J. Mol. Evol. 9, 121–130 (1977)

Journal of
Molecular Evolution

# Alignment Statistic for Identifying Related Protein Sequences

G. William Moore[1] and Morris Goodman[2]

[1] Department of Pathology, The Johns Hopkins Hospital,
Baltimore, Md. 21205, USA

[2] Department of Anatomy, Wayne State University, School of Medicine,
540 E. Canfield Ave., Detroit, Michigan 48201, USA

**Summary.** Closely related proteins show an obvious kinship by having numerous matching amino acids in their aligned sequences. Kinship between anciently separated proteins requires a statistical evaluation to rule out fortuitous similarities. A simple statistic is developed which assumes equal probability for all codon pairs, and a table of critical values for amino acid sequence alignments of length 200 or less is presented. Applying this statistic to *V* and *C* regions of immunoglobulin chains, aligned on the basis of shared features of three-dimensional structure, provides evidence that the *V* and *C* sequences descended from a common ancestor. Similarly the distant evolutionary relationship of dehydrogenases, flavdoxin, and subtilisin, suggested by structural alignments, is verified. On the other hand, the statistic does not verify a common evolutionary origin for the heme binding pocket in globins and cytochrome $b_5$. Empirical evidence from the distribution of MMD values of amino acid pairs in comparisons of misaligned polypeptide chains and from Monte Carlo trials of sequences aligned with arbitrary gaps supports the validity of the statistic.

**Key words:** Structural alignments – Minimum mutation distance – Evolutionary relationship – Significance test – *V* and *C* immunoglobulin sequences – Dehydrogenases.

## 1. Introduction

Proteins recently separated from a common evolutionary progenitor show kinship by having numerous matching amino acids in their sequence structures. Anciently separated proteins usually do not reveal their kinship so easily, as many substitutions, insertions, and deletions of amino acids have accumulated in them. Sensitive criteria are needed to detect any sequence homologies which remain and to decide if they are of sufficient scope to support the hypothesis of descent from a common ancestor. We develop here an alignment statistic which successfully identifies such distant relationships. We present this statistic in the form of a significance table for pairs of aligned amino acid sequence chains ranging from one to two hundred amino acid positions. We then use the table to illustrate a major application in the study of protein evolution. This application is to test sequence alignments proposed for functionally different protein chains, or portions of the chains, from structural similarities revealed by x-ray crystallography. In this regard, we find that the structural alignments proposed by Poljak et al. (1974) for the two kinds of regions in the antibody molecule named variable *(V)* and constant *(C)* show true sequence homology as judged by our significance table. This is also the case for stretches of dehydrogenases, flavodoxin, and subtilisin aligned on structural grounds by Rossman et al. (1974). In contrast the equivocal suggestion of homology on comparing the heme binding pocket in globins and cytochrome $b_5$ (Rossman and Argos, 1975) is not supported by the alignment statistic.

Our test for common ancestry in an amino acid sequence alignment joins a number of others currently in the literature. Fitch (1970, 1975) and Jukes and Cantor (1969) determine the minimum mutation distance for each subalignment of a given length (say, 30) within the alignment as a whole, and compare this to expectations based upon random rearrangements of the same amino acid composition. Needleman and Wunsch (1970) determine the amino acid difference (i.e., number of nonidentical amino acids) in an entire alignment, and compare this to expectation. Sankoff (1972) and Barker and Dayhoff (1972) use modified versions of the Needleman and Wunsch method. Haber and Koshland (1970) employ a functional approach: they determine the number of identical amino acid pairs and electrochemically similar amino acid pairs in an alignment, and compare this to random expectation. McLachlan (1971, 1972) has further developed this functional approach. Our method employs the minimum mutation distance for the alignment as a whole, and compares this to expectations based upon an equal and independent probability of each nucleotide pair.

## 2. The Alignment Statistic

In the significance test to determine whether an aligned pair of sequences of length $= n$ (hereinafter denoted, "$n$-alignment") share a common ancestry, we wish to distinguish between the *null hypothesis* (absence of common ancestry) and the most plausible *alternative hypothesis* (presence of common ancestry). If an $n$-alignment satisfies the null hypothesis, then we expect its *minimum mutation distance* (Fitch and Margoliash, 1967) or MMD, to belong to a random collection of $n$-alignments. Alternatively, if the $n$-alignment has a sufficiently recent common ancestry, then its MMD would be smaller than expected from the null hypothesis, because of the inherited matched alignment positions. The null hypothesis is rejected for an $n$-alignment if the probability that that n-alignment was generated randomly is less than an arbitrary cutoff point $\alpha$ ($\alpha = 0.05, 0.01$, etc.).

For a random 1-alignment (consisting of a single codon pair), there are 61 possible non-terminating codons for one sequence and 61 for the other, 61 x 61 = 3721 possible codon pairs all told. We are not assuming that nucleotide pair alignments are equiprobable in DNA known to share a common ancestry, and indeed such has been called into question by the empirical studies of Barker and Dayhoff (1972). Here we are setting up the *null hypothesis,* for the case in which common ancestry is assumed to be *absent.* For a random $n$-alignment, there are $3721^n$ possible sequence pairs. In generating a random distribution we assume that each of the $3721^n$ $n$-alignments has equal probability of occurrence, namely $\frac{1}{3721^n}$ apiece. Some of the $n$-alignments are perfectly matched, some have a single nucleotide difference, . . ., and some are perfectly mismatched (i.e., with $3^n$ nucleotide differences). The probability that an $n$-alignment has exactly $i$ observed nucleotide differences is denoted $p\ (i,\ n)$.

