References and Notes

1. V. N. Zharkov and V. P. Trubitsyn, *Physics of Planetary Interiors* (Pachart, Tucson, Ariz., 1978), p. 308.
2. W. B. Hubbard, J. J. MacFarlane, J. D. Anderson, G. W. Null, E. D. Biller, *J. Geophys. Res.* **85**, 5909 (1980); W. B. Hubbard and J. J. MacFarlane, *ibid.*, p. 225.
3. W. B. Hubbard, *Rev. Geophys. Space Phys.* **18**, 1 (1980).
4. ———, W. L. Slattery, C. L. DeVito, *Astrophys. J.* **199**, 504 (1975).
5. W. B. Hubbard, *Astron. Zh.* **51**, 1052 (1974).
6. H. Mizuno, *Prog. Theor. Phys.* **64**, 544 (1980).
7. F. Perri and A. G. W. Cameron, *Icarus* **22**, 416 (1974).
8. S. K. Sharma, H. K. Mao, P. M. Bell, *Phys. Rev. Lett.* **44**, 886 (1980); *Carnegie Inst. Washington Yearb.* **79**, 358 (1980).
9. D. J. Stevenson and E. E. Salpeter, in *Jupiter*, T. Gehrels, Ed. (Univ. of Arizona Press, Tucson, 1976), pp. 85–112.
10. D. J. Stevenson, *Phys. Rev. B* **12**, 3999 (1975).
11. G. S. Orton and A. P. Ingersoll, *J. Geophys. Res.* **85**, 5871 (1980).
12. R. Hanel *et al.*, *Science* **212**, 192 (1981).
13. G. W. Null, E. L. Lau, E. D. Biller, J. D. Anderson, *Astron. J.* **86**, 456 (1981).
14. J. L. Elliot, R. G. French, J. A. Frogel, J. H. Elias, D. J. Mink, W. Liller, *ibid.*, p. 444.
15. A. C. Mitchell and W. J. Nellis, *High Pressure Sci. Technol.* **1**, 428 (1979).
16. L. Wallace, *Icarus* **43**, 231 (1980).
17. M. Ross and F. Ree, *J. Chem. Phys.* **73**, 6146 (1980).
18. W. B. Hubbard and J. J. MacFarlane, *Icarus* **44**, 676 (1980).

19. Supported by NASA grant NSG-7045.

Similar Amino Acid Sequences: Chance or Common Ancestry?

Russell F. Doolittle

Summary. The systematic comparison of every newly determined amino acid sequence with all other known sequences may allow a complete reconstruction of the evolutionary events leading to contemporary proteins. But sometimes the surviving similarities are so vague that even computer-based sequence comparison procedures are unable to validate relationships. In other cases similar sequences may appear in totally alien proteins as a result of mere chance or, occasionally, by the convergent evolution of sequences with special properties.

The ultimate goal in the study of protein evolution is the reconstruction of past events that have given rise to the vast inventory of proteins in existence today. It is altogether likely that the overwhelming majority of extant proteins—and certainly most enzymes—have evolved from a very small number of archetypal proteins. The premise is based on the notion that it is simpler to duplicate and modify proteins genetically than it is to assemble appropriate amino acid combinations de novo from random beginnings. In present-day living systems the invention of new proteins depends on gene duplications that lead to specific amino acid sequences being coded for by more than one segment of DNA (or RNA) in a given genome. The duplications are the results of various breakage and reunion events that occur more or less randomly in the genetic material (*1*).

Amino acid sequence studies have revealed that gene duplications occur in all kinds of organisms, prokaryotic and eukaryotic alike. The duplicated segment of nucleic acid may be short, corresponding to a few amino acids, or extensive enough that microscopically visible pieces of chromosome are involved. Depending on whether or not the duplicated portions of the base sequence coexist within the boundaries set by the start and stop signals for protein synthesis, the duplication may lead either to (i) an elongated polypeptide chain or (ii) two separate copies of the protein. The distinction between the two kinds of duplication—contiguous and discrete—is an important one. In the first case, the result is a larger protein fashioned at the expense of the preexisting gene product. Many examples of this phenomenon are recognizable in existing protein sequences (*2*), and there is little doubt that this process has been the major route to larger proteins. In the second case, two independent gene products result, for one of which there ought to be a relaxation of the evolutionary restraints imposed by natural selection. As such, it is free to mutate, most often to random oblivion, but occasionally to a form adapted to some new role. The mutations that lead to divergence are mostly single base substitutions that engender individual amino acid replacements, although other events leading to deletions or insertions also occur.

Examples of creating a new protein with a new function by this route suggest that the new protein usually retains many of its preexisting features, the structural adaptations for new roles often being quite subtle, such as the sundry polypeptide chains that constitute the vertebrate hemoglobin system (*3*). The general shapes and folding patterns of these proteins are similar, and they all bind heme in essentially the same way. But small differences in their structures affect their interactions and their oxygen-binding properties. Similarly, examination of the deep-rooted phylogenetic tree of serine proteases reveals that the fundamental catalytic machinery is virtually identical for all these enzymes, but differences in the substrate binding region allow for an elegant selectivity of action for the diverse gene products that have descended from a host of duplications in the past (*4*). Even when the function of the "new" protein changes radically, as in the case of haptoglobin, a vertebrate transport protein clearly descended from serine protease stock (*5*), key structural features are retained. In this case the protein, whose present role is the salvaging of spent globins, has sharpened its ability to bind specific polypeptides but has lost its capability for hydrolyzing them (*6*, *7*). Similarly, α-lactalbumin, a cofactor in the lactose synthetase system that has evolved from the polysaccharide-splitting enzyme lysozyme, has retained its ability to bind a saccharide component but has lost its hydrolytic capability (*8*). By comparison, it ought to be much more difficult to fashion a new protein with a specific function de novo. In the case of enzymes, the likelihood of assembling a stable constellation of amino acids that

The author is a professor in the Department of Chemistry, University of California, San Diego, La Jolla 92093.

0036-8075/81/1009-0149$01.00/0

can catalyze a given reaction at a rate comparable to those produced by other modern enzymes must be very small (9). Thus, gene duplications leading to amino acid sequence redundancies have been a major driving force in both the elongation of small primitive polypeptides and in expanding the available repertoire of gene products.

Many investigators have considered the possibility of tracing protein evolution back to its primeval roots by comparing present-day sequences. By focusing on pairs of amino acid sequences at various divergence stages subsequent to a duplication and noting the degree of surviving correspondence, it should be possible to reconstruct a protein genealogy in the same way that sequences of the "same" protein from different species have been utilized to give a phylogeny of organisms. However, there are a number of complications and technical problems that have hindered efforts in this regard, not the least of which is the very large number of proteins, only a relatively few of which have been sequenced. In this article, I consider some of the approaches to the problem and some of the factors that have most impeded progress so far—The article is especially concerned with those distant relationships in which the sequences being compared fall between 15 and 25 percent identity where it is difficult to decide between chance similarity and genuine common ancestry.

The Current Data Base

It can be estimated that there ought to be at least 10^6 "unique" proteins in existence today, give or take an order of magnitude (10). For how many of these do we have sequences at this point? For a start, the currently available *Atlas of Protein Sequence and Structure* (11) lists 1081 entries. Of these, however, many are redundant, including the special case of the immunoglobulins (124 sequences listed). Entries for the "same" protein from different species (a vital part of the *Atlas*, it must be added) further lower the number of genuinely different entries. For example, there are 106 cytochrome c entries, as well as 43 fibrinopeptides A and 35 fibrinopeptides B. In fact, exclusive of the immunoglobulins and species differences, the *Atlas* only contains slightly more than 300 different peptide or protein sequences (Table 1).

Another limitation to our current data base has to do with the sizes of the proteins whose sequences have been determined. In this regard, the proteins listed in the current *Atlas* are significantly smaller than the "average" protein, a natural consequence of investigators' undertaking the most manageable problems first (12, 13). Also, the organisms represented are not necessarily the most appropriate for the reconstruction of a system that can be rooted back to ancestral types. Ideally we would like to have at our disposal sequence data for various mainstream enzymes and proteins that exist in both prokaryotic and eukaryotic organisms (14).

Any attempt at proving the basic premise that all enzymes and proteins are descended from a small number of prototypes will be at risk if the data base is not representative and sufficiently large. Accordingly, I have assembled a supplementary atlas consisting of recently published sequences, called *Newat* (Table 1). Redundant structures were purposely de-emphasized in this collection, to the virtual exclusion of immunoglobulins and neurotoxins. Apart from expanding the data base significantly, the supplementary atlas also provides an independent control for other aspects of comparing protein sequences that might be biased by a particular kind of data base, including the types of organisms represented, sizes of the proteins, and the like (15).

Comparing Distantly Related Sequences: State of the Art

The comparison and matching of distantly related, or unrelated, amino acid sequences can be hazardous. Some of the obstacles include problems associated with the lengths of sequences, their amino acid compositions, the stopping and starting places, and, most frustrating of all, the occurrence of interruptions in either or both of the sequences being compared. As a result of these complications, relationships have been described that do not hold up to statistical verification. For example, there have been numerous reports, pro and con, of an alleged relation between the sequences of vertebrate lysozymes and ribonucleases (16). For the most part the alignments were achieved by overzealous "gapping." In contrast, during the last two decades, we have witnessed the discovery of a number of unexpected but significant resemblances between functionally dissimilar proteins that a priori were not suspected of having common ancestry. Some of the more interesting of these have included avian lysozymes and mammalian lactalbumins (8), plasma albumin and α-fetoprotein (17), β-thromboglobulin and platelet factor 4 (18), insulin and relaxin (19), parvalbumin and troponin C (20), plastocyanin and azurin (21), proinsulin and nerve growth factor (22), the serine proteases and haptoglobulin (5), and ovalbumin and antithrombin III (23).

