	1·84	0·82	2·04	0·91	2·28	1·02	2·42	1·08	2·20	0·98
Ile	3·30	0·60	3·64	0·67	3·01	0·55	3·49	0·64	3·59	0·66	2·91	0·53	3·46	0·63
Leu	5·72	0·62	5·77	0·63	4·85	0·53	5·67	0·62	5·48	0·60	5·87	0·64	6·03	0·66
Lys	7·03	1·20	7·03	1·20	6·88	1·17	6·69	1·14	5·87	1·00	6·11	1·04	5·26	0·89
Met	1·55	0·69	1·26	0·56	1·50	0·67	1·50	0·67	1·89	0·84	1·89	0·84	1·49	0·67
Phe	2·28	0·57	2·23	0·58	2·13	0·53	2·28	0·57	2·71	0·67	3·20	0·80	2·76	0·69
Pro	8·14	1·56	6·40	1·22	7·76	1·48	6·50	1·24	8·05	1·54	7·90	1·51	9·08	1·74
Ser	7·95	1·15	9·89	1·43	9·74	1·41	10·23	1·48	7·90	1·14	7·85	1·14	7·88	1·14
Thr	6·25	1·06	5·91	1·01	6·25	1·06	5·43	0·93	6·30	1·07	5·87	1·00	5·47	0·93
Trp	0·87	0·64	0·78	0·58	0·73	0·54	0·63	0·47	0·73	0·54	0·92	0·68	0·96	0·71
Tyr	1·99	0·61	2·13	0·66	2·04	0·63	1·60	0·49	2·42	0·74	2·57	0·79	2·57	0·79
Val	4·46	0·68	4·56	0·70	4·51	0·69	4·31	0·66	6·79	1·04	6·40	0·98	5·22	0·80
Unk	0·00		0·10		0·00		0·00		0·00		0·00		0·00	
r.m.s.	1·58		1·57		1·75		1·75		1·76		1·67		1·72	
E. *PAM bounded between 40 and 54·3* 219,384 matches read of which 219,384 were within given bounds: 749 connected components with a suitable match out of 3219														
Ala	7·54	1·00	6·92	0·92	7·99	1·06	8·03	1·06	6·62	0·88	7·54	1·00	7·19	0·95
Arg	5·77	1·14	5·03	0·99	4·88	0·96	4·33	0·85	4·81	0·95	4·77	0·94	4·98	0·98
Asn	3·74	0·87	5·21	1·21	5·03	1·16	5·62	1·30	4·55	1·05	4·47	1·03	4·65	1·08
Asp	5·88	1·14	5·73	1·11	5·66	1·09	5·70	1·10	5·33	1·03	5·62	1·09	5·42	1·05
Cys	1·63	0·88	1·15	0·62	1·41	0·76	1·00	0·54	1·07	0·58	1·00	0·54	1·10	0·59
Gln	5·51	1·33	4·92	1·19	5·29	1·27	5·29	1·27	4·96	1·20	4·33	1·04	4·85	1·17
Glu	6·58	1·05	7·03	1·12	6·99	1·12	7·25	1·16	6·69	1·07	6·62	1·06	7·13	1·14
Gly	9·28	1·25	10·17	1·37	9·06	1·22	10·21	1·38	10·39	1·40	9·87	1·33	9·71	1·31
His	1·92	0·86	1·89	0·84	1·70	0·76	2·00	0·89	2·03	0·91	2·14	0·96	1·91	0·85
Ile	3·55	0·65	3·66	0·67	3·26	0·60	2·92	0·53	3·25	0·59	3·62	0·66	3·62	0·66
Leu	6·25	0·68	6·40	0·70	6·18	0·67	6·36	0·69	7·47	0·81	7·77	0·84	7·18	0·78
Lys	7·03	1·20	6·92	1·18	6·59	1·12	6·96	1·18	5·84	0·99	5·77	0·98	5·72	0·97
Met	1·04	0·46	1·44	0·64	1·29	0·58	0·85	0·38	1·78	0·79	2·03	0·91	1·71	0·76
Phe	3·00	0·75	2·66	0·66	2·74	0·68	2·74	0·68	3·70	0·92	3·51	0·87	2·89	0·72
Pro	8·36	1·60	6·95	1·33	8·36	1·60	7·33	1·40	8·39	1·60	8·21	1·57	8·39	1·60
Ser	8·47	1·23	9·47	1·37	10·03	1·45	9·40	1·36	7·10	1·03	6·77	0·98	8·27	1·20
Thr	6·07	1·03	6·69	1·14	6·70	1·14	5·99	1·02	6·18	1·05	5·77	0·98	6·29	1·07
Trp	0·44	0·33	0·63	0·47	0·52	0·39	0·67	0·50	0·74	0·55	0·78	0·58	0·90	0·67
Tyr	2·40	0·74	2·26	0·70	1·89	0·58	1·81	0·56	2·85	0·88	3·11	0·96	2·88	0·89
Val	5·55	0·85	4·77	0·73	4·40	0·67	5·55	0·85	6·25	0·95	6·29	0·96	5·22	0·80
Unk	0·00		0·11		0·04		0·00		0·00		0·00		0·00	
r.m.s.	1·38		1·42		1·59		1·53		1·23		1·06		1·24	
F. *PAM bounded between 54·3 and 73·7* 333,060 matches read of which 333,060 were within given bounds: 740 connected components with a suitable match out of 2941														
Ala	7·05	0·93	7·14	0·95	8·12	1·08	7·36	0·97	6·22	0·82	6·57	0·87	7·44	0·99
Arg	4·64	0·91	4·90	0·96	4·52	0·89	4·77	0·94	5·18	1·02	4·90	0·96	5·55	1·09
Asn	4·39	1·02	4·52	1·05	5·09	1·18	5·06	1·17	3·82	0·88	4·04	0·94	4·40	1·02
Asp	6·19	1·20	6·60	1·28	6·35	1·23	6·13	1·19	5·69	1·10	5·94	1·15	5·68	1·10
Cys	1·36	0·74	1·77	0·96	1·11	0·60	1·67	0·90	1·39	0·75	1·52	0·82	1·49	0·81
Gln	4·49	1·08	4·11	0·99	3·51	0·85	4·01	0·97	3·82	0·92	3·76	0·91	4·33	1·04
Glu	6·92	1·11	6·76	1·08	6·92	1·11	6·98	1·12	6·35	1·01	5·81	0·93	6·68	1·07
Gly	8·91	1·20	8·63	1·17	8·28	1·12	8·53	1·15	10·33	1·40	10·39	1·40	8·92	1·21
His	2·34	1·04	2·02	0·90	2·40	1·07	2·27	1·01	2·09	0·93	2·72	1·21	2·35	1·05
Ile	4·23	0·77	3·57	0·65	3·70	0·68	3·38	0·62	3·92	0·72	3·79	0·69	3·75	0·69
Leu	7·27	0·79	7·93	0·86	6·57	0·71	7·27	0·79	8·69	0·94	7·74	0·84	7·37	0·80
Lys	7·42	1·26	6·22	1·06	7·24	1·23	5·97	1·02	5·53	0·94	6·64	1·13	5·85	0·99
Met	1·77	0·79	1·71	0·76	1·58	0·71	1·58	0·71	1·74	0·78	1·23	0·55	1·77	0·79
Phe	3·98	0·99	3·19	0·79	3·29	0·82	2·72	0·68	3·44	0·86	3·25	0·81	3·10	0·77
Pro	6·79	1·30	7·05	1·35	6·95	1·33	6·95	1·33	7·84	1·50	8·63	1·65	7·92	1·51