Under those conditions, we find in the 1-alignment problem, 61 occurrences of perfectly matched alignments, 526 occurrences of alignments with a single nucleotide difference, 1568 occurrences with 2 differences, and 1566 occurrences with 3 differences; 61 + 526 + 1568 + 1566 = 3721 occurrences all told. We say that the probability of actual difference = 0 is $\frac{61}{3721} = 0.01639$, probability of 1 is $\frac{526}{3721} = 0.14136$, probability of 2 is $\frac{1568}{3721} = 0.42139$, and probability of 3 is $\frac{1566}{3721} = 0.42085$.

Since our observations consist of amino acid pairs, we will tend to underestimate the true nucleotide difference. For example, codons AUU and AUA have an actual distance = 1, but since both codons specify the same amino acid (isoleucine), at the amino acid level the observed distance between them is 0. The observed difference between a pair of codons is obtained by mapping each codon into its corresponding amino acid by means of the genetic code, and then finding the *minimum mutation distance* (MMD) for that pair of amino acids (Table 1). The MMD for a pair of amino acids is the minimum number of nucleotide steps necessary to convert one amino acid into the other. By examining the 3721 possible 1-alignments in light of *observed differences,* we find the probability of observed difference = 0 is $\frac{235}{3721} = 0.06316$, proba-

**Table 1.** The observed difference between a pair of codons is obtained by mapping each non-terminating codon into its corresponding amino acid, and then finding the MMD for that pair of amino acids.

| | Codon Pair | | Amino Acid Pair | | Actual Difference | Observed Difference |
|---|---|---|---|---|---|---|
| 1. | AAA | AAA | LYS | LYS | 0 | 0 |
| 2. | AAA | AAC | LYS | ASN | 1 | 1 |
| 3. | AAA | AAG | LYS | LYS | 1 | 0 |
| 4. | AAA | AAU | LYS | ASN | 1 | 1 |
| 5. | AAA | ACA | LYS | THR | 1 | 1 |
| 6. | AAA | ACC | LYS | THR | 2 | 1 |
| . | | | | | | |
| . | | | | | | |
| . | | | | | | |
| 3721. | UUU | UUU | PHE | PHE | 0 | 0 |

bility of 1 is $\frac{1706}{3721} = 0.45848$, probability of 2 is $\frac{1698}{3721} = 0.45633$, and probability of 3 is $\frac{82}{3721} = 0.02204$. Note that we are not altering the actual probability, say, of a 0-nucleotide difference (it remains at 61/3721); we are merely stating that an additional 174 codon comparisons which actually have higher difference values are *observed* as having 0 differences. In other words, our test compensates for the fact that the full nucleotide difference implicit in certain amino acid pairs is not actually seen in the data. This blunting effect of the MMD statistic tends to render our test more conservative – some alignments which are actually similar by reason of common ancestry will be missed by our statistic.

For $n = 1$, we have already shown that $p\,(0,1) = 0.06316$, $p\,(1,1) = 0.45848$, $p\,(2,1) = 0.45633$, and $p\,(3,1) = 0.02204$. For $n = 2$, it is possible to calculate $p\,(j,2)$ by listing out all possible 2-alignments ($3721^2 = 13,845,841$ all told), and tabulating how many have observed difference = 0, how many have observed difference = 1, etc. There is a *recursion formula* which allows us to build up a table of $p$'s in a stepwise fashion. In a recursion formula, if we know the answer for $n = 1$ (which we do), we can get the answer for $n = 2$; if we know $n = 2$, we can get $n = 3$; etc. To calculate $p\,(i,\,n)$, we separate the n-alignment into a *first part* of length $= n - 1$ and a *second part* of length = 1. There are only four possible ways that the $n$-alignment could have observed difference $= i$: (a) if the first part has difference $i$ and the second part has difference 0; (b) first part $i - 1$, second part 1; (c) first part $i - 2$, second part 2; (d) first part $i - 3$, second part 3. The second part cannot have length greater than 3 because it is a single codon pair. Thus we calculate (Table 2); $p(i,\,n) = 0.06316 \cdot p(i,\,n - 1) + 0.45848 \cdot p(i - 1,\,n - 1) + 0.45633 \cdot p(i - 2,\,n - 1) + 0.02204 \cdot p(i - 3,\,n - 1)$. For example, the probability of obtaining 3 differences from a 2-alignment ($i = 3$, $n = 2$) is $p(3,2) = 0.06316 \cdot 0.02204 + 0.45848 \cdot 0.45633 + 0.45633 \cdot 0.45848 + 0.02204 \cdot 0.06316 = 0.42122$.

Each row in Table 2 can be expressed as a histogram, which establishes the cutoff point for the significance test. Figure 1a shows the histogram for $n = 1$; Figure 1b shows the histogram for $n = 4$. When we state that the 5 % significance level for a 4-alignment is 3, we mean that there is *less than* a 5 % chance (precisely, a 4.035 % chance) that a 4-alignment will have 3 or fewer nucleotide differences. The cutoff point 3 is established by cutting the histogram for the 4-alignment (Fig. 1b) at the right most point such that the left tail of the histogram sums to less than 5 %. For any significance level $\alpha$ and any $n$-alignment, the cutoff point $m$ is established by

**Table 2.** Recursive evaluation of $p(i,\,n)$ by the formula
$p(i,\,n) = 0.06316 \cdot p(i,\,n - 1) + 0.45848 \cdot p(i - 1,\,n - 1) + 0.45633 \cdot p(i - 2,\,n - 1) + 0.02204 \cdot p(i - 3,\,n - 1)$.
$p\,(i,\,n)$ is the probability that an $n$-alignment has exactly $i$ observed nucleotide differences

| $n$ | $p(0,n)$ | $p(1,n)$ | $p(2,n)$ | $p(3,n)$ | $p(4,n)$ | $p(5,n)$ | $p(6,n)$ | |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.06316 | 0.45848 | 0.45633 | 0.02204 | – | – | – | |
| 2 | 0.00399 | 0.05791 | 0.26784 | 0.42122 | 0.22844 | 0.02011 | 0.00049 | |
| 3 | 0.00025 | 0.00549 | 0.04529 | 0.17592 | 0.33105 | 0.30412 | 0.12278 | etc. |
| 4 | 0.00002 | 0.00046 | 0.00549 | 0.03438 | 0.12235 | 0.25226 | 0.30213 | etc. |