Many of these relationships were chance findings and not the products of a systematic search aimed at establishing families of protein sequences. Now a determined computer-based search of all available sequences is being pressed on several fronts. There are delicate judgments to make in the search for relationships, however, and it is not merely a matter of searching the data base for one sequence that resembles another. Definitive criteria are necessary for deciding what constitutes an authentic relationship.

The Gap Problem

Consider an idealized comparison of two hypothetical sequences. The aim is to distinguish between cases of authentic relationship resulting from (i) prior gene duplications and (ii) spurious similarities resulting from chance or convergence. For the moment let us assume that the 20 amino acids occur with equal frequencies (.05 each) and that proteins come in definite sizes—that is, they start and stop at specific points—and that descendants of a given gene duplication always have exactly the same lengths. A corollary of this last stipulation is that deletions and insertions are not permitted in this hypothetical case.

Let us consider the limits expected for distinguishing genuine relationships in such a situation. Obviously, two unrelated sequences of the same lengths and overall compositions will exhibit, *on the average*, 5 percent identity. The distribution of percent identities, however, if we consider a large number of such comparisons, ought to approximate a normal curve, and the spread encompassed by two standard deviations on either side of the 5 percent mean ought to be calculable. As it happens, this spread, or dispersion, is a function of the lengths of the two sequences being compared. Thus, if the two sequences are each 50 residues long, then 95 percent of the comparisons will have identities in the 0 to 11 percent range, but if the two sequences are each 200 residues long, then the same 95 percent range will only be from 0 to 9 percent. The point is that the significance of a "percent identity" in assessing the authenticity of a relationship is very much a function of the lengths

Table 1. Breakdown of various data sources by type of protein.

Source	Data bank* (crystal structures)		*Atlas of Protein Sequence and Structure*†		*Newat*‡	
	Unique	Total§	Unique	Total§	Unique	Total§
Enzymes	33	64	62	118	52	76
Redox proteins	11	15	23	166	5	12
Toxins, inhibitors	4	4	59	143	6	6
Transport, binding (not immunoglobulins)	11	28	26	139	10	15
Immunoglobulins	1	4	—	124	1	1
Hormones	3	6	65	158	2	2
Structural proteins	1	1	15	79	9	14
Miscellaneous	1	1	60	154	39	58
Total	65	123	310‖	1081	114	184

*From (70). †From (11) as supplied on purchased magnetic tape; covers volume 5 and supplements 1 to 3. ‡Compiled from the original literature, mostly covering the period 1978–mid-1980. Does not include more than 50 ribosomal protein sequences compiled in a separate index (32). §The "total" is different from the "unique" in that it includes various forms or derivatives of some proteins, as well as the same protein from various species. ‖Not including immunoglobulins.

of the sequences being compared (24).

Proteins do not come in neatly defined packages of exactly the same lengths, and it will often happen that the descendants of a given gene duplication do not have sequences that start and stop at exactly the same points. One or the other may have lost a few residues at either end, leading to terminal overhangs. In such a case it is necessary to shift the two sequences relative to each other in order to put their common residues in register. But suppose the two sequences being compared are not the descendants of a gene duplication; what does the shifting do to our expectations for random percent identities? Naturally the number of coincidences will be higher. If the two sequences are each 100 residues long, and if it is permitted to shift either of them up to five residues in either direction, then the average percent identity expected is raised from 5 to 8 percent; and 95 percent of random comparisons will fall between 4 and 12 percent (24). Viewed the other way, one comparison in every 50 will likely exceed 12 percent identity even when only random sequences are being examined.

If we move another step closer to reality and permit the existence of internal deletions and insertions (gaps), the situation becomes more complicated. It is obvious that gaps increase the matching of unrelated sequences as well as related ones, and if unlimited gaps are allowed, two unrelated sequences that are very long can be arranged in a fashion that achieves virtual identity over their aligned portions. Clearly, some sort of penalty has to be imposed every time a gap is allowed in a sequence. But deletions and insertions *do* occur during the evolution of proteins, and gaps must be introduced in many cases if a proper alignment of sequences is to be achieved. Finding the appropriate balance point for when a gap is warranted and when it is not is a part of the present art.

Convergence and Divergence

How likely is it that resemblances in amino acid sequences are the result of evolutionary convergence, as opposed to divergence from a common ancestor, and what impact would such adaptive events have on statistical expectations for chance similarities? Convergence, as used here, implies natural selection for a set of amino acids that can provide a particular structure, as opposed to the chance sequence resemblances that can be expected in any large-scale comparison. For example, there might be particular constellations of amino acids that are repeatedly selected for use as "elbows" or turns. Alternatively, there might be some combinations of amino acids that do not occur because of structural instability or steric problems.

A further problem is that the 20 amino acids do not occur with equal frequencies. King and Jukes (25) showed that, in a set of 53 mammalian proteins, the occurrence of the amino acids was roughly proportional to the average occurrence of codons expected from the base composition of mammalian DNA. At this point all available data indicate that the distribution of amino acids is similar for all groups of organisms (Fig. 1). Thus, the three most frequently occurring amino acids—glycine, alanine, and leucine—account for a quarter of all residues and occur four times as often as the least frequent amino acids—tryptophan, histidine, cysteine, and methionine. Accordingly, a chance match of one of the last named is much less likely than a match for the first named. On the average, however, the 5 percent identity expected if all the amino acids occurred equally is only shifted upward to about 6 percent when the observed distribution is taken into account. Moreover, in comparisons of sequences more than 100 residues long, the results obtained with scoring procedures that take account of amino acid frequencies differ very little from those that do not.

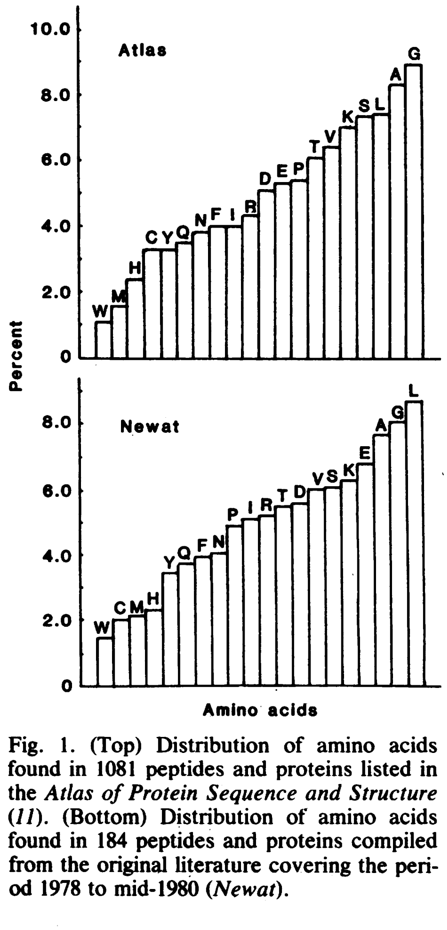


Fig. 1. (Top) Distribution of amino acids found in 1081 peptides and proteins listed in the *Atlas of Protein Sequence and Structure* (11). (Bottom) Distribution of amino acids found in 184 peptides and proteins compiled from the original literature covering the period 1978 to mid-1980 (*Newat*).

Other complications can be anticipated, however. For example, the surface covering of many globular proteins is composed of α-helical segments in which every third or fourth residue is nonpolar and is directed toward the interior of the protein, whereas most of the other side chains tend to be polar and project into the surrounding solvent. The natural rhythm of residues in these α-helices may be common enough to offset the expectations of how often the same amino acid would be expected to be matched on a strictly random basis. Similarly, peptide sequences associated with membranes or protein interiors will have disproportionate numbers of nonpolar residues and will resemble each other more than would be expected by chance alone if it were presumed that all 20 amino acids were contributing with their universal frequencies (26).

Shuffling Segments: Complications of Splicing

Many cells have the wherewithal to engage in the rearrangement of segments of the genome (*27*). The rearrangements may involve segments coding for entire polypeptide chains or, alternatively, segments that code for only a part of a polypeptide chain. Rearrangements of the latter type could conceivably result in different parts of various enzymes and other proteins becoming genetically fused. As such, various combinations of binding sites and catalytic units might be put to good advantage in the construction of new enzymes or other proteins. In other cases, the elongation of some polypeptide chains may have come about as the result of gene fusions rather than by gene duplication per se (*28*). Finally, in eukaryotes gene rearrangements and splicing appear to be intimately associated with the existence of untranslated intervening nucleotide sequences.

The notion that enzymes having the same cofactors may all be related is long-standing, and one might expect a priori that all the dehydrogenases that use pyridine nucleotide cofactors would share a common ancestry, or that all the heme binders might stem from a common structure, or, similarly, biotin binders or pyridoxal binders, and so forth. Thus, if the cofactor binding portions of a set of enzymes have similar structures or amino acid sequences, the question then arises whether the entire enzyme protein is derived from a common source in every case, or whether the various components of an enzyme were assembled by a series of gene fusions (*29*).