Table 5 *(continued)*

	1	1/*f*	2	2/*f*	3	3/*f*	4	4/*f*	5	5/*f*	6	6/*f*	Z	Z/*f*
Ser	8·37	1·21	8·94	1·29	9·64	1·40	10·24	1·48	6·79	1·48	6·03	0·87	7·53	1·09
Thr	5·21	0·89	6·48	1·10	6·22	1·06	6·86	1·17	6·41	1·09	5·88	1·00	6·18	1·05
Trp	0·88	0·65	1·14	0·84	0·82	0·61	0·95	0·70	1·14	0·84	1·14	0·84	1·07	0·79
Tyr	2·91	0·90	2·12	0·65	2·84	0·87	1·77	0·54	3·57	1·10	3·51	1·08	3·02	0·93
Val	4·74	0·72	4·68	0·71	4·83	0·74	5·50	0·84	6·00	0·91	6·51	0·99	5·57	0·85
Unk	0·13		0·54		0·03		0·03		0·03		0·00		0·03	
r.m.s.	1·03		1·08		1·26		1·26		1·05		1·24		0·98	
G. PAM bounded between 73·7 and 100														
524,639 matches read of which 524,639 were within given bounds; 708 connected components with a suitable match out of 2573														
Ala	6·81	0·90	6·75	0·89	7·36	0·97	7·73	1·02	6·49	0·86	5·75	0·76	6·73	0·89
Arg	4·80	0·94	4·71	0·93	4·89	0·96	5·17	1·02	4·60	0·91	5·11	1·01	4·82	0·95
Asn	4·68	1·08	5·06	1·17	5·80	1·34	5·80	1·34	4·43	1·03	4·17	0·97	4·57	1·06
Asp	6·52	1·26	6·47	1·25	6·64	1·28	5·52	1·07	6·24	1·21	5·83	1·13	6·15	1·19
Cys	1·67	0·90	1·70	0·92	1·09	0·59	1·64	0·89	1·41	0·76	1·41	0·76	1·43	0·77
Gln	4·14	1·00	3·88	0·93	3·85	0·93	4·31	1·04	4·48	1·08	4·11	0·99	4·58	1·10
Glu	7·13	1·14	6·58	1·05	7·21	1·15	6·87	1·10	6·29	1·00	6·64	1·06	7·07	1·13
Gly	8·91	1·20	9·08	1·23	8·02	1·08	8·79	1·19	10·14	1·37	9·63	1·30	9·01	1·22
His	2·41	1·08	2·59	1·16	2·67	1·19	2·07	0·92	2·21	0·99	2·47	1·10	2·31	1·03
Ile	4·43	0·81	4·05	0·74	3·68	0·67	4·11	0·75	3·82	0·70	4·08	0·75	4·14	0·76
Leu	7·39	0·80	7·50	0·82	6·58	0·72	7·47	0·81	8·02	0·87	8·42	0·92	7·78	0·85
Lys	7·21	1·23	6·98	1·19	6·98	1·19	6·67	1·13	6·87	1·17	6·87	1·17	6·67	1·13
Met	1·32	0·59	1·38	0·62	1·52	0·68	1·18	0·53	1·98	0·88	2·27	1·01	1·80	0·80
Phe	3·53	0·88	3·39	0·84	3·22	0·80	2·99	0·74	3·76	0·94	3·97	0·99	3·34	0·83
Pro	6·09	1·16	6·44	1·23	5·95	1·14	7·04	1·35	7·33	1·40	7·33	1·40	7·20	1·38
Ser	7·64	1·11	7·39	1·07	9·05	1·31	8·45	1·22	5·86	0·85	5·46	0·79	6·97	1·01
Thr	6·32	1·08	5·92	1·01	6·55	1·12	5·78	0·98	5·09	0·87	5·49	0·94	5·74	0·98
Trp	1·21	0·90	1·15	0·85	0·92	0·68	1·09	0·81	1·87	1·39	1·38	1·02	1·29	0·96
Tyr	2·82	0·87	2·99	0·92	2·70	0·83	2·36	0·73	3·76	1·16	3·62	1·11	3·04	0·94
Val	4·97	0·76	6·01	0·92	5·32	0·81	4·97	0·76	5·34	0·81	6·01	0·92	5·36	0·82
Unk	0·00		0·00		0·00		0·00		0·00		0·00		0·00	
r.m.s.	0·92		0·88		1·15		1·03		1·09		0·99		0·89	

Collected from the MIPS Version 64 database. Entries indicate the frequency of occurrence of the designated amino acid at the designated position, and this frequency divided by *f*, the frequency of occurrence of the designated amino acid in the database as a whole. These frequencies are given in Table 7. Unk is unknown amino acid. Entries at positions in the insert and flanking regions defined below:

...XXX1__...__2XXX...
...YYY35ZZZZZ64YYY...

(b) *Probability of a gap as a function of evolutionary distance*

As is evident from the data in Table 1, the probability of an indel increases with increasing PAM distance. The relation is linear only at short PAM distances, however. Equation (8) fits these data as an exponential ($\Sigma\Delta x^2 = 2{\cdot}2$):

$$\text{indel/amino acid} = 0{\cdot}0224 - 0{\cdot}0219 \times \exp(-0{\cdot}0102 \cdot \text{PAM}). \quad (8)$$

A remarkably linear relationship exists, however, between the average number of amino acid residues between indel and the reciprocal of PAM distance (× 100. Fig. 2), which represents the average amino acid residues per mutation. Extrapolating this relationship to the *y* axis, representing the point where the protein pairs have accumulated an infinite number of accepted point mutations, yields an intercept of 30 amino acid residues. This result suggests (*vide infra*) that segments of proteins on average 30 amino acid residues in length remain undisrupted by divergent evolution even after extended periods of time. Extrapolation of the exponential approximation yields an undisrupted unit approximately 40 amino acid residues in length.