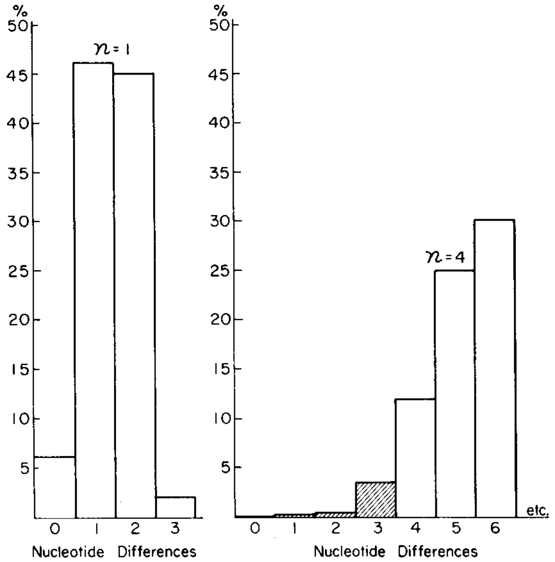

**Fig. 1.** **a** The significance histogram for $n = 1$. **b** The significance histogram for $n = 4$. The hatched area in this histogram represents the maximum part of the left tail of the histogram which sums to less than 5 %

cutting the $n$-alignment histogram at the rightmost (i.e., maximum) point such that the left tail of the histogram (i.e., all histogram bars up to and including $m$) sums to less than $\alpha$. This is achieved by maximizing $m$, where

$$\sum_{i=0}^{m} p(i, n) < \alpha$$

Table 3 shows the maximum observed difference for $n$ between 1 and 200 at selected significance levels ($\alpha = 0.95, 0.50, 0.05, 0.01, 0.001, 0.0001$).

## 3. Structural Alignments

*V and C Sequences.* Each different kind of immunoglobulin chain can be divided into $V$ and $C$ regions which are thought to share a remote common ancestry. Strong sequence homology exists among the different $V$ regions, also among the different $C$ regions (Jukes and Cantor, 1969; Barker and Dayhoff, 1972). Homology between $V$ and $C$ sequences, however, has not been obvious. The assumption of a remote ancestry between them rests on such features as similar location of disulfide-bonded cysteine residues and similar chain folding patterns. Poljak et al. (1974) have now used the common three-dimensional structural features revealed by x-ray crystallography at $2.0\text{A}^{\circ}$ resolutions for two $V$ and two $C$ regions in a particularly well studied human myeloma immunoglobulin to align these $V$ and $C$ sequences. We find significant homology between them by our alignment statistic as shown in Table 4. To our knowledge this is the first clear statistical evidence from sequence data for the hypothesis of common genetic ancestry of $V$ and $C$ regions.

An immunoglobulin molecule of the IgG class contains two identical light chains, each about 220 amino acids long, and two identical heavy chains, each about 440 amino

**Table 3.** Critical values for *n*-alignments where $n \leqslant 200$. At each significance level, the greatest observed MMD at which the null hypothesis of no common ancestry can still be rejected is given. For example, for a 14-alignment, the null hypothesis can be rejected at the 5 % level for an observed MMD of 15 or less.