The possibility of shuffled segments presents a number of problems, foremost of which is the assignment of boundaries in a given sequence. Given a polypeptide chain, portions of which may have descended from two or more different ancestral proteins, where do the different sequences start and stop? Moreover, the rearranged segments may be relatively short, a factor that bears heavily on the statistics of comparisons. Thus, although it may be easier to find a match for a short sequence, it will be proportionately more difficult to prove its validity. On another count, it has been suggested that the existence of intervening sequences in eukaryotic genes may be the mechanistic basis for internal gaps in proteins, small deletions resulting from base substitutions changing the splice points (*30*). If this proves to be so, then the junction between introns and exons will be blurred and the comparisons will become more difficult. Worse, extrachromosomal elements can transpose segments between the genomes of different organisms (*31*), further confusing the situation.

Computer-Assisted Comparison: Methodology

Explorative studies for establishing sequence relationships generally assume three aspects. First, the available data base (such as the *Atlas* or *Newat*) must be searched with a regimen that is purposely set in a liberal mode so that all possibly related sequences can be identified. Different investigators use somewhat different schemes, but in our studies (*32*) we have employed a sliding segment approach similar to that introduced by Fitch (*33*). It can be set to count any number of identities, x, over the course of any segment, y, so long as $x \leq y$. It should be understood that such searches usually retrieve a number of sequences that are only marginally similar to the sequence being tested and resemble it merely by chance.

Once two sequences have been selected for rigorous comparison, the next task is to align them optimally. Several procedures have been devised for determining the optimum match of two sequences (*34–36*); most impose a "gap penalty" every time a skip is made in one or the other sequences in order to improve the degree of matching. The matching itself may be of a sort whereby only identities are scored, or it may involve a weighted scale that gives partial credit for matched amino acids that are structurally similar or that are genetically or evolutionarily favored. All these factors are taken into account—that is, matched identities or similar residues counting positively and gaps counting negatively—and an alignment score may be computed.

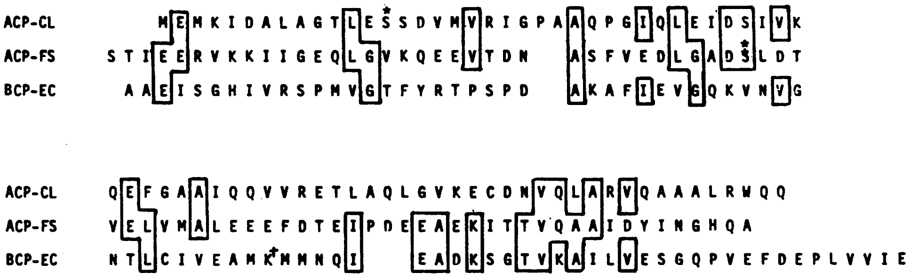
The optimum alignment of two sequences does not prove that they are evolved from a common ancestor. In order to ascertain whether the two sequences are more similar than could reasonably be expected on the basis of chance alone, some variation on the following procedure is usually undertaken. Sets of scrambled sequences whose compositions and lengths are identical to those of the two proteins under study are generated by the computer and subjected to the same alignment and scoring procedure used for the authentic alignment. The maximum alignment scores of all the scrambled comparisons are determined and averaged, and a standard deviation is computed. The alignment score obtained with the genuine sequences is then compared with the mean scrambled score, and the number of standard deviations above (or below) is noted. Scores that are 3.0 or more standard deviations above the scrambled mean scores can reasonably be expected to represent authentic relationships (*37*). In many cases the results are indecisive and fall between 0 and 3 standard deviations above the mean value of the scramble comparisons. This does not necessarily mean that the two proteins have not diverged from a common ancestral protein, but merely that any similarities between them are too weak to be statistically significant. The empirical process of computer scrambling takes into account two major variables in sequence matching—lengths of sequences and amino acid composition.

Acyl Carrier Proteins: Chance, Convergence, or Common Ancestry?

Clearly, there must be some point beyond which a resemblance for two historically related sequences cannot be verified statistically. For example, consider the amino acid sequences of two different acyl synthetase carrier proteins that operate in two different systems. One of these is the acyl carrier protein for the citrate-lyase system, a 78-residue protein that has been isolated from *Klebsiella aerogenes*. The other is the 77-residue acyl carrier protein that participates in the fatty acid synthetase system in *Escherichia coli* (Fig. 2). A number of biochemical observations suggest that the two proteins are descended from a common ancestor, including their near-identical sizes, large amount (> 50 percent) of α-helix in both proteins, and equivalent roles in their respective enzyme systems. In both cases a serine residue serves as the attachment site for a phosphoribosyl dephospho-coenzyme A prosthetic group, although the serine residues are located at different positions (*38, 39*).

Optimal alignment of the two sequences, including a single gap, results in only a 16 percent identity (Fig. 2). Worse, 10 of the 12 identities involve the six most commonly occurring residues (Fig. 1). Moreover, an empirical scrambling approach for determining statistical significance, with a simple identity scoring system, reveals that the alignment score falls 1.0 standard deviations *below* the mean of the scramble-comparisons

(Table 2). In fact, a biotin carrier protein is at least as similar to the fatty acid synthetase acyl carrier protein as is the citrate-lyase protein (Fig. 2). Has the point of proving common ancestry been passed for these sequences? Or can we, by various devices, establish common ancestry for them on the basis of amino acid sequence data alone?



ACP-CL M E M K I D A L A G T L E S S D V M V R I G P A A Q P G I Q L E I D S I V K
ACP-FS S T I E E R V K K I I G E Q L G V K Q E E V T D N A S F V E D L G A D S L D T
BCP-EC A A E I S G H I V R S P M V G T F Y R T P S P D A K A F I E V G Q K V N V G
ACP-CL Q E F G A A I Q Q V V R E T L A Q L G V K E C D N V Q L A R V Q A A A L R W Q Q
ACP-FS V E L V M A L E E E F D T E I P D E E A E K I T T V Q A A I D Y I N G H Q A
BCP-EC N T L C I V E A M K M M N Q I E A D K S G T V K A I L V E S G Q P V E F D E P L V V I E
*
†

Fig. 2. Amino acid sequences of three small "carrier proteins": *ACP-CL*, acyl carrier protein, citrate lyase system from *Klebsiella aerogenes* (*38*); *ACP-FS*, acyl carrier protein, fatty acid synthetase system from *E. coli* (*39*); *BCP-EC*, biotin carrier protein from *E. coli* (*71*). The asterisks (*) denote the active sites of the acyl carrier proteins at which the serine residues (*S*) become phosphorylated. The dagger (†) indicates the lysine residue (*K*) at which biotin becomes attached in the biotin carrier protein. None of the sequences is homologous with any of the others at a statistically significant level. The single letter code is that recommended by IUPAC (*72*).

Attempts at Image Enhancement

In theory there are several possibilities for improving the sensitivity of sequence comparison schemes so that genuine ancestral relationships can be distinguished from chance resemblances. For example, one can extend the comparison beyond the simple matching of identities, awarding positive credit for matched amino acids that are structurally similar or are known to interchange frequently during evolution (*6, 37, 40*). Such weighted scales have their drawbacks. For example, the rewarding of paired residues that are interchangeable as a result of single-base substitutions (*33*) ignores the fact that only about one-third of the 75 replacements so allowed involve structurally similar amino acids (*6*). Even with a system that weights differences on the basis of their apparent frequency among amino acid replacements observed in the "same" protein from different species (*37*), there is no assurance that any improvement will result. Thus, whereas in the case of similar sequences it is differences that are interesting, it is identities that are most significant in comparisons of very dissimilar sequences. As a result, such weighting systems may give rise to a background interference that can actually mask significant similarities. In the case of the acyl carrier proteins, use of a weighted matrix failed to establish a valid relationship.

Comparing ancestral sequences. Pauling and Zuckerkandl (*41*) long ago suggested that homologies that were not readily apparent might be revealed by comparisons of ancestral sequences. Indeed, it seems only logical that a truer gauge of relationship would be obtained if the sequences compared were those existing more closely in time to the putative duplication. By using present-day species differences, a topology can be derived that goes back to a hypothetical ancestral sequence. This approach suffers from several problems, however, and it has not yet provided the general solutions that were originally hoped for. First, data have to be available from a variety of species. Second, even then the method is exceedingly error-prone, calculated ancestral sequences always favoring the most commonly occurring amino acids (*42*).

Multiple comparison methods. Sometimes, as in the case of the acyl carrier proteins, pairwise comparisons fail to give statistically significant scores, even when other considerations strongly imply a relationship. The availability of a third distantly related sequence can sometimes be used to prove a relationship (*32, 43*). In these cases it is the coincidence of residues occurring at the same locations in all three of the sequences that can be statistically most meaningful. As it happens, the acyl carrier protein for a third enzyme system, citramalate lyase, has been isolated from *Clostridia tetanomorphum* (*44*), and it will be interesting to see whether its sequence provides a connection for establishing an overall evolutionary relationship. Acyl carrier proteins have also been isolated from plants (*45*), and it is possible that the sequences of these proteins could reveal a provable connection.

Comparing DNA sequences. It might seem that direct recourse to the DNA sequence, when it is available, might offer better data for establishing homologies. Unfortunately, there are inherent drawbacks to the comparison of DNA sequences, starting with the fact that there are only four bases. Random sequences ought to exhibit 25 percent identity on the average. But, as in the case of all sequence comparisons, there will be a distribution of resemblances. Add to this the well-known degeneracy of the code, whereby different triplets code for the same amino acid, and a situation develops wherein comparisons of the order of lengths that are of most interest (roughly 100 to 1000 nucleotides), the dispersion is usually sufficient to mask the subtle relationships we are seeking (*46*). Finally, the gap problem is at least as much of a stumbling block in DNA comparisons as it is in amino acid sequence studies.