(c) *Computing the probability of a gap and a gap penalty for aligning sequences*

These observations can be combined to yield an equation for computing the probability of an indel of length *k*, and the corresponding penalty that should be assigned to the gap of this length when found in an alignment of two homologous protein sequences. Assuming a linear dependence of gap probability on PAM distance (accurate only at short PAM distances), this equation is:

$$\text{Probability}\{\text{indel of length } k\} \approx (c_1) \cdot \text{PAM}/k^6. \quad (9)$$

In a typical dynamic programming alignment, costs are traditionally expressed at ten times the logarithm (base 10) of probability. Conforming to this tradition, the cost of a gap in an alignment is expressed by the following formula:

$$\text{Cost}\{\text{indel of length } k\} \approx c_2 + 10 \cdot \log_{10}(\text{PAM}) - 17 \cdot \log_{10}(k). \quad (10)$$

Table 6

Normalized distribution of amino acids in and around gap as a function of PAM distance

		Position relative to the deletion						
Amino acid	Mid PAM	1	2	3	4	5	6	*Z*
Ala	13·9	1·33	1·20	1·42	1·16	1·03	1·20	1·12
Ala	18·85	1·33	1·18	1·35	1·21	1·21	1·22	1·21
Ala	25·6	1·24	1·43	1·46	1·46	1·37	1·26	1·06
Ala	34·75	1·10	0·98	1·09	1·14	1·13	0·98	1·10
Ala	47·15	1·00	0·92	1·06	1·06	0·88	1·00	0·95
Ala	64·0	0·93	0·95	1·08	0·97	0·82	0·87	0·99
Ala	86·85	0·90	0·89	0·97	1·02	0·86	0·76	0·89
Arg	13·9	1·05	0·81	0·71	0·71	0·95	0·87	0·98
Arg	18·85	1·03	0·92	0·90	0·90	0·95	0·80	0·86
Arg	25·6	0·88	1·04	0·91	0·80	0·88	0·93	0·89
Arg	34·75	1·04	0·96	1·00	0·81	0·89	0·97	0·94
Arg	47·15	1·14	0·99	0·96	0·85	0·95	0·94	0·98
Arg	64·0	0·91	0·96	0·89	0·94	1·02	0·96	1·09
Arg	86·85	0·94	0·93	0·96	1·02	0·91	1·01	0·95
Asn	13·9	1·30	1·05	1·28	1·12	1·11	0·93	0·93
Asn	18·85	0·93	1·10	0·95	1·14	0·83	0·81	0·84
Asn	25·6	1·16	0·79	1·14	1·02	0·88	1·02	1·04
Asn	34·75	1·16	1·23	1·27	1·20	0·91	1·01	1·02
Asn	47·15	0·87	1·21	1·16	1·30	1·05	1·03	1·08
Asn	64·0	1·02	1·05	1·18	1·17	0·88	0·94	1·02
Asn	86·85	1·08	1·17	1·34	1·34	1·03	0·97	1·06
Asp	13·9	1·11	0·96	1·07	0·91	0·70	0·78	1·00
Asp	18·85	1·34	0·89	1·25	1·00	0·81	0·90	0·83
Asp	25·6	1·09	0·98	1·21	0·99	0·95	1·04	1·03
Asp	34·75	1·08	1·12	1·20	1·17	0·79	0·86	0·88
Asp	47·15	1·14	1·11	1·09	1·10	1·03	1·09	1·05
Asp	64·0	1·20	1·28	1·23	1·19	1·10	1·15	1·10
Asp	86·85	1·26	1·25	1·28	1·07	1·21	1·13	1·19
Cys	13·9	0·79	0·58	0·72	0·72	0·36	0·36	0·46
Cys	18·85	0·66	0·71	0·66	0·57	0·62	0·75	0·76
Cys	25·6	0·61	0·75	0·46	0·75	0·39	0·75	0·58
Cys	34·75	0·89	0·89	0·86	0·76	0·44	0·76	0·74
Cys	47·15	0·88	0·62	0·76	0·54	0·58	0·54	0·59
Cys	64·0	0·74	0·96	0·60	0·90	0·75	0·82	0·81
Cys	86·85	0·90	0·92	0·59	0·89	0·76	0·76	0·77
Gln	13·9	1·51	1·22	1·62	1·23	1·61	1·35	1·37
Gln	18·85	1·08	1·52	1·07	1·54	1·20	1·42	1·78
Gln	25·6	1·34	1·33	1·43	1·41	1·44	1·41	1·48
Gln	34·75	1·14	1·17	1·17	1·35	1·33	1·25	1·39
Gln	47·15	1·33	1·19	1·27	1·27	1·20	1·04	1·17
Gln	64·0	1·08	0·99	0·85	0·97	0·92	0·91	1·04
Gln	86·85	1·00	0·93	0·93	1·04	1·08	0·99	1·10
Glu	13·9	0·94	0·96	0·99	1·07	0·94	0·83	0·85
Glu	18·85	1·24	1·41	1·23	1·44	1·23	1·17	1·13
Glu	25·6	1·19	1·22	1·13	1·23	0·93	0·86	1·07
Glu	34·75	1·19	1·25	1·22	1·27	0·96	0·92	1·04
Glu	47·15	1·05	1·12	1·12	1·16	1·07	1·06	1·14
Glu	64·0	1·11	1·08	1·11	1·12	1·01	0·93	1·07
Glu	86·85	1·14	1·05	1·15	1·10	1·00	1·06	1·13
Gly	13·9	1·26	1·44	1·14	1·07	1·66	1·64	1·48
Gly	18·85	1·20	1·47	1·29	1·46	1·69	1·58	1·34
Gly	25·6	1·48	1·45	1·38	1·39	1·58	1·53	1·33
Gly	34·75	1·41	1·40	1·31	1·45	1·70	1·70	1·61
Gly	47·15	1·25	1·37	1·22	1·38	1·40	1·33	1·31
Gly	64·0	1·20	1·17	1·12	1·15	1·40	1·40	1·21
Gly	86·85	1·20	1·23	1·08	1·19	1·37	1·30	1·22
His	13·9	0·83	1·25	0·90	1·08	1·25	1·49	1·05
His	18·85	0·88	0·58	0·62	0·62	0·91	0·95	1·14
His	25·6	1·53	0·94	1·17	1·06	0·88	0·76	0·99
His	34·75	0·91	0·95	0·82	0·91	1·02	1·08	0·98
His	47·15	0·86	0·84	0·76	0·89	0·91	0·96	0·85
His	64·0	1·04	0·90	1·07	1·01	0·93	1·21	1·05
His	86·85	1·08	1·16	1·19	0·92	0·99	1·10	1·03
Ile	13·9	0·44	0·37	0·56	0·42	0·76	0·61	0·77
Ile	18·85	0·55	0·48	0·49	0·49	0·72	0·63	0·63
Ile	25·6	0·49	0·52	0·49	0·38	0·49	0·47	0·61