| N | .01% | .1% | 1% | 5% | 10% | 50% | 95% |
|---|---|---|---|---|---|---|---|
| 1 | -1 | -1 | -1 | -1 | 0 | 0 | 1 |
| 2 | -1 | -1 | 0 | 0 | 1 | 2 | 3 |
| 3 | -1 | 0 | 1 | 1 | 2 | 3 | 5 |
| 4 | 0 | 1 | 2 | 3 | 3 | 5 | 7 |
| 5 | 1 | 2 | 3 | 4 | 4 | 6 | 8 |
| 6 | 2 | 3 | 4 | 5 | 6 | 8 | 10 |
| 7 | 3 | 4 | 5 | 6 | 7 | 9 | 12 |
| 8 | 4 | 5 | 6 | 7 | 8 | 11 | 13 |
| 9 | 5 | 6 | 7 | 9 | 9 | 12 | 15 |
| 10 | 6 | 7 | 9 | 10 | 11 | 13 | 17 |
| 11 | 7 | 8 | 10 | 11 | 12 | 15 | 18 |
| 12 | 8 | 9 | 11 | 13 | 13 | 16 | 20 |
| 13 | 9 | 10 | 12 | 14 | 15 | 18 | 21 |
| 14 | 10 | 12 | 13 | 15 | 16 | 19 | 23 |
| 15 | 11 | 13 | 15 | 16 | 17 | 21 | 25 |
| 16 | 12 | 14 | 16 | 18 | 19 | 22 | 26 |
| 17 | 13 | 15 | 17 | 19 | 20 | 23 | 28 |
| 18 | 14 | 16 | 18 | 20 | 21 | 25 | 29 |
| 19 | 16 | 17 | 20 | 22 | 23 | 26 | 31 |
| 20 | 17 | 19 | 21 | 23 | 24 | 28 | 32 |
| 21 | 18 | 20 | 22 | 24 | 25 | 29 | 34 |
| 22 | 19 | 21 | 24 | 26 | 27 | 31 | 36 |
| 23 | 20 | 22 | 25 | 27 | 28 | 32 | 37 |
| 24 | 21 | 24 | 26 | 28 | 29 | 34 | 39 |
| 25 | 23 | 25 | 27 | 30 | 31 | 35 | 40 |
| 26 | 24 | 26 | 29 | 31 | 32 | 36 | 42 |
| 27 | 25 | 27 | 30 | 32 | 34 | 38 | 43 |
| 28 | 26 | 29 | 31 | 34 | 35 | 39 | 45 |
| 29 | 27 | 30 | 33 | 35 | 36 | 41 | 46 |
| 30 | 29 | 31 | 34 | 36 | 38 | 42 | 48 |
| 31 | 30 | 32 | 35 | 38 | 39 | 44 | 49 |
| 32 | 31 | 34 | 36 | 39 | 40 | 45 | 51 |
| 33 | 32 | 35 | 38 | 40 | 42 | 46 | 52 |
| 34 | 33 | 36 | 39 | 42 | 43 | 48 | 54 |
| 35 | 35 | 37 | 40 | 43 | 44 | 49 | 56 |
| 36 | 36 | 39 | 42 | 44 | 46 | 51 | 57 |
| 37 | 37 | 40 | 43 | 46 | 47 | 52 | 59 |
| 38 | 38 | 41 | 44 | 47 | 49 | 54 | 60 |
| 39 | 40 | 42 | 46 | 48 | 50 | 55 | 62 |
| 40 | 41 | 44 | 47 | 50 | 51 | 57 | 63 |
| 41 | 42 | 45 | 48 | 51 | 53 | 58 | 65 |
| 42 | 43 | 46 | 50 | 52 | 54 | 59 | 66 |
| 43 | 44 | 48 | 51 | 54 | 55 | 61 | 68 |
| 44 | 46 | 49 | 52 | 55 | 57 | 62 | 69 |
| 45 | 47 | 50 | 54 | 57 | 58 | 64 | 71 |
| 46 | 48 | 51 | 55 | 58 | 59 | 65 | 72 |
| 47 | 49 | 53 | 56 | 59 | 61 | 67 | 74 |
| 48 | 51 | 54 | 58 | 61 | 62 | 68 | 75 |
| 49 | 52 | 55 | 59 | 62 | 64 | 69 | 77 |
| 50 | 53 | 57 | 60 | 63 | 65 | 71 | 78 |
| 51 | 54 | 58 | 61 | 65 | 66 | 72 | 80 |
| 52 | 56 | 59 | 63 | 66 | 68 | 74 | 81 |
| 53 | 57 | 60 | 64 | 67 | 69 | 75 | 83 |
| 54 | 58 | 62 | 65 | 69 | 71 | 77 | 84 |
| 55 | 59 | 63 | 67 | 70 | 72 | 78 | 86 |
| 56 | 61 | 64 | 68 | 72 | 73 | 80 | 87 |
| 57 | 62 | 66 | 70 | 73 | 75 | 81 | 89 |
| 58 | 63 | 67 | 71 | 74 | 76 | 82 | 90 |
| 59 | 65 | 68 | 72 | 76 | 77 | 84 | 92 |
| 60 | 66 | 70 | 74 | 77 | 79 | 85 | 93 |
| 61 | 67 | 71 | 75 | 78 | 80 | 87 | 95 |
| 62 | 68 | 72 | 76 | 80 | 82 | 88 | 96 |
| 63 | 70 | 73 | 78 | 81 | 83 | 90 | 98 |
| 64 | 71 | 75 | 79 | 82 | 84 | 91 | 99 |
| 65 | 72 | 76 | 80 | 84 | 86 | 92 | 101 |
| 66 | 73 | 77 | 82 | 85 | 87 | 94 | 102 |
| 67 | 75 | 79 | 83 | 87 | 89 | 95 | 104 |
| 68 | 76 | 80 | 84 | 88 | 90 | 97 | 105 |
| 69 | 77 | 81 | 86 | 89 | 91 | 98 | 107 |
| 70 | 78 | 83 | 87 | 91 | 93 | 100 | 108 |
| 71 | 80 | 84 | 88 | 92 | 94 | 101 | 110 |
| 72 | 81 | 85 | 90 | 93 | 95 | 103 | 111 |
| 73 | 82 | 87 | 91 | 95 | 97 | 104 | 113 |
| 74 | 84 | 88 | 92 | 96 | 98 | 105 | 114 |
| 75 | 85 | 89 | 94 | 98 | 100 | 107 | 116 |
| 76 | 86 | 91 | 95 | 99 | 101 | 108 | 117 |
| 77 | 87 | 92 | 96 | 100 | 102 | 110 | 119 |
| 78 | 89 | 93 | 98 | 102 | 104 | 111 | 120 |
| 79 | 90 | 95 | 99 | 103 | 105 | 113 | 122 |
| 80 | 91 | 96 | 100 | 104 | 107 | 114 | 123 |
| 81 | 92 | 97 | 102 | 106 | 108 | 115 | 125 |
| 82 | 94 | 99 | 103 | 107 | 109 | 117 | 126 |
| 83 | 95 | 100 | 105 | 109 | 111 | 118 | 128 |
| 84 | 96 | 101 | 106 | 110 | 112 | 120 | 129 |
| 85 | 98 | 103 | 107 | 111 | 114 | 121 | 131 |
| 86 | 99 | 104 | 109 | 113 | 115 | 123 | 132 |
| 87 | 100 | 105 | 110 | 114 | 116 | 124 | 134 |
| 88 | 101 | 107 | 111 | 115 | 118 | 126 | 135 |
| 89 | 103 | 108 | 113 | 117 | 119 | 127 | 137 |
| 90 | 104 | 109 | 114 | 118 | 120 | 128 | 138 |
| 91 | 105 | 110 | 115 | 120 | 122 | 130 | 140 |
| 92 | 107 | 112 | 117 | 121 | 123 | 131 | 141 |
| 93 | 108 | 113 | 118 | 122 | 125 | 133 | 143 |
| 94 | 109 | 114 | 119 | 124 | 126 | 134 | 144 |
| 95 | 111 | 116 | 121 | 125 | 127 | 136 | 146 |
| 96 | 112 | 117 | 122 | 127 | 129 | 137 | 147 |
| 97 | 113 | 118 | 124 | 128 | 130 | 138 | 149 |
| 98 | 114 | 120 | 125 | 129 | 132 | 140 | 150 |
| 99 | 116 | 121 | 126 | 131 | 133 | 141 | 152 |
| 100 | 117 | 122 | 128 | 132 | 134 | 143 | 153 |
| 101 | 118 | 124 | 129 | 133 | 136 | 144 | 155 |
| 102 | 120 | 125 | 130 | 135 | 137 | 146 | 156 |
| 103 | 121 | 127 | 132 | 136 | 139 | 147 | 158 |
| 104 | 122 | 128 | 133 | 138 | 140 | 148 | 159 |
| 105 | 124 | 129 | 134 | 139 | 141 | 150 | 161 |
| 106 | 125 | 131 | 136 | 140 | 143 | 151 | 162 |