The three-dimensional predictive schemes. Resemblances between and among proteins are often (but not necessarily always) more definitively perceived on the basis of their crystal structures than by comparisons of their amino acid sequences (*47*). Unfortunately, the number of crystal structures available is small (Table 1). Indeed, the DNA sequence explosion that is now just beginning is making available sequences of proteins that have never been isolated, let alone crystallized. If we were able to predict accurately the three-dimensional structure of a protein from its amino acid sequence, however, we might have a more reliable guide to ancestral relationships. At this point, however, current predictive schemes (*48*) are not nearly accurate or precise enough to be of any usefulness.

Survey of Some Evolutionarily Related Proteins

Even though ancillary biochemical evidence may strongly support common ancestry, two sequences may have diverged so much that current comparison procedures cannot demonstrate a statistically certifiable relationship. In such a case, the possibility of independent (functionally convergent) evolution cannot be ruled out, at least not until additional sequence data, either from other potentially related proteins or the same proteins from other species, provide bridges that are statistically valid. But many protein relationships do bear up to scrutiny, and these can be tallied with an eye to eventual ancestral reconstruction.

Table 2. A range of protein pairs thought to be related by common ancestry. This table contains 55 comparisons of 80 different protein or peptide sequences. Of these, 53 sequences were taken from the *Atlas of Protein Sequence and Structure* (*11*), and 27 were obtained from original literature sources appearing in the period 1978 to 1980 (*50–52*). In a few cases [trypsin(ogen), (pro)phospholipase, cytochrome b_5, t-antigen and viral coat proteins], a given protein is represented more than once (that is, from different species). Also, eight serine proteases are compared in ten different situations in order to show a wide range of change within a single group. Similarly, eight globin sequences are used in seven different comparisons, since these are landmarks familiar to many readers. Beyond that, the intent is to emphasize the degree of change that occurs in related proteins as they evolve different functions. The percent identity was calculated from aligned regions only; unmatched segments were not included. Similarly, in the case of gaps, "terminal gaps" (elongations or shortenings at either the amino or carboxyl termini) were not included. The gaps per 100 values were calculated with the average number of residues in the two proteins as the denominator. Alignment scores were determined by a computer program that optimally aligns two sequences by an algorithm similar to that described by Needleman and Wunsch (*34*) and imposes a penalty for every gap. The gap penalty was equal to 2.5 times the benefit gained for a match of two identical residues. Matched cysteines were given double weight in the scoring system. The statistical significance of the similarity of the two sequences was achieved by comparing the alignment scores with the mean of 36 comparisons of jumbled sequences of the same lengths and compositions as the two proteins being compared, the difference between them being expressed in standard deviations (S.D.). Usually a score of 3.0 S.D. or more is taken as demonstrating an authentic relationship (*37*). Normalized alignment scores (NAS) were calculated by dividing alignment scores by the number of aligned residues in a given comparison and multiplying by 100. In a few cases the "same" comparisons appear more than once, except with sequences from different species, in order to illustrate how the choice of species can influence the degree of similarity. Numbers in parentheses indicate sequence length.

Protein I	Protein II	Percent identity	Gaps	Gaps/ 100	NAS	*n* (S.D.)
Hemoglobin β, human (146)	Hemoglobin δ, human (146)	93.2	0	0	945	54.0
Chymotrypsinogen A, bovine (245)	Chymotrypsinogen B, bovine (245)	79.2	0	0	873	65.8
Lactate dehydrogenase M, pig (333)	Lactate dehydrogenase H, pig (331)	75.2	1	0.3	760	75.0
Hemoglobin β, human (146)	Hemoglobin γ, human (146)	73.3	0	0	740	40.7
SG-protease A, *S. griseus* (182)	SG-protease B, *S. griseus* (185)	61.8	2	1.1	612	40.9
Carbonic anhydrase B, human (260)	Carbonic anhydrase C, human (259)	60.6	1	0.4	579	56.8
β-Thromboglobulin, human (81)	Platelet factor 4, human (70)	55.7	1	1.3	579	21.9
Glucagon, human (29)	Secretin, pig (27)	51.9	0	0	519	14.1
Carboxypeptidase A, bovine (307)	Carboxypeptidase B, bovine (306)	47.7	1	0.3	475	37.5
Chymotrypsinogen A, bovine (245)	Trypsinogen, bovine (229)	46.1	6	2.6	425	22.6
Macromycin, *S. macromycetius* (45)*	Neocarzinostatin, *S. carzinostaticus* (109)	46.7	1	2.2	411	6.1
Hemoglobin β, human (146)	Hemoglobin α, human (141)	43.6	2	1.4	400	17.0
λ-Constant, immuno-, human (102)	κ-Constant, immuno-, human (104)	42.1	3	3.0	399	13.1
Lysozyme (egg white), chicken (129)	Lactalbumin (milk), human (123)	38.2	3	2.4	378	10.7
Chymotrypsinogen B, bovine (245)	Trypsinogen, bovine (229)	38.8	5	2.1	363	23.5
Viral coat protein, PF1 (46)	Viral coat protein, Xf (44)	40.5	1	2.3	345	5.0
Flagellin switcher, *Salmonella* (190)	Transposase, TN3 plasmid (185)	35.5	2	1.1	328	15.9
Insulin, elephant (51)	Relaxin, pig (48)	24.4	1	2.0	322	4.1
Bacterial trypsin, *S. griseus* (221)	Vertebrate trypsin, dogfish (222)	38.3	8	3.6	316	16.9
SG-protease B, *S. griseus* (185)	α-Lytic protease, *Myxobacter* (198)	38.9	7	3.7	314	15.9
SG-protease A, *S. griseus* (182)	α-Lytic protease, *Myxobacter* (198)	38.4	7	3.7	305	10.0
Fibrinogen β, human (461)	Fibrinogen γ, human (411)	32.6	5	1.2	304	31.2
Cardiotoxin, *Bungarus fasciatus* (118)	Prophospholipase, pig (131)	31.5	5	4.0	282	5.0
t Antigen, SV40 (174)	t Antigen, polyoma (211)	27.6	3	1.6	279	10.9
Hemoglobin α, human (141)	Myoglobin, human (153)	27.0	1	0.7	259	9.3
Neocarzinostatin, *S. carzinostaticus* (109)	Antibacterial sub. A, *S. carzinostaticus* (87)	34.7	3	3.5	257	2.1
Phycocyanin A, cyanobacter (162)	Phycocyanin B, cyanobacter (172)	28.1	2	1.2	256	7.2
Epidermal growth factor, human (53)	Pancreatic trypsin inhibitor, human (56)	24.4	1	1.9	256	3.2
Protocatchuate dioxygenase A, *Pseudomonas* (200)	Protocatchuate dioxygenase B, *Pseudomonas* (237)	32.1	7	3.2	249	11.5
Follitropin α, human (92)	Thyrotropin β, human (112)	22.0	2	2.4	244	2.9
Cardiotoxin, *Bungarus fasciatus* (118)	Phospholipase, Gaboon adder (118)	24.8	3	2.5	243	4.2
Sulfite oxidase, chicken mitochondria (97)*	Cytochrome b_5, chicken (84)	31.3	2	2.3	241	6.2
Ovalbumin, chicken (386)	Antithrombin III, human (423)	28.1	6	1.6	241	14.3
Hemoglobin β, human (146)	Myoglobin, human (153)	25.5	1	0.7	238	8.8
Parvalbumin, carp (108)	Troponin C, bovine (161)	27.0	2	2.0	230	6.1
Cytochrome c, pig (104)	Cytochrome f, *Spirulina maxima* (89)	27.0	3	2.9	213	7.2
Sulfite oxidase, chicken mitochondria (97)*	Cytochrome b_5, rabbit (116)	26.6	2	2.1	213	5.8
Viral coat protein, FD (49)	Viral coat protein, Xf (44)	25.6	1	2.3	198	1.1
β2-Microglobulin, human (100)	κ-Constant, immuno-, human (104)	23.7	2	2.0	196	4.4
Invertebrate hemoglobin, midge (152)	Vertebrate hemoglobin, hagfish (148)	26.4	3	2.0	187	3.7

Ribosomal protein L32, *E. coli* (56)	Ribosomal protein L33, *E. coli* (54)	18.5	0	0	185	2.3
α-Lytic protease, myxobacter (198)	Vertebrate trypsin, dogfish (222)	26.5	8	3.8	183	2.7
β2-Microglobulin, human (100)	κ-Constant, immuno-, human (102)	18.8	1	1.0	182	3.5
Plastocyanin, spinach (99)	Azurin, *Pseudomonas* (128)	27.3	4	4.0	182	2.9
Cytochrome c, human (104)	Cytochrome f, *Euglena gracilis* (87)	25.0	3	3.0	181	2.9
Ribosomal protein, L30, *E. coli* (58)	Ribosomal protein, L32, *E. coli* (56)	17.9	0	0	179	2.1
Viral coat protein, PF1 (46)	Viral coat protein, FD (49)	23.3	1	2.3	174	−0.2
Histocompatibility antigen, mouse (173)	λ chain, immunoglobulin SH (183)*	16.7	3	1.7	170	4.1
Leghemoglobin, yellow lupin (153)	Invertebrate hemoglobin, midge (151)	22.0	3	2.0	167	2.6
Nerve growth factor, mouse (118)	Proinsulin, human (89)	16.0	1	1.2	167	1.6
Chymotrypsinogen B, bovine (245)	Haptoglobin β, human (245)	19.4	4	1.6	165	5.4
Fibrinogen α, human (239)†	Fibrinogen β, human (239)†	18.5	5	2.1	153	5.7
Fibrinogen α, human (239)†	Fibrinogen γ, human (239)†	15.5	2	0.8	151	5.4
Chymotrypsinogen A, bovine (245)	Haptoglobin β, human (245)	18.9	5	2.0	149	3.2
ACP, citrate lyase, *Klebsiella* (78)	ACP, fatty acid synthetase, *E. coli* (77)	16.0	1	1.3	127	−1.0

*Only represents portion of entire sequence. †Only the NH_2-terminal regions (amounting to one-third to one-half) of the fibrinogen chains were used in these two comparisons. The complete α-chain contains 610 residues (73), the COOH-terminal two-thirds of which is radically different from the β-chain (461 residues) and γ-chain (411 residues).