Table 6 *(continued)*

		Position relative to the deletion						
Amino acid	Mid PAM	1	2	3	4	5	6	Z
Ile	34·75	0·60	0·67	0·55	0·64	0·66	0·53	0·63
Ile	47·15	0·65	0·67	0·60	0·53	0·59	0·66	0·66
Ile	64·0	0·77	0·65	0·68	0·62	0·72	0·69	0·69
Ile	86·85	0·81	0·74	0·67	0·75	0·70	0·75	0·76
Leu	13·9	0·52	0·55	0·48	0·64	0·83	0·73	0·71
Leu	18·85	0·53	0·53	0·60	0·52	0·64	0·69	0·73
Leu	25·6	0·69	0·61	0·59	0·67	0·76	0·78	0·71
Leu	34·75	0·62	0·63	0·53	0·62	0·60	0·64	0·66
Leu	47·15	0·68	0·70	0·67	0·69	0·81	0·84	0·78
Leu	64·0	0·79	0·86	0·71	0·79	0·94	0·84	0·80
Leu	86·85	0·80	0·82	0·72	0·81	0·87	0·92	0·85
Lys	13·9	0·79	0·79	0·89	0·87	0·64	0·75	0·80
Lys	18·85	0·97	1·13	1·04	1·10	0·76	0·82	0·85
Lys	25·6	0·97	0·96	0·96	1·02	1·03	0·85	1·01
Lys	34·75	1·20	1·20	1·17	1·14	1·00	1·04	0·89
Lys	47·15	1·20	1·18	1·12	1·18	0·99	0·98	0·97
Lys	64·0	1·26	1·06	1·23	1·02	0·94	1·13	0·99
Lys	86·85	1·23	1·19	1·19	1·13	1·17	1·17	1·13
Met	13·9	0·60	1·25	0·42	1·02	0·54	1·01	0·83
Met	18·85	0·51	0·66	0·55	0·73	0·69	0·80	0·72
Met	25·6	0·53	0·50	0·47	0·53	0·73	0·79	0·67
Met	34·75	0·69	0·56	0·67	0·67	0·84	0·84	0·67
Met	47·15	0·46	0·64	0·58	0·38	0·79	0·91	0·76
Met	64·0	0·79	0·76	0·71	0·71	0·78	0·55	0·79
Met	86·85	0·59	0·62	0·68	0·53	0·88	1·01	0·80
Phe	13·9	0·53	0·83	0·40	0·74	0·60	0·70	0·84
Phe	18·85	0·59	0·55	0·51	0·47	0·45	0·43	0·56
Phe	25·6	0·55	0·78	0·64	0·59	0·62	0·52	0·68
Phe	34·75	0·57	0·58	0·53	0·57	0·67	0·80	0·69
Phe	47·15	0·75	0·66	0·68	0·68	0·92	0·87	0·72
Phe	64·0	0·99	0·79	0·82	0·68	0·86	0·81	0·77
Phe	86·85	0·88	0·84	0·80	0·74	0·94	0·99	0·83
Pro	13·9	1·63	1·71	1·64	1·34	1·66	1·35	1·47
Pro	18·85	1·67	1·34	1·50	1·13	1·58	1·56	1·53
Pro	25·6	1·53	1·27	1·52	1·26	1·56	1·59	1·54
Pro	34·75	1·56	1·22	1·48	1·24	1·54	1·51	1·74
Pro	47·15	1·60	1·33	1·60	1·40	1·60	1·57	1·60
Pro	64·0	1·30	1·35	1·33	1·33	1·50	1·65	1·51
Pro	86·85	1·16	1·23	1·14	1·35	1·40	1·40	1·38
Ser	13·9	1·56	1·49	1·81	1·63	1·22	1·18	1·16
Ser	18·85	1·17	1·51	1·36	1·72	1·23	1·04	1·18
Ser	25·6	1·17	1·25	1·31	1·53	1·01	1·07	1·22
Ser	34·75	1·15	1·43	1·41	1·48	1·14	1·14	1·14
Ser	47·15	1·23	1·37	1·45	1·36	1·03	0·98	1·20
Ser	64·0	1·21	1·29	1·40	1·48	1·48	0·87	1·09
Ser	86·85	1·11	1·07	1·31	1·22	0·85	0·79	1·01
Thr	13·9	1·07	1·00	1·17	0·98	1·05	1·16	1·05
Thr	18·85	1·21	0·92	1·31	0·94	1·36	1·29	1·12
Thr	25·6	0·96	1·02	1·05	0·87	1·12	1·19	1·08
Thr	34·75	1·06	1·01	1·06	0·93	1·07	1·00	0·93
Thr	47·15	1·03	1·14	1·14	1·02	1·05	0·98	1·07
Thr	64·0	0·89	1·10	1·06	1·17	1·09	1·00	1·05
Thr	86·85	1·08	1·01	1·12	0·98	0·87	0·94	0·98
Trp	13·9	0·50	0·59	0·50	0·30	0·39	0·39	0·61
Trp	18·85	0·42	0·36	0·61	0·30	0·42	0·79	0·77
Trp	25·6	0·44	0·39	0·34	0·29	0·78	0·97	0·82
Trp	34·75	0·64	0·58	0·54	0·47	0·54	0·68	0·71
Trp	47·15	0·33	0·47	0·39	0·50	0·55	0·58	0·67
Trp	64·0	0·65	0·84	0·61	0·70	0·84	0·84	0·79
Trp	86·85	0·90	0·85	0·68	0·81	1·39	1·02	0·96
Tyr	13·9	0·92	1·03	0·78	1·12	1·03	1·07	0·94
Tyr	18·85	0·78	0·73	0·81	0·73	0·58	0·78	0·71
Tyr	25·6	0·49	0·83	0·65	0·73	0·61	0·67	0·77
Tyr	34·75	0·61	0·66	0·63	0·49	0·74	0·79	0·79
Tyr	47·15	0·74	0·70	0·58	0·56	0·88	0·96	0·89
Tyr	64·0	0·90	0·65	0·87	0·54	1·10	1·08	0·93
Tyr	86·85	0·87	0·92	0·83	0·73	1·16	1·11	0·94

Table 6 *(continued)*

		Position relative to the deletion						
Amino acid	Mid PAM	1	2	3	4	5	6	Z
Val	13·9	0·59	0·69	0·49	0·66	0·67	0·92	0·87
Val	18·85	0·79	0·72	0·66	0·60	0·77	0·82	0·78
Val	25·6	0·74	0·74	0·64	0·74	0·87	0·84	0·78
Val	34·75	0·68	0·70	0·69	0·66	1·04	0·98	0·80
Val	47·15	0·85	0·73	0·67	0·85	0·95	0·96	0·80
Val	64·0	0·72	0·71	0·74	0·84	0·91	0·99	0·85
Val	86·85	0·76	0·92	0·81	0·76	0·81	0·92	0·82

Collected from the MIPS Version 64 database. Entries indicate the frequency of occurrence of the designated amino acid at the designated position, and this frequency divided by f, the frequency of occurrence of the designated amino acid in the database as a whole. These frequencies are given in Table 7. Entries at positions in the insert and flanking regions defined below:

...XXX1__...__2XXX...
...YYY35ZZZZZ64YYY...

To fit the data, the constant c_2 must have a value of $-38{\cdot}08$. Adjusting the data to allow the coefficient of the $\log_{10}(\mathrm{PAM})$ term to vary (to accommodate the non-linearity of gap probability with PAM length at longer PAM distances), the cost equation becomes:

$$\mathrm{Cost}\{\text{indel of length } k\} \approx -35{\cdot}03 + 6{\cdot}88 \cdot \log_{10}(\mathrm{PAM}) + 17{\cdot}02 \cdot \log_{10}(k). \quad (11\mathrm{a})$$

The root-mean-square deviation is 0·7 of data for the eight sample sets at PAM windows above 8·7 for the range of deletions up to length 60. This error is small, for example, when compared with the entries in a typical Dayhoff matrix, which range from −8 to +17.