| 107 | 126 | 132 | 137 | 142 | 144 | 153 | 164 |
| 108 | 127 | 133 | 139 | 143 | 146 | 154 | 165 |
| 109 | 129 | 135 | 140 | 145 | 147 | 156 | 167 |
| 110 | 130 | 136 | 141 | 146 | 148 | 157 | 168 |
| 111 | 131 | 137 | 143 | 147 | 150 | 159 | 170 |
| 112 | 133 | 139 | 144 | 149 | 151 | 160 | 171 |
| 113 | 134 | 140 | 145 | 150 | 153 | 161 | 173 |
| 114 | 135 | 141 | 147 | 151 | 154 | 163 | 174 |
| 115 | 137 | 143 | 148 | 153 | 155 | 164 | 176 |
| 116 | 138 | 144 | 149 | 154 | 157 | 166 | 177 |
| 117 | 139 | 145 | 151 | 156 | 158 | 167 | 179 |
| 118 | 140 | 147 | 152 | 157 | 160 | 169 | 180 |
| 119 | 142 | 148 | 154 | 158 | 161 | 170 | 182 |
| 120 | 143 | 149 | 155 | 160 | 162 | 171 | 183 |
| 121 | 144 | 151 | 156 | 161 | 164 | 173 | 185 |
| 122 | 146 | 152 | 158 | 163 | 165 | 174 | 186 |
| 123 | 147 | 153 | 159 | 164 | 167 | 176 | 188 |
| 124 | 148 | 155 | 160 | 165 | 168 | 177 | 189 |
| 125 | 150 | 156 | 162 | 167 | 169 | 179 | 190 |
| 126 | 151 | 157 | 163 | 168 | 171 | 180 | 192 |
| 127 | 152 | 159 | 165 | 170 | 172 | 182 | 193 |
| 128 | 153 | 160 | 166 | 171 | 174 | 183 | 195 |
| 129 | 155 | 161 | 167 | 172 | 175 | 184 | 196 |
| 130 | 156 | 163 | 169 | 174 | 176 | 186 | 198 |
| 131 | 157 | 164 | 170 | 175 | 178 | 187 | 199 |
| 132 | 159 | 165 | 171 | 176 | 179 | 189 | 201 |
| 133 | 160 | 167 | 173 | 178 | 181 | 190 | 202 |
| 134 | 161 | 168 | 174 | 179 | 182 | 192 | 204 |
| 135 | 163 | 170 | 175 | 181 | 183 | 193 | 205 |
| 136 | 164 | 171 | 177 | 182 | 185 | 194 | 207 |
| 137 | 165 | 172 | 178 | 183 | 186 | 196 | 208 |
| 138 | 167 | 174 | 180 | 185 | 188 | 197 | 210 |
| 139 | 168 | 175 | 181 | 186 | 189 | 199 | 211 |
| 140 | 169 | 176 | 182 | 188 | 190 | 200 | 213 |
| 141 | 170 | 178 | 184 | 189 | 192 | 202 | 214 |
| 142 | 172 | 179 | 185 | 190 | 193 | 203 | 216 |
| 143 | 173 | 180 | 186 | 192 | 195 | 205 | 217 |
| 144 | 174 | 182 | 188 | 193 | 196 | 206 | 219 |
| 145 | 176 | 183 | 189 | 195 | 197 | 207 | 220 |
| 146 | 177 | 184 | 191 | 196 | 199 | 209 | 222 |
| 147 | 178 | 186 | 192 | 197 | 200 | 210 | 223 |
| 148 | 180 | 187 | 193 | 199 | 202 | 212 | 225 |
| 149 | 181 | 188 | 195 | 200 | 203 | 213 | 226 |
| 150 | 182 | 190 | 196 | 202 | 204 | 215 | 228 |
| 151 | 184 | 191 | 197 | 203 | 206 | 216 | 229 |
| 152 | 185 | 193 | 199 | 204 | 207 | 217 | 230 |
| 153 | 186 | 194 | 200 | 206 | 209 | 219 | 232 |
| 154 | 187 | 195 | 202 | 207 | 210 | 220 | 233 |
| 155 | 189 | 197 | 203 | 209 | 211 | 222 | 235 |
| 156 | 190 | 198 | 204 | 210 | 213 | 223 | 236 |
| 157 | 191 | 199 | 206 | 211 | 214 | 225 | 238 |
| 158 | 193 | 201 | 207 | 213 | 216 | 226 | 239 |
| 159 | 194 | 202 | 208 | 214 | 217 | 228 | 241 |
| 160 | 195 | 203 | 210 | 215 | 218 | 229 | 242 |
| 161 | 197 | 205 | 211 | 217 | 220 | 230 | 244 |
| 162 | 198 | 206 | 213 | 218 | 221 | 232 | 245 |
| 163 | 199 | 207 | 214 | 220 | 223 | 233 | 247 |
| 164 | 200 | 209 | 215 | 221 | 224 | 235 | 248 |
| 165 | 202 | 210 | 217 | 222 | 226 | 236 | 250 |
| 166 | 203 | 212 | 218 | 224 | 227 | 238 | 251 |
| 167 | 204 | 213 | 220 | 225 | 228 | 239 | 253 |
| 168 | 206 | 214 | 221 | 227 | 230 | 240 | 254 |
| 169 | 207 | 216 | 222 | 228 | 231 | 242 | 256 |
| 170 | 208 | 217 | 224 | 229 | 233 | 243 | 257 |
| 171 | 210 | 218 | 225 | 231 | 234 | 245 | 259 |
| 172 | 211 | 220 | 226 | 232 | 235 | 246 | 260 |
| 173 | 212 | 221 | 228 | 234 | 237 | 248 | 262 |
| 174 | 213 | 222 | 229 | 235 | 238 | 249 | 263 |
| 175 | 215 | 224 | 231 | 236 | 240 | 251 | 265 |
| 176 | 216 | 225 | 232 | 238 | 241 | 252 | 266 |
| 177 | 217 | 226 | 233 | 239 | 242 | 253 | 267 |
| 178 | 219 | 228 | 235 | 241 | 244 | 255 | 269 |
| 179 | 220 | 229 | 236 | 242 | 245 | 256 | 270 |
| 180 | 221 | 231 | 237 | 243 | 247 | 258 | 272 |
| 181 | 222 | 232 | 239 | 245 | 248 | 259 | 273 |
| 182 | 224 | 233 | 240 | 246 | 249 | 261 | 275 |
| 183 | 225 | 235 | 242 | 248 | 251 | 262 | 276 |
| 184 | 226 | 236 | 243 | 249 | 252 | 263 | 278 |
| 185 | 228 | 237 | 244 | 250 | 254 | 265 | 279 |
| 186 | 229 | 239 | 246 | 252 | 255 | 266 | 281 |
| 187 | 230 | 240 | 247 | 253 | 256 | 268 | 282 |
| 188 | 231 | 241 | 248 | 255 | 258 | 269 | 284 |
| 189 | 233 | 243 | 250 | 256 | 259 | 271 | 285 |
| 190 | 234 | 244 | 251 | 257 | 261 | 272 | 287 |
| 191 | 235 | 246 | 253 | 259 | 262 | 274 | 288 |
| 192 | 237 | 247 | 254 | 260 | 263 | 275 | 290 |
| 193 | 238 | 248 | 255 | 262 | 265 | 276 | 291 |
| 194 | 239 | 250 | 257 | 263 | 266 | 278 | 293 |
| 195 | 240 | 251 | 258 | 264 | 268 | 279 | 294 |
| 196 | 242 | 252 | 260 | 266 | 269 | 281 | 296 |
| 197 | 243 | 254 | 261 | 267 | 271 | 282 | 297 |
| 198 | 244 | 255 | 262 | 269 | 272 | 284 | 298 |
| 199 | 245 | 256 | 264 | 270 | 273 | 285 | 300 |
| 200 | 247 | 258 | 265 | 271 | 275 | 286 | 301 |