For example, Table 2 shows 55 different pairs of proteins whose sequences indicate common ancestry.

The comparisons drew heavily from the approximately 200 sequences assembled from the original literature appearing in the period 1978 to 1980 (*Newat*), although a number of previously compared and familiar pairs are included that are meant to serve as benchmarks for the reader, including eight different and wide-ranging hemoglobin comparisons and ten different serine protease-type comparisons. All comparisons were made with a simple computer-comparison scheme that emphasizes identities (*49*). The computer generates an optimal alignment, a normalized alignment score (NAS), and a measure of the statistical significance of the match.

All the protein pairs listed in Table 2, in my view, are genuinely related and have evolved from common ancestors, although in a few cases, such as the acyl carrier proteins and certain of the viral coat proteins, the evidence is biological or biochemical and is not based on the sequence comparisons alone. Of the more than two dozen new comparisons, most of the resemblances were noted by the investigators themselves (*50*). In some cases, however, they pointed out some possible similarities of sequence but were hesitant about the significance of the match (*51*), and in still others, the resemblances to previously reported sequences were apparently overlooked (*52*).

The data in Table 2 suggest that an NAS greater than 200, which is equivalent to an adjusted 20 percent identity, almost always indicates a genuine relationship. The significance of scores between 140 and 200 depends on the lengths of the two sequences being compared; if they are longer than 200 residues, the evidence for common ancestry is very strong (Table 2). When the NAS drops below 140, however, a relationship can only be alleged if there is strong ancillary evidence.

Not unexpectedly, there is an inverse relation between the number of gaps per unit length and the degree of identity (Fig. 3). The data indicate that for all but the shortest (in residue length) of comparisons, it would be unreasonable if gaps did not occur once the identity falls below 50 percent. But is there a general rule for how many gaps can be expected? Obviously, the tolerance for deletions or insertions will be dependent on structure-function aspects of the particular proteins being compared. Of the 55 sequence pairs listed in Table 2, the highest number of gaps observed per unit length in any comparison is approximately 4 out of 100 residues. Any comparison involving that number of gaps ought to achieve a minimum of 25 percent identity in order to be statistically significant, even if it involves very long sequences, inasmuch as the introduction of every gap means that the cutoff mark for significance is shifted upward.

The listing in Table 2 also provides some insight into the size of gaps introduced in protein sequences as a function of their relatedness (*53*). A complete survey of the gaps introduced in all the comparisons revealed that small gaps were, not surprisingly, more frequent than large gaps (Fig. 4). Larger gaps were only slightly more frequent in the comparison of very distantly related sequences than in cases of moderately distant relationships. On another note, the frequency of gaps in prokaryotic sequences was not significantly different from that in eukaryotic ones, contradicting notions (*30*) that intervening sequences play a major role in the occurrence of internal deletions (*54*).

Less Likely and Unlikely Relationships

In contrast to those in Table 2, the sequence pairs listed in Table 3 constitute a group in which the relationships, if any, are considerably more subtle, although in most of the cases the two sequences compared have been reported to be "homologous." Approximately half of the comparisons gave NAS's high enough to warrant serious scrutiny. Several were quite significant when tested by the scrambled sequence approach, although some others involved short sequences for which that test is not appropriate. In some cases the sequences, taken alone, appear to meet the test of statistical significance, but the biology of the situation engenders caution if not skepticism.

At the top of the list (Table 3), with both the highest normalized score (NAS) and the greatest statistical likelihood (n = 5.2 standard deviations; see legend of Table 2), is a pair consisting of diphtheria toxin fragment B and a human plasma apolipoprotein (*55*). The resemblance is striking, indeed, there being 18 identities over a stretch of 72 aligned residues, with only a single gap of one residue having been introduced. The NAS is a commanding 215. Since the initial observation, however, the sequence of the adjacent region of the diphtheria toxin sequence (fragment A) has been determined (*56*), and in this case there is no apparent homology with

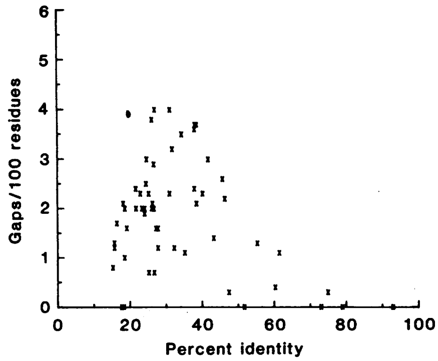


Fig. 3. Distribution of gaps per 100 residues as a function of the percent identity of compared sequences. The 100-residue unit length is based on the mean lengths of the two residues being compared. The data are taken from the 55 comparisons listed in Table 2.

the corresponding region of the apolipoprotein. When the combined fragment A and fragment B segments are aligned with the apolipoprotein, the NAS drops to an unexciting 141, and the statistical significance disappears. Is the match listed in Table 3 the result of a gene splice? At present we have no precedent for bacterial invaders being able to pirate pieces of DNA (or messenger RNA) from their eukaryotic hosts, although it is interesting that the diphtheria toxin is made on an extrachromosomal element that is thought not to be a part of the corynebacterium genome (*57*). For the time being, however, it is probably best to consider the resemblance between the mammalian lipid-binding protein and the corynebacterium-associated toxin material to be an extraordinary statistical coincidence. It would also be premature to consider this a case of convergence on the basis of common functional requirements.

In a prescient paper on the subject of the origin of life, Egami (*58*) suggested that the cytochromes and ferredoxins might have evolved from a common ancestor. In view of the premise of this article—that all extant proteins are related to a small number of root stocks—this would have been a very important observation. As it happens, the NAS's obtained with the exact examples he chose, cytochromes c_3 from *Desulfovibrio vulgaris* and ferredoxins from two species of *Clostridium*, are quite impressive (Table 3). In contrast, the statistical significance as determined by the scrambled sequence approach is nonexistent. This is not to say that the two problems have not evolved from a common ancestor: only that the extant sequences compared have not yet provided the evidence. It may very well be that a better choice of species (*59, 60*) or the use of multiple-comparison methods (*32, 43*) will provide connections between the two groups, although it should be borne in mind that the crystal structures of the two proteins exhibit no obvious resemblances (*61*).

The β subunit of *Vibrio cholera* toxin has been reported to be homologous with various mammalian glycoprotein hormones (*62*). The low significance of the best possible matches of these proteins makes the purported relationship highly questionable, however (Table 3). In another case, it has been reported that κ casein, from either sheep or cattle, is homologous to the γ chain of human fibrinogen (*54*). In neither case does the alignment hold up to statistical evaluation (*63*) (Table 3). Ironically, however, when this problem was being studied, the fibrinogen γ-chain sequence was sub-

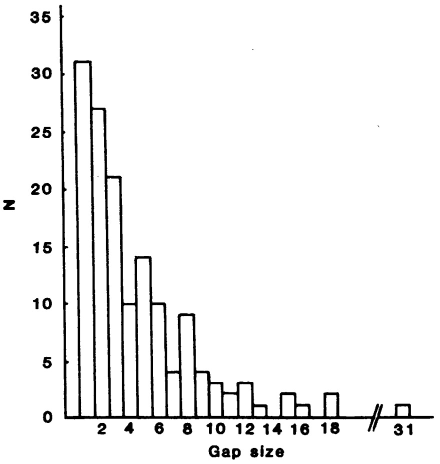


Fig. 4. Distribution of gap sizes observed in 55 sequence comparisons listed in Table 2. Although there is some redundancy in the data, inasmuch as some sequences were involved in more than one comparison, a more restricted tally of gaps in comparisons in which no sequence was used more than once had a similar distribution.

Table 3. Some protein pairs for which there are reports suggesting common ancestry. The comparison procedure is described in legend to Table 2.