Equation (11) describes a gap penalty that is not linear in k, the length of the gap. Application of a non-linear gap penalty to a dynamic programming alignment is problematic. Therefore, we have calculated the best linear fit to equation (11) for routine use, emphasizing again that such a linear equation is a less satisfactory approximation than equation (11) itself. As more than 95% of the gaps are shorter than 20 amino acid residues, the approximation was restricted to this range. Adjusting by maximum likelihood to this range, we obtain:

$$\mathrm{Cost}\{\text{indel of length } k\} \approx -37{\cdot}31 + 6{\cdot}88 \log_{10}(\mathrm{PAM}) - 1{\cdot}47(k-1). \quad (12\mathrm{a})$$

To be used in standard alignment routines (at PAM = 250), equation (12) simplifies to:

$$\mathrm{Cost}\{\text{indel of length } k\} \approx -20{\cdot}8 - 1{\cdot}47(k-1). \quad (13\mathrm{a})$$

These parameters are sufficiently different from those found as defaults on commonly used alignment programs as to warrant their examination in special cases, although once again we must caution that a non-linear gap penalty is the only one that is grounded in empirical data.

To test the reliability of these results, an analogous equation was obtained from data derived from Version 19 of SwissProt. Further, as inspection of the data suggested that a primary source of error is the inclusion of fragments of protein sequences and precursor as separate entries, any entry with the word "fragment" or "precursor" in the description fields was automatically excluded before calculation. The following equations were obtained:

$$\mathrm{Cost}\{\text{indel of length } k\} \approx -35{\cdot}72 + 7{\cdot}22 \cdot \log_{10}(\mathrm{PAM}) + 16{\cdot}96 \cdot \log_{10}(k). \quad (11\mathrm{b})$$

A linear approximation of this is given by the equation (12b):

$$\mathrm{DelCost}(k) = -38{\cdot}07 + 7{\cdot}22 \log(p) - 1{\cdot}44(k-1), \quad (12\mathrm{b})$$

which, at PAM 250, gives:

$$\mathrm{DelCost}(k) = -20{\cdot}8 - 1{\cdot}44(k-1). \quad (13\mathrm{b})$$

(d) *Influence of protein type on gap length distribution*

To explore whether the exponential parameter θ of the Zipfian distribution is influenced by the type of protein, the database was separated into two classes of proteins, the immunoglobulins and the non-immunoglobulins, and the length distribution again examined. Divergent evolution within immunoglobulins is presumably dominated by functional variation, with deletion-prone splicing a presumed mechanism. In contrast, divergent evolution in non-immunoglobulins is dominated by point mutation, a large fraction of which is (again presumably) approximately neutral (Benner & Ellington, 1988). The rate of accumulation and types of indels accumulated during divergent evolution of immunoglobulins and non-immunoglobulins could conceivably differ for these reasons.

Formulae describing the probability and length distribution of gaps in the two sets of proteins were not greatly different. These are shown below (eqn (14) for immunoglobulins; eqn (15) for non-immunoglobulins; and eqn (16) for immunoglobulins and non-immunoglobulins together).

Table 7
Relative frequency of occurrence of amino acids in the database

Ala	7·55	Leu	9·20
Arg	5·08	Lys	5·88
Asn	4·32	Met	2·24
Asp	5·17	Phe	4·02
Cys	1·85	Pro	5·23
Gln	4·15	Ser	6·91
Glu	6·26	Thr	5·87
Gly	7·40	Trp	1·35
His	2·24	Tyr	3·25
Ile	5·47	Val	6·56

$$\text{Cost}\{\text{indel of length } k\}_{\text{immunoglob}} \approx -29{\cdot}1 + 3{\cdot}07[\log_{10}(\text{PAM})] - 17{\cdot}4[\log_{10}(k)]. \quad (14)$$

$$\text{Cost}\{\text{indel of length } k\}_{\text{non-immunoglob}} \approx -35{\cdot}3 + 7{\cdot}03[\log_{10}(\text{PAM})] - 17{\cdot}0[\log_{10}(k)]. \quad (15)$$

$$\text{Cost}\{\text{indel of length } k\}_{\text{both}} \approx -36{\cdot}8 + 7{\cdot}98[\log_{10}(\text{PAM})] - 17{\cdot}0[\log_{10}(k)]. \quad (16)$$

The parameter describing the gap length distribution is remarkably constant (−17·4, −17·0 and −17·0 for the 3 families), suggesting that the gap length distribution arises from features inherent to protein structure. In evaluating the other parameters of these equations, it must be remembered that the preponderance of matches within the immunoglobulin family is between sequences separated by low PAM distances, and the number of deletions examined for immunoglobulins is only approximately 1% that for non-immunoglobulins. This is the case despite the very large number of matches involving immunoglobulins, as the large number of repetitive matches within the immunoglobulin tree renders many matches redundant.

(e) *Amino acids flanking gap and within the insert*

Table 5 shows the probability of different amino acids being found at positions in and surrounding a gap, both in absolute terms and after normalization for the frequency of occurrence of the designated amino acid residue in the database as a whole (Table 7). Data are again tabulated for gaps appearing in alignments between protein sequences with different PAM distances. Table 6 shows a summary of these data at varying PAM distances by amino acid type, normalized for the frequency of occurrence of amino acid residues in the the database (Table 7).

4. Discussion

These results allow the construction of a formal empirical model describing insertions and deletions during divergent evolution. In this model, indels accumulate over evolutionary time, with units with an average length of 30 to 40 amino acid residues remaining undisturbed even after large amounts of divergence. The probability of an indel of length k is proportional to $1/k^{-\theta}$, where $\theta \approx 1{\cdot}7$. This relationship applies over the entire range of PAM distance. Thus, once created, the indel remains unchanged (or perhaps is slightly shortened); the region suffering an indel is not much more likely to suffer subsequent indels than is the rest of the protein.

These empirical observations are unlikely to need substantial revision as the database grows. The empirical laws expressing gap probability were derived from the entire database, not a subset of the database. The only sampling biases, therefore, are those that influenced the selection of proteins in the database itself. While we cannot exclude at this point the possibility that empirical laws will be different in some special proteins (e.g. membrane proteins or viral coat proteins), these proteins were represented in the database used to define the empirical laws reported here.

We next asked what these empirical laws might suggest about protein structure in general. For example, the fit of gap probability to the inverse 1·7 power of the gap length is quite good. Therefore, it is appropriate to search for a structural explanation for this empirical fact. In principle, mechanisms operating at both the DNA level and the protein level must be considered in explaining the gap length distribution. Although events occurring at the DNA level cannot be ruled out as factors influencing the gap length distribution, one explanation, based on assumptions concerning how natural selection operates at the level of proteins, proved to be particularly interesting.

Virtually all entries in the protein sequence database correspond to proteins that are functional in living organisms. Thus, for an insertion or deletion event to be represented in the database, it must be accepted by natural selection, subject to functional constraints. Like accepted point mutations, accepted indels are those that maintain the function of the protein within limits, themselves determined by the environment of the protein within the host organism. This implies that indels during the divergent evolution leading to sequences in contemporary protein sequence databases cannot have greatly disrupted the folded structure of the corresponding proteins.

To avoid disrupting the folded structure of a protein, deletions or insertions generally must extract or insert polypeptide segments whose ends are close in space in the folded structure. Prior to the insertion, the amino acid residues flanking the insert are joined by a covalent bond, and therefore must lie together in three-dimensional space. Conversely, the amino acid residues flanking a deleted segment must, after the deletion, be joined by a covalent bond. To obtain such a covalent bond without major reorganization of the protein fold, these amino acid residues must lie near in space prior to removal of the insert.