**Table 4.** Comparison of $V$ and $C$ Regions of Fab' New for Sequence Homology by the Alignment Statistic

| Sequence Pairs[a] | Number of Compared Residue Positions | MMD Values | Significance Level | |
|---|---|---|---|---|
| $V_L - V_H$ | 102 | 98 | $<$ 0.01 % | |
| Misaligned $V_L - V_H$ | 99 | 139 | $>$ 10 % | $<$ 50 % |
| $V_L - C_L$ | 84 | 97 | $>$ 0.1 % | $<$ 0.1 % |
| Misaligned $V_L - C_L$ | 82 | 116 | $>$ 10 % | $<$ 50 % |
| $V_L - C_H1$ | 82 | 100 | $>$ 0.1 % | $<$ 1 % |
| Misaligned $V_L - C_H1$ | 83 | 121 | $>$ 50 % | $<$ 95 % |
| $V_L$-Myoglobin | 103 | 149 | $>$ 50 % | $<$ 95 % |
| $V_H - C_L$ | 91 | 111 | $>$ 0.1 % | $<$ 1 % |
| $V_H - C_H1$ | 89 | 114 | $>$ 1 % | $<$ 5 % |
| Misaligned $C_H1 - V_H$ | 89 | 123 | $>$ 10 % | $<$ 50 % |
| $V_H$-Myoglobin | 117 | 174 | $>$ 50 % | $<$ 95 % |
| $C_L - C_H1$ | 101 | 92 | $<$ 0.01 % | |
| Misaligned $C_H1 - C_L$ | 92 | 134 | $>$ 50 % | $<$ 95 % |
| $C_L$-Myoglobin | 105 | 167 | $>$ 95 % | |
| $C_H1$-Myoglobin | 103 | 155 | $>$ 50 % | $<$ 95 % |

[a] When $V_L$ was used as a misaligned sequence in the comparisons its alignment was placed out of register by shifting each of its residues one position over to the right. When $C_H1$ was used as a misaligned sequence, its alignment was placed out of register by shifting each of its residues two positions over to the right. Sequences were also compared to an unrelated protein, dolphin myoglobin.

acids long. The $V$ region of each light chain ($V_L$) consist of the N-terminal half of the chain and the $C$ region ($C_L$) consists of the C-terminal half. The $V$ region of each IgG heavy chain ($V_H$) consists of the N-terminal quarter of the chain and is thus about the same lengths as $V_L$. Moreover the remaining 3 quarters of the chain can be divided into three homology domains $C_H1$, $C_H2$, and $C_H3$, each showing clear genetic relationship to $C_L$ and each at about the same length as a $V$ region. A fragment of IgG, called Fab' because it is the antigen binding fragment, consists of the two complete light chains and the N-terminal half ($V_H + C_H1$) of the heavy chains. The amino-acid sequences aligned by Poljak et al. (1974) were from the Fab' fragment from the IgG myeloma immunoglobulin NEW. Thus the homologies tested in Table 4 are for $V_L$, $V_H$, $C_L$, and $C_H1$ regions of IgG.