Protein I	. Protein II	Percent identity	Gaps	Gaps/100	NAS	*n* (S.D.)
Lipid binding protein, AI, human (125/245)*	Diphtheria toxin fragment, *Corynebacter* (125)	25.0	1	0.8	215	5.2
β-Thromboglobulin, human (81)	Fibrinogen γ-chain, human (88/411)*	28.0	3	3.7	193	2.0
Cytochrome c_3, *Desulfovibrio vulgaris* (107)	Ferredoxin, *Clostridium pasteurianum* (55)	20.0	1	1.8	191	−0.6
Cytochrome c_3, *D. vulgaris* (107)	Ferredoxin, *Clostridium butyricum* (55)	27.3	3	3.7	191	−0.3
Brain-specific protein, bovine (91)	Troponin C, bovine (161)	24.7	2	2.4	185	1.1
Phosphoglyceromutase, yeast (151/241)*	Myoglobin, sheep (153)	20.7	3	2.0	155	2.4
Cholera toxin β, *Vibrio cholera* (103)	Thyrotropin β, bovine (113)	17.7	2	1.9	146	1.2
Kappa casein, sheep (171)	Fibrinogen γ-chain, human (179/411)*	20.9	5	2.9	142	1.4
Lysozyme, chicken (129)	Ribonuclease, pig (124)	14.3	2	1.6	134	0.4
ATP-Ribosyl transferase, *Salmonella typh.* (299)	Lac repressor, *E. coli* (347)	18.4	7	2.3	127	0.8
Cytochrome oxidase II, bovine mitochondria (137/227)*	Plastocyanin, spinach (99)	19.2	3	2.0	126	1.5
ATP-Ribosyl transferase, *Salmonella typh.* (299)	Phosphoglyceraldehyde dehydrogenase, pig (331)	15.0	4	1.3	119	1.6
Colicin E3, *E. coli* (97)	Ribonuclease U2, *Ustilago sphaerogena* (113)	14.4	1	1.0	119	−0.1
Lipotropin, human (91)	Follitropin α, human (92)	14.6	1	1.1	116	0.0
Aldolase A, rabbit (360)	Lactate dehydrogenase, dogfish (329)	14.7	4	1.2	115	1.0
Kappa casein, bovine (169)	Fibrinogen γ, human (179/411)*	15.0	3	1.8	111	−0.8
Cytochrome oxidase V, bovine mitochondria (109)	Cytochrome c_3, *D. vulgaris* (107)	12.0	1	1.0	105	
Cytochrome oxidase II, bovine mitochondria (137/227)*	Azurin, *Pseudomonas fluorescens* (128)	11.3	0	0.0	93	0.1

*Only the fraction of the total sequence shown was used in the comparison.

jected to the search routine and another candidate, β-thromboglobulin, a protein found in blood platelets, turned up. The alignment procedure produced a rather high NAS and a not trivial statistical significance (n = 2.0 S.D.). Given the functional interdependence of proteins involved in mammalian blood coagulation, this may indeed be a genuine relationship. In any event, the alignment is considerably more striking than when the same γ-chain sequence is compared with κ caseins.

Table 3 also contains a number of other comparisons where the original investigators thought they perceived homology or common ancestry, including the case of ATP (adenosine triphosphate) ribosyltransferase from *Salmonella* that has been reported to resemble both the lac repressor and phosphoglyceraldehyde dehydrogenase (*64*). Either of these would have been of considerable interest in our search for root stocks, but both fall below the level of likelihood. Similarly, colicin E3 from *Escherichia coli* has been reported to be a potential ribonuclease, and comparisons of its sequence with those of other ribonucleases hinted that it might resemble the ribonuclease U2 from the fungus *Ustilago sphaerogena* (*65*). The resemblance is below the level of verification, however. Although cytochrome oxidase, subunit II, has been reported (*66*) to be homologous to plastocyanin and azurin, this result does bear up to scrutiny, even in a tailor-made comparison where a large unmatchable segment of cytochrome oxidase was omitted, and may be the result of overgapping. A possible relationship of aldolase A to lactate dehydrogenase (*67*) does not endure either. Finally, the perennial comparison of lysozyme and ribonuclease was performed, and it also fails to achieve significance.

Again, while it is possible to express the likelihood of two proteins being genuinely related on the basis of their amino acid sequences, it is not possible to prove that two sequences are not derived from a common ancestor. The diverging sequences may simply have changed so much that the correlation cannot be proved statistically.

Protein Families and Superfamilies

Several years ago Zuckerkandl proposed that the number of "classes of proteins" found in nature is probably less than a thousand (*68*). He arrived at this conclusion more or less intuitively on the basis of the rate at which "different" proteins were being found to be homologous. Dayhoff *et al.* (*69*), using the terms "families" and "superfamilies," arrived at a similar conclusion regarding superfamilies. The arithmetic leading to their conclusion was not altogether clear, however, and their assignment of protein sequences to families and superfamilies was arbitrary, indeed. Thus, by the definition of Dayhoff *et al.*, proteins belong to the same family if they differ at fewer than half of their amino acid positions, whereas superfamilies comprise those sequences that are demonstrably homologous but differ at more than half of their positions.

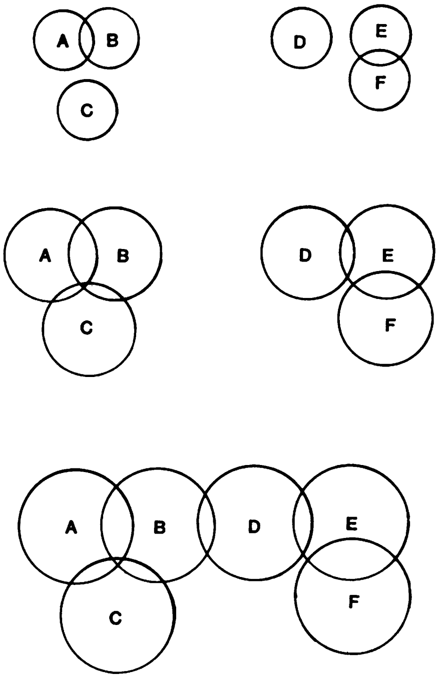


Fig. 5. Symbolic representation of sequence families and superfamilies. A family is defined as any set of sequences that are demonstrably homologous, regardless of the percent identity; they are denoted by the sets A, B, C, D, E and F. Superfamilies are defined as comprising two or more families, not all of whose members are demonstrably homologous with all the other members of each family. The top section of the figure includes the superfamilies A-B and E-F. The middle section, which depicts the situation after more sequences have been added to the data base, includes the superfamilies A-B-C and D-E-F. Finally, the bottom section shows the situation when a very large number of sequences have been determined and a number of intersections found for the six families, a single superfamily now being evident.

It is questionable whether sharp percent identity cutoffs in this context provide any useful information, and I would like to propose a completely different usage for the terms "protein family" and "protein superfamily." In this new system, let any sequences that are demonstrably related to each other belong to the same family, regardless of the percent identity that they exhibit. Thus, haptoglobin would belong to the same family as the serine proteases (Table 2). Moreover, if the sequence of a particular protein from any species meets the statistical requirements for an authentic relationship, then let that automatically provide family membership for that protein from all species, unless other considerations render the purported relationships illogical (*59*).

Superfamilies, on the other hand, would be composed of two or more families, not all of whose different proteins are demonstrably homologous with all the members of the other families in the set:

Superfamily I = Family A ∪ Family B

Obviously, this scheme allows a given protein to belong to more than one family (Fig. 5), a not unreasonable prospect given the notion that all proteins are derived from a small number of starter types.

Do the data accumulated so far support such a notion? Of the 424 "unique" proteins listed in Table 1, upward of 15 percent have already been found to resemble "other" proteins, not including the "same" protein in other species. More than 70 of these proteins appear in Table 2. Although many of these relationships were more or less anticipated on the basis of similar functions, as in the case of the serine proteases or the globins or glycopeptide hormones, others were wholly unexpected. As the data base expands we can reasonably expect more of the latter.

How many superfamilies can we expect? Certainly many will be identified as more sequences are accumulated and various families found to have common members. But eventually the number of observed superfamilies ought to reach a maximum and then decline as the combined sets themselves begin to intersect (Fig. 5). In the limit, the number of superfamilies as manifested by this scheme should be the same as the number of archetypal proteins that served as starter types.

References and Notes

1. Actually, breaks may occur randomly, but misalignment most frequently involves base-pairing of similar, as opposed to random, base sequences. As a result, gene duplication often begets more gene duplication.
2. Among proteins that have been lengthened by contiguous gene duplication are the bacterial and plant ferredoxins, immunoglobulins, calmodulin, plasma albumin, plasminogen, vertebrate fibrinogen chains, transferrin, various ribosomal proteins, and β-galactosidase.
3. H. A. Itano, *Adv. Protein Chem.* **12**, 215 (1957); V. M. Ingram, *The Hemoglobins in Genetics and Evolution* (Columbia Univ. Press, New York, 1963); M. F. Perutz, *Proc. R. Soc. London Ser. B* **173**, 113 (1969).
4. K. A. Walsh and H. Neurath, *Proc. Natl. Acad. Sci. U.S.A.* **52**, 884 (1964); D. M. Shotton and B. S. Hartley, *Nature (London)* **225**, 802 (1970); A.

D. McLachlan and D. M. Shotton, *Nature (London) New Biol.* **229**, 202 (1971); L. T. J. Delbaere *et al.*, *Nature (London)* **257**, 758 (1975).