Next, we assume that only random coils are deleted or inserted. Further, we make the assumption, not entirely obvious, that the behaviour of a randomly coiled component of a folded polypeptide is governed by laws governing the statistical mechanics of isolated randomly coiled polymers (Flory, 1953).

For an ideal unidimensional randomly coiled polymer, the probability that the two ends lie together in three-dimensional space is inversely proportional to the mean volume occupied by the polymer. This volume is proportional to the cube of the mean radius of the polymer. As the mean radius of a sphere occupied by a randomly coiled polymer is proportional to the square route of the length of the polymer (Flory, 1953), the probability that two ends of a randomly coiled unidimensional polymer lie near in space is proportional to the length of the polymer raised to the three-halves (or 1·5) power. Thus, given these assumptions, we might expect the probability of an indel of length k will vary with $k^{-1\cdot5}$, remarkably close to the $k^{-1\cdot7}$ dependence observed empirically.

This calculation is appropriate only for an idealized unidimensional polymer, of course. Real polymers fill the second and third dimensions in space orthogonal to the dimension along the polymer chain, giving rise to an excluded volume. The excluded volume of a real polymer chain increases the exponent in the formula relating mean volume to length. This exponent is an experimentally measurable quantity, and depends to some extent on the composition of the polymer. For a typical polypeptide, the volume of a random coil is a function of $(\text{length})^{1\cdot7 \text{ to } 1\cdot8}$ (Brant & Flory, 1965), remarkably close to that observed in the gap length distribution reported here.

Thus, the Zipfian distribution of gap lengths observed here can be explained, both qualitatively and quantitatively, as the consequence of two hypotheses relating to the folded structure of proteins. First, gaps are flanked by amino acid residues that lie close in space in the folded structure of proteins. Second, the insert added or removed in the insertion or deletion event adopts a random coil structure, with the random coil structure behaving much as a free random coil might.

We fully recognize that by treating the insert as an independent folding unit (Flory, 1953), this explanation assigns a greater role to the insert in determining the overall conformation of the polypeptide chain than is generally accepted. This will undoubtedly make this explanation controversial. However, the hypotheses underlying the explanation have proven to be quite useful in predicting *de novo* the folded structure of proteins from a set of aligned sequences of homologous proteins (Crawford *et al.*, 1987; Benner, 1989; Benner & Gerloff, 1991). For example, the folded structure of protein kinase was recently predicted in advance of any crystallographic data (Benner & Gerloff, 1991); the prediction later shown by crystallography to be remarkably accurate (Knighton *et al.*, 1991; Thornton *et al.*, 1991; Benner, 1992). In this prediction, a randomly coiled structure was assigned to any segments that were deleted in any of the homologous proteins, and the alignment parsed at this point. Further, in assembling the predicted secondary structural units into supersecondary structures, the amino acid residues flanking deletions were brought together in space. Thus, the hypotheses used to explain the gap length distribution can be directly applied to the problem of *de novo* prediction of the folded structure of proteins.

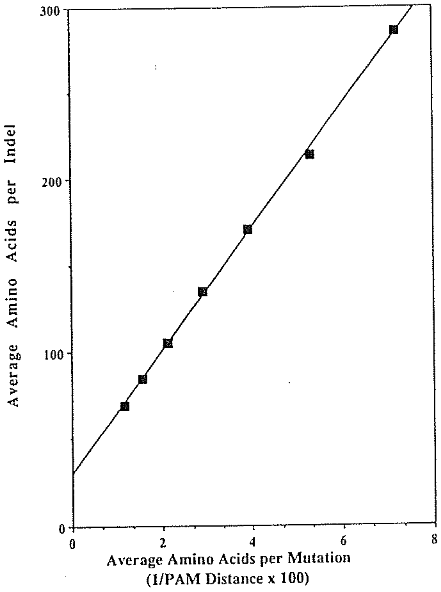

Figure 2. Probability of indel as a function of evolutionary distance. Data from analysis of the Swiss-Prot Version 19 database (Table 1B, PAM bounds 11·8 to 100). (■) Slope = 35·4, intercept 29·8, $R^2 = 0{\cdot}999$. A similar plot for data obtained from the analysis of the MIPS Version 64 database yields slope = 35·7, intercept 25·2, $R^2 = 0{\cdot}997$. Data at PAM distances greater than 10 were selected as they display the smallest variance, and are likely to be suitable for extrapolation to infinite evolutionary distance. Linear fits for the entire data set were also obtained (Swiss-Prot: slope = 31·3, intercept 43·4, $R^2 = 0{\cdot}997$; MIPS: slope = 25·1, intercept 59·4, $R^2 = 0{\cdot}978$).

A linear relationship was observed between the average number of amino acid residues between indel and the reciprocal of PAM distance (multipled by 100, Fig. 2). This choice of axes is not wholly arbitrary, as the reciprocal of PAM distance (the abscissa) is directly proportional to the average number of amino acid residues between mutations in an alignment, while the ordinate describes the average number of amino acid residues between indels.

The plot in Figure 2, containing data collected with lower and upper PAM bounds of 11·8 and 100, displays two notable features. First, it is remarkably linear. Second, as the average distance between residues suffering mutation falls to zero (that is, as evolutionary distance measured in PAM units becomes infinite), the average distance between indels does not also fall to zero. Rather, it appears that segments of proteins, on average 30 to 40 amino acid residues in length, remain undisrupted by indels even after extended periods of divergent evolution.

Such extrapolations are, of course, prone to error and should be treated with caution. For example, Pascarella & Argos (1992) used an analogous extrapolation from a much smaller set of data to draw a different conclusion that the average size of the undisrupted peptide unit is between seven and eight amino acid residues long. Thus, interpretations of the extrapolation presented here should be made with caution. Nevertheless, if we assume that the extrapolation in Figure 2 is accurate, it is worth noting that peptides 30 to 40 amino acids in length are often presumed to be the smallest that can fold to adopt a stable folded structure in aqueous solution (Wetlaufer, 1981; Thomas & Luisi, 1986; Patthy, 1991). For example, pancreatic polypeptide, the smallest naturally occurring peptide that has been demonstrated by crystallography to form a stable folded structure, contains 36 amino acids (Glover *et al.*, 1983). Work in this laboratory has suggested that designed peptides with 32 amino acids form stable structures (Allemann, 1989; Johnsson *et al.*, 1990) as dimers. While a detailed discussion of this hypothesis must focus on individual protein structures and therefore is beyond the scope of this paper, it is intriguing to suggest that during divergent evolution, units important for folding longer than a single helix or strand remain undisrupted by insertion and deletions during divergent evolution.