The table shows that the probability that these sequences belong to a random collection is very low for the $V_L - V_H$ and $C_L - C_H1$ pairs, less than 0.01 %. While not that low for the several $V - C$ pairs it is still small, falling between 1% and 5% for $V_H - C_H1$, between 0.1 % and 1 % for $V_L - C_H1$, and also for $V_H - C_L$ and falling between 0.01 % and 0.1 % for $V_L - C_L$.

Thus if we take the usual cutoff point for rejecting a null hypothesis in statistical tests, 5 % probability, we find that in all cases significant homology exists among these $V$ and $C$ region sequences. Moreover the probabilities are all greater than 10 % and cluster about 50 %, supporting the null hypothesis, when the sequences are misaligned or when they are compared to myoglobin, an unrelated protein.

In order to align the $V$ and $C$ regions against one another on the basis of matched features of three-dimensional structure, Poljak et al. (1974) interspersed the alignments with gaps and insertions. Such a procedure reduces the number of compared residues between sequences. This actually makes it harder for these residues to pass the significance test for homology in that with a decreasing *n*-alignment the MMD value must decrease to a proportionately greater extent to support the hypothesis of common ancestry. For example, as Table 3 shows, at *n*-alignment of 100 an MMD value 1.32 times n achieves the 5 % significance level, but at an *n*-alignment of 10 the MMD value must be reduced to 1.0 times *n* to have this significance level. The structural alignment for V and C regions loses only about 15 % of its total number of residues to insertions and we see in Table 3 the *n*-alignments which remain easily pass the test for homology. We have also restricted our *n*-alignments to stretches of contiguous amino acids between gaps or insertions as the case may be, and these too pass the significance test for homology, although usually at a somewhat borderline level. For example, $V_L$ and $C_L$ were compared over three such uninterrupted stretches starting at $V_L$ residue positions 1, 15, and 18 with *n*-alignments of 14, 17 and 24 respectively. The first of these *n*-alignments has MMD of 13, which achieves the 1 % significance level, but when moved out of register, or misaligned (each residue of $V_L$ was moved one position to the left), it has an MMD value of 19, or 50 % probability of being random. The 18 *n*-alignment has an MMD of 19, between a 1 % and 5 % probability of being random, but misaligned its MMD is 28, a probability of randomness between 50 % and 95 %. The 24 *n*-alignment has an MMD of 27, a probability between 1 % and 5 %, but misaligned its MMD is 38, probability between 50 % and 95 %. We consider such results further evidence that aligned *V* and *C* region sequences in fact do show a real evolutionary relationship.

*Dehydrogenases, Flavodoxin, and Subtilisin.* Analogous work to that on immunoglobulin *V* and *C* regions has been done by Rossmann et al. (1974) on dehydrogenases, flavodoxin, and subtilisin. The dehydrogenases share a common structural domain whose function is to bind nicotinamide adenine dinucleotide (NAD). The same structure is utilized to bind flavin monoucleotide in flavodoxin, and it is also similar to an aromatic pocket of subtilisin. This permitted Rossmann et al. (1974) to recognize corresponding amino acids when sequence comparisons alone would fail. Their best alignment was found by comparing a particular stretch of 37 residues (residues 22 –58) representing the most conserved part of the adenine binding pocket in lactate dehydrogenase (LDH) to corresponding residues in the other proteins, glyceraldehyde-3-phosphate dehydrogenase (GAPDH), liver alcohol dehydrogenase (LADH), glutamate dehydrogenase (GluDH), flavodoxin, and subtilisin. As can be seen in Table 5 these comparisons pass the alignment statistic test for homology. This supports the claim of Rossmann et al. that the conserved structures in these proteins are evolutionarily related.

*Heme-Binding Pocket in Globins and Cytochrome $b_5$.* Rossman and Argos (1975) have also used their method of comparing the three-dimensional structures of folded polypeptide chains to investigate the possible evolutionary derivation of cytochrome $b_5$ and the globins from a common primordial heme-binding protein. They found that up to 52 residues of 85 three-dimensionally characterized residues in calf liver cytochrome $b_5$ are structurally and topologically equivalent to the globin fold in horse oxyhemoglobin. On determining minimum base changes (or MMD values in our terminology) for the numbers of "equivalenced" residues in their best "fits" between

**Table 5.** Comparison of the most conserved part of LDH to corresponding residues on other dehydrogenases, flavodoxin, and subtilisin.

| Sequence Pair[a] | Number of Compared Residue Positions | MMD Value | Significance Level | |
|---|---|---|---|---|
| Dogfish LDH-Yeast GAPDH | 37 | 40 | 0.1 % | |
| Misaligned Dogfish LDH-Yeast GAPDH | 36 | 50 | > 10 % | < 50 % |
| Dogfish LDH-Horse LADH | 36 | 38 | > 0.01 % | < 0.1 % |
| Misaligned Dogfish LDH-Horse LADH | 35 | 49 | > 0 % | < 50 % |
| Dogfish LDH-Bovine GluDH | 36 | 43 | > 1 % | < 5 % |
| Misaligned Dogfish LDH-Bovine GluDH | 35 | 46 | > 10 % | < 50 % |
| Dogfish LDH-*Clostridium* flavodoxin | 33 | 39 | > 1 % | < 5 % |
| Misaligned Dogfish LDH-*Clostridium* flavodoxin | 32 | 48 | > 50 % | < 95 % |
| Dogfish LDH-subtilisin | 36 | 41 | > 0.1 % | < 1 % |
| Misaligned Dogfish LDH-subtilisin | 35 | 46 | > 10 % | < 50 % |

[a] When dogfish LDH was used as a misaligned sequence in these comparisons, its alignment was placed out of register by shifting each of its residues one position over to the left

the two proteins, they felt that these mutation values were in the range expected for evolutionarily related proteins, but still not low enough to prove it.