5. D. R. Barnett, T.-H. Lee, B. H. Bowman, *Biochemistry* **11**, 1189 (1972).
6. R. F. Doolittle, in *The Proteins*, H. Neurath and R. L. Hill, Eds. (Academic Press, New York, ed. 2, 1979), vol. 4, pp. 1–118.
7. A. Kurosky *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **77**, 3388 (1980).
8. K. Brew, *Essay Biochem.* **6**, 93 (1970).
9. This is not to say that similarly acting enzymes have not evolved more than once. In fact, the independent invention of lysozymes on more than one occasion seems quite certain [R. J. Simpson, G. S. Begg, D. S. Dorow, F. J. Morgan, *Biochemistry* **19**, 1814 (1980)], and serine proteases have evidently arisen from two different stocks, as evidenced by the subtilisin group, and the chymotrypsinogen type [J. Kraut *et al.*, *Cold Spring Harbor Symp. Quant. Biol.* **36**, 117 (1971)].
10. In this context we are not interested in species variations of the "same" protein. Cytochrome c, for example, would only be counted once, even though there are thousands of versions expressed among extant organisms. The estimate of 10^6 is a compromise between two extremes: first, the maximum number of individual and different proteins that a prokaryotic organism can make, given the size of its genome, is about 10^4 [J. D. Watson, *The Molecular Biology of the Gene* (Benjamin, New York, ed. 3, 1976), p. 219]. In a world in which all organisms made only variations of the same proteins this would be the number of unique proteins. Eukaryotic genomes are much larger; and, although much of their DNA is not translated into protein, the number of proteins made by a given eukaryotic organism might be as high as 10^5. If we include a few orders of magnitude for the multiplicity of different proteins made by different organisms (that is, all organisms do not make the same proteins), then the number of existing protein types may be of the order of 10^7 to 10^8. A much smaller number results from an estimate of the number of different enzymes in existence based on a tabulation of their activities and characteristic properties [*Enzyme Nomenclature 1978*. Recommendations of the Nomenclature Committee of the International Union of Biochemistry on the Nomenclature and Classification of Enzymes (Academic Press, New York, 1979)]. The last official count was 2122, and a decline in the rate of new reports in recent years suggests that there may be as few as 3000 definably different enzymes. If we now estimate what fraction of known proteins are enzymes, it would seem that the number of unique proteins may be as few as 10^4 to 10^5. The compromise estimate of 10^6 is meant only to give us some idea of a count.
11. A tape containing the 1081 sequences listed in *The Atlas of Protein Sequence and Structure* [(National Biomedical Research Foundation, Washington, D.C., 1978), vol. 5 and supplements 1, 2, and 3] was purchased from the publisher.
12. The average sequence listed in the *Atlas* (*11*) is less than 100 residues in length, whereas the average protein subunit length in nature is about 350 residues long (*13*).
13. S. S. Sommer and J. E. Cohen, *J. Mol. Evol.* **15**, 37 (1980).
14. Some proteins that have been shown to be evolutionarily related in prokaryotes and eukaryotes on the basis of amino acid sequence resemblances include phosphoglyceraldehyde dehydrogenase, triosephosphate dehydrogenase, dihydrofolate reductase, phosphorylase, ferredoxin, pyruvate carboxylase, cytochromes, ribosomal proteins, and ribonuclease.
15. In addition to being longer on the average, the sequences in *Newat* include more prokaryotic and viral proteins, as well as many nonglobular (for example, membrane) proteins. Computer printouts of *Newat* may be obtained by sending a self-addressed and suitably stamped manila envelope (10 by 13 inches) to me.
16. Reports favoring a relationship include C. Manwell, *Comp. Biochem. Physiol.* **23**, 383 (1967) and D. J. Strydom, *J. Mol. Evol.* **9**, 349 (1970); reports denying a relationship include J. E. Haber and D. E. Koshland, Jr., *J. Mol. Biol.* **50**, 617 (1970) and W. M. Fitch, *Annu. Rev. Genet.* **7**, 343 (1973).
17. E. Ruosolahti and W. D. Terry, *Nature (London)* **260**, 803 (1976).
18. G. S. Begg, D. S. Pepper, C. N. Chesterman, F. J. Morgan, *Biochemistry* **17**, 1739 (1978).
19. R. James, H. Niall, S. Kwok, G. Bryant-Greenwood, *Nature (London)* **267**, 544 (1977).
20. J. H. Collins, *Biochem. Biophys. Res. Commun.* **58**, 301 (1974); A. G. Woods and A. D. McLachlan, *Nature (London)* **252**, 646 (1974).
21. L. Ryden and J. O. Lundgreen, *Nature (London)* **261**, 344 (1976).
22. W. A. Frazier, R. H. Angeletti, R. A. Bradshaw, *Science* **176**, 482 (1972).
23. L. T. Hunt and M. O. Dayhoff, *Biochem. Biophys. Res. Commun.* **95**, 864 (1980).
24. K. Shuler, G. Weiss, K. Lindenberg, unpublished work.
25. J. L. King and T. H. Jukes, *Science* **164**, 788 (1969).
26. An examination of the frequency distribution of the 400 possible dipeptidyl sequences occurring in all the proteins listed in both the *Atlas* and *Newat* revealed that the observed counts were not significantly different from what would be expected in a random situation, given the overall frequency of the individual amino acids. The only real anomaly observed was that the dipeptidyl sequence Cys-His (Cys, cysteine; His, histidine) occurs two and a half times as often in the *Atlas* as had been expected (255 instead of 99). This imbalance, which was not observed in *Newat*, was mainly attributable to the fact that Cys-His occurs in all cytochromes c, a disproportionate number (> 100) of which are listed in the *Atlas*.
27. H. Sakano, R. Maki, Y. Kurosawa, W. Roeder, S. Tonegawa, *Nature (London)* **286**, 676 (1979); G. S. Roeder, P. J. Farabaugh, D. T. Chaleff, G. R. Fink, *Science* **209**, 1375 (1980); M. Simon, J. Zieg, M. Silverman, G. Mandel, R. Doolittle, *ibid.*, p. 1370.
28. J. M. Hood, A. V. Fowler, I. Zabin, *Proc. Natl. Acad. Sci. U.S.A.* **75** 113 (1978).
29. The answer to the question has a direct bearing on how we approach the evolution of many key enzymes and deserves close scrutiny. Thus although there are well-known cases of clusters of independent genes being fused so that a continuous polypeptide chain is expressed [J. Yourna, T. Kohno, R. Roth, *Nature (London)* **228**, 820 (1970)], this is a quite different situation from where the different component pieces of a single enzyme are the result of genetic exchange. There are cases where a single cofactor appears in several independently evolved systems, including heme-binding domains that occur in conjunction with and fused to various oxidases [B. Guiard and F. Lederer, *J. Mol. Biol.* **135**, 639 (1979)], but here again each of the joined units can exist as an independent gene product, a factor that influences the chances of success during natural selection. This is also significantly different from the popular notion of splicing together, for example, a pyridine nucleotide-binding segment, a dehydrogenating engine, and a particular substrate-binding component, thereby inventing a new enzyme from parts with no apparent survival value when on their own. For a somewhat different view of the problem, see A. Lilljas and M. G. Rossmann [*Annu. Rev. Biochem.* **43**, 475 (1974)].
30. E. Smith, in *Frontiers of Bioorganic Chemistry and Molecular Biology*, S. N. Ananchenko, Ed. (Pergamon, New York, 1980), pp. 39–51.
31. Extrachromosomal elements have clearly played a role in gene transfer in prokaryotes [D. Jones and P. H. A. Sneath, *Bacteriol. Rev.* **34**, 40 (1970); D. Reanney, *ibid.* **40**, 552 (1976)].
32. R. A. Jue, N. W. Woodbury, R. F. Doolittle, *J. Mol. Evol.* **15**, 129 (1980).
33. W. M. Fitch, *J. Mol. Biol.* **16**, 9 (1966).
34. S. B. Needleman and C. D. Wunsch, *J. Mol. Evol.* **48**, 443 (1970).
35. D. Sankoff, *Proc. Natl. Acad. Sci. U.S.A.* **69**, 4 (1972); P. H. Sellers, *SIAM (Soc. Ind. Appl. Math.) J. Appl. Math.* **26**, 787 (1974).
36. T. C. Elleman, *J. Mol. Evol.* **11**, 143 (1978).
37. W. C. Barker and M. O. Dayhoff, in *Atlas of Protein Sequence and Structure* (National Biomedical Research Foundation, Washington, D.C., 1972), vol. 5, pp. 101–110.
38. K. Beyreuther, H. Bohmer, P. Dimroth, *Eur. J. Biochem.* **87**, 101 (1978).
39. T. C. Vanaman, S. J. Wakil, R. L. Hill, *J. Biol. Chem.* **243**, 6420 (1968).
40. A. D. McLachlan, *J. Mol. Evol.* **61**, 409 (1971).
41. L. Pauling and E. Zuckerkandl, *Acta Chem. Scand.* **17**, 59 (1963).
42. J. L. King, *J. Mol. Evol.* **15**, 73 (1980).
43. M. S. Waterman, T. F. Smith, W. A. Beyer, *Adv. Math.* **20**, 367 (1976).
44. P. Dimroth, W. Buckel, R. Loyal, H. Eggerer, *Eur. J. Biochem.* **80**, 469 (1977).
45. R. D. Simoni, R. S. Criddle, P. K. Stumpf, *J. Biol. Chem.* **242**, 573 (1967).
46. As a case in point, a comparison of the DNA sequences of the sarcoma oncogenes from the avian and murine sarcoma viruses led to the conclusion that there was no homology between the two [A. P. Czernilofsky, A. D. Levinson, H. E. Varmus, J. M. Bishop, *Nature (London)* **287**, 198 (1980)]. A subsequent comparison of the putative amino acid sequences derived from those same DNA sequences revealed that the two gene products are as homologous as myoglobin and the β chain of hemoglobin [C. A. Van Beveren, J. A. Galleshaw, V. Jonas, A. J. M. Berns, R. F. Doolittle, D. J. Donoghue, I. M. Verma, *Nature (London)* **289**, 58 (1981)].
47. For example, in a study of relationships between prokaryotic and eukaryotic serine proteases, comparisons of α-carbon backbone coordinates were about twice as effective as the use of sequence matching as an index to homology [M. N. G. James, L. T. J. Delbaere, G. D. Brayer, *Can. J. Biochem.* **56**, 396 (1978)].
48. For a review of predictive schemes, see P. Y. Chou and G. D. Fasman, *Annu. Rev. Biochem.* **47**, 51 (1978).
49. The effectiveness of the program was attested to by the fact that the sequences of human cytochrome c and *Euglena gracilis* cytochrome f were found to be related (Table 2), a relationship that has been used in the past as an indicator of the sensitivity of a procedure for detecting distant relationships (*36*). In a few cases comparisons were also made with a matching scale that was inversely weighted to the universal frequencies of the amino acids; in some others a matrix based on structure-function similarities was employed (*6*). In most of these cases there was no significant difference in the alignment scores achieved, and in the end we found the simple identity matching program to be as effective as any we tried.
50. The following reports noted similarities of the named sequences to previously reported sequences: sulfite-oxidase heme-binding domain, B. Guiard and F. Lederer, *Eur. J. Biochem.* **100**, 441 (1979); phycocyanin c subunits, G. Frank, W. Sidler, H. Widmer, H. Zuber, *Hoppe-Seyler's Z. Physiol. Chem.* **359**, 1491 (1978); SV40 and polyoma t antigens, T. Friedmann, R. F. Doolittle, G. Walter, *Nature (London)* **274**, 291 (1978); flagellin switcher protein, J. Zeig and M. Simon, *Proc. Natl. Acad. Sci. U.S.A.* **77**, 4196 (1980); macromycin, T. H. Sawyer, K. Gustzow, M. O. J. Olson, H. Busch, A. W. Prestayko, S. T. Crooke, *Biochem. Biophys. Res. Commun.* **86**, 1133 (1979); β-thromboglobulin (*15*); cardiotoxin, H.-S. Lu and T.-B. Lo, *Int. J. Pept. Protein Res.* **12**, 181 (1978); hagfish hemoglobin, G. Liljequist, G. Braunitzer, S. Paleus, *Hoppe-Seyler's Z. Physiol. Chem.* **360**, 125 (1979); mouse histocompatibility factor, H. Uehara, B. M. Ewenstein, J. M. Martinko, S. G. Nathenson, J. E. Coligan, T. J. Kindt, *Biochemistry* **19**, 306 (1980); fibrinogen chains, A. Henschen and F. Lottspeich, *Thromb. Res.* **11**, 869 (1977); K. W. K. Watt, T. Takagi, R. F. Doolittle, *Proc. Natl. Acad. Sci. U.S.A.* **75**, 1731 (1978); R. F. Doolittle, K. W. K. Watt, B. A. Cottrell, D. D. Strong, M. Riley, *Nature (London)* **280**, 464 (1979); nerve growth factor (*22*); ovalbumin (*23*).
51. Inexplicable uncertainty was expressed about the very significant relationship of viral coat proteins by B. Frangione, Y. Nakashima, W. Konigsberg and R. L. Wiseman [*FEBS Lett.* **96**, 381 (1978)].
52. The most astonishing case involved the α and β subunits of protocatchuate 3,4-dioxygenase [N. A. Kohlmiller and J. B. Howard, *J. Biol. Chem.* **254**, 7309 (1979); M. Iwaki, H. Kagamiyama, M. Nozake, *J. Biochem. (Tokyo)* **86**, 1159 (1979)].
53. The penalty imposed by the alignment program employed in this study did not make a distinction between gaps on the basis of size, the logic being that the loss (gain) of any multiple of three nucleotides can effect a gap of any corresponding number of amino acids in a single event.
54. It is generally conceded that intervening sequences are restricted to eukaryotic genomes. See, for example, W. Gilbert, in *Eukaryotic Gene Regulation*, R. Axel, T. Maniatis, C. F. Fox, Eds. (Academic Press, New York, 1979).
55. P. Lambotte, C. Capiau, J. M. Ruysschaert, P. Falmagne, J. Kirkx, *Arch. Int. Physiol. Biochim.* **87**, 819 (1979).
56. R. J. DeLange, L. C. Williams, R. E. Drazin, R. J. Collier, *J. Biol. Chem.* **254**, 5838 (1979).
57. R. J. Collier, *Bacteriol. Rev.* **39**, 54 (1975).
58. F. Egami, *Origins Life* **5**, 405 (1974).
59. The use of the most appropriate species is not without hazard, and sequence-comparers must not abuse this method just as they must guard against the promiscuous use of gaps. For example, during the course of a search for sequences that resemble phosphoglyceromutase (*60*), a