An analysis of the types of amino acids found flanking indels and within the insert itself (Tables 5 and 6) shows that seven amino acid residues (Ile, Leu, Met, Phe, Trp, Tyr and Val) are strongly underrepresented both within the insert and in regions flanking the insert at all PAM distances. Cys is modestly underrepresented in these regions. Two (Gly and Pro) are strongly overrepresented both within and flanking the insert at all PAM distances. Ser is strikingly overrepresented in the flanking regions. but not within the insert. Asn is normally distributed, except at positions 3 and 4, defined as:

$$\begin{array}{l} \ldots XXX1\_\_\ \ldots\ \_\_2XXX \\ \ldots YYY35ZZZZZ64YYY. \end{array} \qquad (18)$$

Both Ala and Gln are overrepresented in low PAM windows. but are normally distributed in higher PAM windows. The seven remaining amino acid residues (Arg, Asp, Glu, His, Lys, Thr) are all approximately normally distributed at all positions within and flanking the insert, with Arg possibly underrepresented and Asp, Glu and Lys possibly overrepresented.

In examining the data in Tables 5 and 6, it is important to remember (Table 1) that the number of indels and deleted amino acids is far larger in the upper PAM windows than in the lower PAM windows. Nevertheless, the patterns observed are consistent with two structural generalizations. First, coils almost always lie on the surface of globular proteins (Cohen *et al.*, 1986). Therefore, hydrophilic amino acid residues are expected to be overrepresented and hydrophobic amino acid residues underrepresented in positions in and around the inserts, should the inserts adopt coil structures and be flanked by coil structures. Consistent with this generalization, all amino acid residues underrepresented in the region of the indel are hydrophobic, while none of those overrepresented are hydrophobic.

Second, Pro and Gly, amino acids often found in coils, are the most strongly overrepresented amino acids in and flanking the insert. This is, again, consistent with the notion that the insert does not usually adopt a standard secondary structure (α helix or β strand) in the folded protein. The abundance of Ala and Gln in the insert at low PAM distance presumably reflects the fact that several classes of proteins contain repetitive sequences involving these amino acids that undergo facile deletion.

The most striking unexpected results here are the distributions observed for Ser and Asn. The difference between the occurrence of Ser within the insert and in the regions flanking the insert is apparently significant and deserved further examination. Likewise, the overrepresentation of Asn in positions 3 and 4 might suggest hypotheses regarding the role of this residue in forming structures that are susceptible to deletion. Both Ser and Asn are classically regarded as "structure disrupters", as reflected in classical secondary structure prediction heuristics (e.g. the Chou and Fasman heuristic).

It is possible, of course, to use these distributions as part of a scheme for scoring gaps, where an estimate of the probability of a gap is based in part on the amino acid residues flanking and within the insert. This approach has not yet been computationally implemented as a part of an alignment program. However, in protein structure prediction recipes described in detail elsewhere (Benner, 1989; Benner & Gerloff, 1991), Pro and Gly within an insert are said to "confirm" the placement of a gap within an alignment, and this confirmation strengthens the reliability of a "parse" based on this placement.

It is important to note that the results of this study differ in several respects from certain results reported from other studies of indels (Demchuk *et al.*, 1989; Pascarella & Argos, 1992). For example, Demchuk *et al.* (1989) suggest from their analysis of indels that pentapeptides may be fundamental units of protein structure. We do not find evidence for this in our work. Pascarella & Argos (1992) found that indels are slightly longer in alignments of proteins with lower residue pairwise indentities than with proteins having higher residue pairwise identities. We report the opposite trend. They suggested that there might be an upper limit to gap size of approximately five residues. The present study suggests no such limit. They found a rather irregular relation between indel probability and percent residue identity. We find a more regular relationship between indel probability and evolutionary distance measured in PAM units. Several of their findings with respect to the frequencies of various amino acid residues within or flanking the insert differ from those reported here.

We believe that the differences between the conclusions of other authors and those presented here can be accounted for by three factors. First, the database used here is large, with over 16,000 evolutionarily independent indels and no selection bias other than that of the protein sequence database as a whole. The database used by Pascarella & Argos (1992) contained 714 evolutionarily independent indels in protein families represented in the crystal database; the database used by Demchuk *et al.* (1989) is still smaller. Second, we constructed alignments using the more advanced Dayhoff matrices and gap deletion penalties obtained from the exhaustive matching of the protein sequence database (Gonnet *et al.*, 1992). Finally, Pascarella & Argos (1992) measured evolutionary distance using a percent residue identity: PAM distance is used here. Although percent residue identity is a good surrogate for PAM distance for proteins very similar in sequence, it is an inaccurate measure of evolutionary distance at large evolutionary distances. An analysis measuring evolutionary distance in PAM units therefore undoubtedly permits more accurate analysis of trends over the entire range of evolutionary divergence.

The field of protein chemistry presents two challenges: *de novo* prediction of folded structure from sequence data, and *de novo* design of polypeptides that fold in solution and catalyse reactions. Substantial progress has now been made both in the design of proteins (Allemann, 1989; Johnsson *et al.*, 1990; Osterhout *et al.*, 1992) and in structure prediction (Crawford *et al.*, 1987; Benner, 1989; Bazan, 1990; Benner & Gerloff, 1991), and a rigorous model of structural and behavioural evolution in proteins has underlaid this progress (Benner & Ellington, 1990). We expect that further evolutionary analyses will enable still more rapid progress to be made, both in these and other laboratories.

We are indebted to Dr E. Zass for expert literature search services, to Sandoz AG and the Swiss National Science Foundation for support. We are also indebted to the Digital Equipment Corporation for donation of computer equipment. M.A.C. was supported by a fellowship from the Wellcome Trust.

References

Allemann, R. K. (1989). Evolutionary guidance as a tool in organic chemistry. Dissertation, E.T.H. no. 8804.

Altschul, S. F. (1989). Gap cost for multiple sequence alignment. *J. Theoret. Biol.* 138, 297–309.

Arratia, R., Gordon, L. & Waterman, M. S. (1986). An extreme value theory for sequence matching. *Ann. Stat.* 14, 971–993.

Bairoch, A. & Boeckmann, B. (1991). The SWISS-PROT protein sequence data bank. *Nucl. Acids Res.* 19, 2247–2249.

Bairoch, A. & Boeckmann, B. (1992). The SWISS-PROT protein sequence data bank. *Nucl. Acids Res.* 20, 2019–2022.

Bazan, J. F. (1990). Structural design and molecular evolution of a cytokine receptor superfamily. *Proc. Nat. Acad. Sci., U.S.A.* 87, 6934–6938.

Benner, S. A. (1989). Patterns of divergence in homologous proteins as indicators of tertiary and quaternary structure. *Advan. Enzyme Reg.* 28, 219–236.

Benner, S. A. (1992). Predicting *de novo* the folded structure of proteins. *Curr. Opin. Struct. Biol.* 2, 402–412.

Benner, S. A. & Ellington, A. D. (1988). Interpreting the behavior of enzymes: purpose or pedigree? *CRC Crit. Rev. Biochem.* 23, 369–426.

Benner, S. A. & Ellington, A. D. (1990). Evolution and structural theory: the frontier between chemistry and biochemistry. *Bioorgan. Chem. Front.* 1, 1–70.

Benner, S. A. & Gerloff, D. (1991). Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: a prediction of the structure of the catalytic domain of protein kinases. *Advan. Enzyme Reg.* 31, 121–181.

Benner, S. A., Ellington, A. D. & Tauer, A. (1989). Modern metabolism as a palimpsest of the RNA world. *Proc. Nat. Acad. Sci., U.S.A.* 86, 7054–7058.