In the six best structural fits, recorded in Table 1 of Rossman and Argos (1975), there were 29, 52, 51, 40, 46 and 48 equivalenced residues yielding MMD values of 36, 67, 70, 50, 60, and 62 respectively. According to our alignment statistic these MMD values are indicative of the match between evolutionarily unrelated or random polypeptide chains. Only the N-alignment of 40 equivalenced residues with an MMD value of 50 achieves the 5 % significance level, i.e. borderline evidence for significant amino acid homology. Otherwise the N-alignments show probabilities of randomness of 10 % or greater. Thus a better case might be made that the similarity the three-dimensional structure between the heme binding pocket in globins and cytochrome $b_5$ was produced by evolutionary convergence rather than derived by common inheritance from some primordial heme binding protein. It is worth emphasizing that our alignment statistic is apparently quite discriminating in providing evidence for evolutionary homology when applied to proposed structural alignments.

## 4. Further Empirical Support for the Statistic

It is unlikely that all conceivable $n$-alignments are equally probable, because, regardless of ancestry, certain conceivable but structurally ill-conditioned polypeptide chains could never be expressed in a living organism. Thus the derivation of our alignment statistic is not strictly valid. On the other hand, it appears that the *proportions* of 0-differences, 1-differences, 2-differences, and 3-differences (6 %, 46 %, 46 %, and 2 %, respectively) match the proportions seen in observed, non-aligned sequences. For example, when $V_L$ and $C_L$ chains are one position out of alignment, we obtain difference proportions of 9 %, 45 %, 42 % and 5 % for the resulting 82 amino acid pairs. As another example, when dogfish LDH and yeast GAPDH, at the most conserved part of the adenine binding pocket, are one position out of alignment we obtain difference proportions of 8 %, 44 %, and 3 % for the resulting 36 pairs. If we combine the data on these two examples, the difference proportions are 8 %, 45 %, 42 %, and 4 % for the resulting 118 amino acid pairs.

The assumption of *independence* of each alignment position to its neighbor is also unlikely on the basis of structural considerations, but again the predicted and actual *proportions* may not be very dissimilar. If we consider sequential *pairs* of residues in the previous examples, we obtain 0-difference, 1-difference, . . ., and 6-difference values of 4 %, 7 %, 27 %, 38 %, 20 %, 4 %, and 0 % respectively for the resulting 55 amino acid pairs. The predicted values (Table 2) are 0 %, 6 %, 27 %, 42 %, 23 %, 2 %, and 0 %, respectively.

Another criticism of the derivation of our alignment statistic is its assumption that no systematic bias enters in the selection of an alignment. In many insertion/deletion problems, the investigator may slide the amino acid residue gap in an effort to minimize the observed MMD value. Obviously, the introduction of an arbitrary number of arbitrary-sized deletions and insertions can be used to lower the observed MMD to any desired value, thus destroying the sense of the test. A simulated distribution of MMD values was generated for every $n$-alignment with $n \leq 12$ and every gap-length $n - 1$, where a *single gap* was permitted to slide along the alignment and settle in position with lowest MMD value. In 5000 Monto Carlo trials apiece, the simulated significance table was never more than one mutation less than predicted in Table 3. Simulation experiments are open to the criticism that they are not exhaustive trials, but our experience suggests that manipulation of insertions and deletions in moderation may not substantially bias the evaluation of significance levels.

*Acknowledgments.* We thank Drs. David Hewett-Emmett and Richard Holmquist for stimulating discussion and for critically reviewing earlier drafts of this manuscript.
This work was supported by grant GB-36157 from the National Science Foundation.

## References

Barker, W.C., Dayhoff, M.O. (1972). Atlas of Protein Sequence and Structure 5, 89
Fitch, W.M. (1975). J. Mol. Biol., 16, 9
Fitch, W.M. (1970). J. Mol. Biol. 49, 1
Fitch, W.M., Margoliash, E. (1967). Science 155, 279
Haber, J.E., Koshland, D.E., Jr. (1970). J. Mol. Biol. 50, 617
Jukes, T.H., Cantor, C.R. (1969). In: Mammalian protein metabolism, H.M. Munro, Ed. New York: Academic Press
McLachlan, A.D. (1971). J. Mol. Biol. 61, 409
McLachlan, A.D. (1972). J. Mol. Biol. 64, 417
Needleman, S.B., Wunsch, C.D. (1970). J. Mol. Biol., 48, 443
Poljak, R.J., Amzel, L.M., Chen, B.L., Phizackerley, R.P., Saul, R. (1974). Proc. Natl. Acad. Sci. (U.S.A.) 71, 3440
Rossman, M.G., Argos, P. (1975). J. Biol. Chem. 250, 7525
Rossman, M.G., Moras, D., Olsen, K.W. (1974). Nature 250, 194
Sankoff, D. (1972). Proc. Natl. Acad. Sci. (U.S.A.) 69, 4

*Received June 25, 1976*