possible match appeared linking that enzyme to sheep myoglobin. There were 23 identities among the 127 aligned residues, and the NAS score positioned it relatively high on the listings (Table 3). The statistical significance was uncertain ($n = 2.4$ S.D.). What puts the lie to common ancestry, quite apart from the distant relationship of yeast and sheep, is that none of the other myoglobin sequences, of which there were more than two dozen explored, had a resemblance anywhere near approaching the sheep comparison. A statistical anomaly occurred in which there were several identities in one small segment, and that had carried the bulk of the "homology."

60. L. A. Fothergill and R. N. Harkins, *Eur. J. Biochem.*, in press. I thank L. A. Fothergill for sending me this sequence in advance of its publication.
61. E. T. Adman, L. C. Sieker, L. H. Jensen, *J. Biol. Chem.* **248**, 3987 (1973); F. R. Salemme, *Annu. Rev. Biochem.* **46**, 299 (1977).
62. A. Kurosky, D. E. Markel, J. W. Peterson, W. M. Fitch, *Science* **195**, 299 (1977); A. Kurosky, D. E. Markel, J. W. Peterson, *J. Biol. Chem.* **252**, 7257 (1977).
63. J. Jolles, A.-M. Fiat, F. Schoetgen, C. Alais, P. Jolles, *Biochim. Biophys. Acta* **365**, 335 (1974); P. Jolles, M.-H. Loucheux-Lefebvre, A. Henschen, *J. Mol. Evol.* **11**, 271 (1978).
64. D. Piszkiewicz, B. E. Tilley, T. Rand-Meir, S. M. Parsons, *Proc. Natl. Acad. Sci. U.S.A.* **76**, 1589 (1979).
65. K. Suzuki and K. Imahori, *J. Biochem. (Tokyo)* **84**, 1031 (1978).
66. G. J. Steffens and G. Buse, *Hoppe-Seyler's Z. Physiol. Chem.* **360**, 613 (1979).
67. E. Stellwagen, *J. Mol. Biol.* **106**, 903 (1976).
68. E. Zuckerkandl, in *Ecole de Roscuff-1974* (Editions du Centre National de la Recherche Scientifique, Paris, 1974); *J. Mol. Evol.* 7, 1 (1975).
69. M. O. Dayhoff, W. C. Barker, L. T. Hunt, R. M. Schwartz in *Atlas of Protein Sequence and Structure, Suppl. 3* (National Biomedical Research Foundation, Washington, 1978).
70. The *Protein Data Bank Newsletter,* No. 13, July 1980 (Chemistry Department, Brookhaven National Laboratory, Upton, N.Y. 11973).
71. M. R. Sutton, R. R. Fall, A. M. Nervi, A. W. Alberts, R. Vagelos, R. A. Bradshaw, *J. Biol. Chem.* **252**, 3934 (1977).
72. The single letter code is A, alanine; C, cysteine; D, aspartic acid; E, glutamic acid; F, phenylalanine; G, glycine; H, histidine; I, isoleucine; K, lysine; L, leucine; M, methionine; N, asparagine; P, proline; Q, glutamine; R, arginine; S, serine; T, threonine; V, valine; W, tryptophan; Y, tyrosine [IUPAC-IUB Commission on Biochemical Nomenclature, *Biochem. J.* **113**, 1 (1969)].
73. K. W. K. Watt, B. A. Cottrell, D. D. Strong, R. F. Doolittle, *Biochemistry* **24**, 5410 (1979).
74. I thank N. W. Woodbury and L. R. Doolittle for writing the computer programs used in this study. I also thank K. Schiar, G. Weiss, and K. Lindenberg for giving me access to their unpublished work, K. Anderson for assistance in compiling original articles for tabulation in *Newat*, W. E. Doolittle for typing the sequences into the computer, and P. E. Geiduschek, J. Kyte, and C. Wills for reading the manuscript and offering critical comments on it.

AAAS–Newcomb Cleveland Prize
To Be Awarded for an Article or a Report Published in *Science*

The AAAS–Newcomb Cleveland Prize is awarded annually to the author of an outstanding paper published in *Science* from August through July. This competition year starts with the 7 August 1981 issue of *Science* and ends with that of 30 July 1982. The value of the prize is $5000; the winner also receives a bronze medal.

Reports and Articles that include original research data, theories, or syntheses and are fundamental contributions to basic knowledge or technical achievements of far-reaching consequence are eligible for consideration for the prize. The paper must be a first-time publication of the author's own work. Reference to pertinent earlier work by the author may be included to give perspective.

Throughout the year, readers are invited to nominate papers appearing in the Reports or Articles sections. Nominations must be typed, and the following information provided: the title of the paper, issue in which it was published, author's name, and a brief statement of justification for nomination. Nominations should be submitted to AAAS–Newcomb Cleveland Prize, AAAS, 1515 Massachusetts Avenue, NW, Washington, D.C. 20005. Final selection will rest with a panel of distinguished scientists appointed by the Board of Directors.

The award will be presented at a session of the annual meeting. In cases of multiple authorship, the prize will be divided equally between or among the authors.

Deadline for nominations: postmarked 16 August 1982

Nomination Form
AAAS–Newcomb Cleveland Prize

AUTHOR: ____

TITLE: ____

____

DATE PUBLISHED: ____

JUSTIFICATION: ____

____

____

____

____

____

____

____

____