Blundell, T. L., Sibanda, B. L., Sternberg, M. J. E. & Thornton, J. M. (1987). Knowledge-based prediction of protein structures and the design of novel molecules. *Nature (London)*, 326, 347–352.

Brant, D. A. & Flory, P. J. (1965). The configuration of random polypeptide chains. *J. Amer. Chem. Soc.* 87, 2788.

Cohen, F. E., Abarbanel, R. M., Kuntz, I. D. & Fletterick, R. J. (1986). Turn predictions in proteins using a pattern-matching approach. *Biochemistry*, 25, 266–275.

Crawford, I. P., Niermann, T. & Kirschner, K. (1987). Prediction of secondary structure by evolutionary comparison: application to the α subunit of tryptophan synthase. *Proteins*, 2, 118–129.

Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. (1978). A model for evolutionary change in proteins. In *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., ed.), vol. 5, suppl. 3, p. 345, National Biomedical Research Foundation, Washington, DC.

Demchuk, E. J., Esipova, N. G. & Tymanyan, V. G. (1989). Regularities in arrangement and evolution of protein primary structures connected with deletions–insertions found upon statistical analysis of data banks. *Studia Biophys.* 129, 193–199.

de Vos, A. M., Ultsch, M. & Kossiakoff, A. A. (1992). Human growth hormone and extracellular domain of its receptor: crystal structure of the complex. *Science*, 255, 306–312.

Doolittle, R. F. (1990). Molecular evolution: computer analysis of protein and nucleic acid sequences. *Methods Enzymol.* 183, 1–736.

Doolittle, R. F. (1991). Counting and discounting the universe of exons. *Science*, 253, 677–679.

Dorit, R. L., Schoenbach, L. & Gilbert, W. (1990). How big is the universe of exons? *Science*, 250, 1377–1382.

Edwards, F. W. & Cavalli-Sforza, L. L. (1963). The reconstruction of evolution. *Annu. Human. Genet.* 27, 104.

Feng, D. F. & Doolittle, R. F. (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.* 25, 351–360.

Feng, D. F., Johnson, M. S. & Doolittle, R. F. (1985). Aligning amino acid sequences: comparison of commonly used methods. *J. Mol. Evol.* 21, 112–125.

Fitch, W. & Margoliash, E. (1967). Construction of phylogenetic trees. *Science,* 155, 279–284.

Fitch, W. & Smith, T. F. (1983). Optimal sequence alignment. *Proc. Nat. Acad. Sci., U.S.A.* 80, 1382–1386.

Flory, P. (1953). *Principles of Polymer Chemistry,* Cornell University Press, Ithaca, NY.

Freund, J. E. (1971). *Mathematical Statistics,* Prentice Hall, Englewood Cliffs, NJ.

Glover, I., Haneef, I., Pitts, J., Wood, S., Moss, D., Tickle, I. & Blundell, T. (1983). Conformational flexibility in a small globular hormone: X-ray analysis of avian pancreatic polypeptide at 0·98 Å resolution. *Biopolymers,* 22, 293–304.

Gonnet, G. H. & Baeza-Yates, R. (1991). *Handbook of Algorithms and Data Structures,* 2nd edit., Addison-Wesley, New York.

Gonnet, G. H. & Benner, S. A. (1991). Computational biochemistry research at ETH. *Technical Report 154, Departement Informatik,* E.T.H., Zurich.

Gonnet, G. H., Cohen, M. A. & Benner, S. A. (1992). Exhaustive matching of the entire protein sequence database. *Science,* 256, 1443–1445.

Hyde, C. C., Ahmed, S. A., Padlan, E. A.. Miles, E. W. & Davies, D. R. (1988). Three-dimensional structure of the tryptophan synthase alpha$_2$beta$_2$ multienzyme complex from *Salmonella typhimurium. J. Biol. Chem.* 263, 17857–17871.

Johnsson, K., Allemann, R. K. & Benner. S. A. (1990). Designed enzymes: new peptides that fold in aqueous solution and catalyse reactions. In *Molecular Mechanisms in Bioorganic Processes* (Bleasdale, C. & Golding, B. T., eds), pp. 166–187, Royal Society of Chemistry, Cambridge, U.K.

Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *CABIOS,* 8, 275–282.

Knighton, D. R., Zheng, J., Eyck, L. F. T., Ashford, V. A., Xuong, N.-H., Taylor, S. S. & Sowadski, J. M. (1991). Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. *Science,* 253, 407–414.

Needleman, S. B. & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48, 443–453.

Niermann, T. & Kirschner, K. (1990). Improving the prediction of secondary structure of "TIM-barrel" enzymes. *Protein Eng.* 4, 137–147.

Oliver, S. G. *et al.* (1992). The complete DNA sequence of yeast chromosome III. *Nature (London),* 357, 38–46.

Osterhout, J. J., Jr, Handel, T., Na, G., Toumadje, A., Long, R. C., Connolly, P. J., Hoch, J. C., Johnson, W. C., Jr, Live, D. & DeGrado, W. F. (1992). Characterization of the structural properties of α_1B, a peptide designed to form a four-helix bundle. *J. Amer. Chem. Soc.* 114, 331–337.

Pascarella, S. & Argos, P. (1992). Analysis of insertions/deletions in protein structures. *J. Mol. Biol.* 224, 461–471.

Patthy, L. (1991). Exons – original building blocks of proteins? *Bioessays,* 13, 187–192.

Sankoff, D. & Kruskal, J. B. (1983). Time warps, string edits, and macromolecules. In *The Theory and Practice of Sequence Comparison,* Addison-Wesley, Reading, MA.

Sellers, P. (1974). On the theory and computation of evolutionary distances. *SIAM J. Appl. Math.* 26. 787–793.

Sulston, J., Du, Z., Thomas, K., Wilson, R., Hillier, L.. Staden, R., Halloran, N., Green, P., Thierry-Mieg, J., Qiu, L., Dear, S., Coulson, A., Craxton, M., Durbin, R., Berks, M., Metzstein, M., Hawkins, T.. Ainscough, R. & Waterston, R. (1992). The *C. elegans* genome sequencing project: a beginning. *Nature (London),* 356, 37–41.

Taylor, W. R. (1990). Hierarchical method to align large numbers of biological sequences. *Methods Enzymol.* 183, 456.

Thomas, R. M. & Luisi, P.-L. (1986). Protein fragment studies: implications for protein folding and protein design. *Gazz. Chim. Ital.* 116, 617–621.

Thorne, J. L., Kishino, H. & Felsenstein, J. (1992). Inching toward reality: an improved likelihood model of sequence evolution. *Mol. Biol. Evol.* 34, 3–16.

Thornton, J. M., Flores, T. P., Jones, D. T. & Swindells, M. B. (1991). Prediction of progress at last. *Nature (London),* 354, 105–106.

Wetlaufer, D. B. (1981). Folding of protein fragments. *Advan. Protein Chem.* 34, 61–92.

Zuckerkandl, E. & Pauling, L. (1965). Evolutionary divergence and convergence in proteins. In *Evolving Genes and Proteins* (Bryson, V. & Vogel, H. J., eds), pp. 97–166, Academic Press, New York.

Edited by F. Cohen