review

On computer-assisted analysis of biological sequences: proline punctuation, consensus sequences, and apolipoprotein repeats

Mark S. Boguski, Mark Freeman, Nabil A. Elshourbagy,* John M. Taylor,* and Jeffrey I. Gordon[1]

Departments of Biological Chemistry and Medicine, Washington University School of Medicine, St. Louis, MO 63110, and Gladstone Foundation Laboratories, Cardiovascular Research Institute, Department of Physiology, University of California, San Francisco, CA 94149*

Abstract During the past several years, the use of computer programs in the analysis of protein and DNA sequences has become commonplace. In all but the simplest procedures, the ability to critically review the results obtained with computer methods requires *i)* a basic knowledge of the algorithms employed (and the assumptions upon which they are based), *ii)* an awareness of the capabilities and limitations of the particular program that implements an algorithm, and *iii)* some familiarity with probability and statistics. We describe a number of computer methods that have been applied to the analysis of apolipoprotein sequences. We discuss the suitability of these methods for particular problems, how the choice of initial "parameters" can affect the results, and what the results can tell us about protein or gene sequences. We also identify some outstanding problems of apolipoprotein sequence analysis where further work is needed. —**Boguski, M. S., M. Freeman, N. A. Elshourbagy, J. M. Taylor, and J. I. Gordon.** On computer-assisted analysis of biological sequences: proline punctuation, consensus sequences, and apolipoprotein repeats. *J. Lipid Res.* 1986. **27:** 1011–1034.

Supplementary key words apolipoproteins • database searching • optimized alignments • comparison matrices • correlation analysis • secondary structure prediction • peptide engineering • molecular evolution and sequence phylogenies

CONTENTS

INTRODUCTION

Computers are essential tools for modern research on the sequences of biological macromolecules. For the apolipoproteins in particular, application of various computer methods has led to important insights about the structure, function, and evolution of this protein family. Historically, Fitch (1), McLachlan (2), Barker and Dayhoff (3), and Segrest and Feldmann (4) were first to apply computer methods to the analysis of apolipoprotein sequences. However, such techniques did not enjoy widespread use until smaller, more powerful and less expensive computer hardware became available, until improved and more user-friendly software was developed, and until the explosive increase in molecular sequence information made computers necessary for efficient data management. Recently, our group (5–9) and others (see below) have employed a variety of computer programs in the analysis of apolipoprotein DNA and amino acid sequences. At their best, these programs can detect subtle (often unsuspected) relationships among sequences and produce experimentally testable hypotheses about protein structure and function. The purpose of the present work is to provide a detailed overview of the computer methods that have proved most useful in studying the apolipoproteins and to critically review some of the results derived by

Abbreviations: LCAT, lecithin:cholesterol acyltransferase; CD, circular dichroism; HDL, high density lipoproteins; LDL, low density lipoproteins.

[1]To whom reprint requests should be addressed.

these methods. The definition and usefulness of "consensus" sequences are described. Finally, we show that the precise conceptualization of "repeated sequences" depends upon the context in which they are considered: the definition of a repeating unit may vary relative to its evolutionary, structural or functional significance.

PHILOSOPHY OF COMPUTATIONAL ANALYSIS

Computers and their programs may be best thought of as simply additional laboratory tools and experimental methods that have their own unique capabilities and limitations. As with any experimental method, reproducibility is of crucial importance and depends on the precise conditions under which the experiment is conducted. It is not widely appreciated that most computer programs operate under a variable set of initial "parameters" (or conditions) that can greatly influence the results obtained. These initial parameters are entirely separate from the input data itself and largely reflect certain assumptions that the user or programmer has made. Specific types of program parameters (e.g., span or window, gap penalty, k-tuple) will be described below along with the programs that employ them. For now, we would just like to make the point that, *for purposes of reproducibility, supplying the parameters under which an analysis has been done is arguably as important as, for example, stating the pH and substrate concentration for an enzyme reaction.* It is surprising that such essential information is frequently absent from published accounts of computational analyses.

Many computer methods for sequence analysis employ probability and statistical techniques that become especially important in assessing the significance of distant or subtle sequence relationships. Statistics just indicate the relative likelihood that a hypothesis is true. Often additional criteria are necessary for evaluating the results of computational analysis (10). For example, when a database search reveals an unlikely relationship between two sequences, is there really any biological reason to believe that the relationship is of functional or evolutionary significance? One can also consider whether alternative computer methods produce mutually consistent results (7). Computer-derived results are best interpreted in the context of direct experimental evidence. For example, are Chou-Fasman structure predictions consistent with information derived from circular dichroism studies? In summary, there are five criteria to be considered when interpreting computer-derived results: *i)* accurate and appropriate data and parameters; *ii)* statistical significance; *iii)* biological context; *iv)* supporting (experimental) evidence; and *v)* mutual consistency. Of course, not all of these criteria will be satisfied in every situation. But they are extremely helpful in the critical appraisal of computational analyses.

METHODS AND APPLICATIONS

Database searching

An increasingly common dilemma for molecular biologists is that knowing the sequence of a macromolecule often precedes any knowledge of its precise function. Indeed, the first or only indication of the biological function of a newly discovered sequence may be its similarity to another sequence of known function identified by searching a database. Database searching is thus a logical and important first step in sequence identification and analysis. However, establishing a relationship between two sequences is not merely a matter of finding, in a database, one sequence that resembles another (11). Once a putative homology has been identified, rigorous quantitative comparisons must be carried out to distinguish genuine ancestral relationships from chance similarity or convergent sequence evolution.[2] This is critically important in cases where the biological context seems dubious.

The two major databases available in this country are the Bolt, Beranek and Newman, Inc. GENETIC SEQUENCE DATA BANK[3] (GenBank™) and the National Biomedical Research Foundation PROTEIN SEQUENCE DATABASE[4] (NBRF). GenBank (Release 42.0) contains over 6.7 megabases representing 9,697 sequences and the NBRF database (Release 9.0, May 1986) contains 862,289 residues representing 3,712 sequences. The information stored in these databases overlaps to an unknown extent.

One important aspect of effective database searching is an awareness of what *is* and what *is not* present in the database (especially when interpreting negative search results). Only a small fraction of all proteins and nucleic acids in nature has been sequenced to date. There are an estimated 30–40 thousand proteins specified by the mammalian genome (12), but the primate file of GenBank, for example, contains only 822 sequence entries with some of these being multiple copies of the same sequence determined in different laboratories. Phylogenetic representation of some sequences is prodigious; other sequences may be represented by only a single species. Sequence collections are

[2] Convergent evolution has been defined by Doolittle (11) as natural selection for a particular structure as opposed to divergence from a common ancestral sequence. An example of convergent evolution is the selection of a particular constellation of amino acids that produce topologic similarities in the active site catalytic groups of the zinc enzymes, thermolysin, carbonic anhydrase, and alcohol dehydrogenase (15).

[3] Bolt, Beranek and Newman Inc., 10 Moulton St., Cambridge, MA 02238.

[4] Protein Identification Resource, National Biomedical Research Foundation, Georgetown University Medical Center, Washington, DC 20007.

biased for historical reasons and because sequences that are abundant and easily purified are over-represented. For example, of the 3,447 sequences in Release 7.0 of the NBRF Database, 162 (4.7%) represent hemoglobin subunits of many different species. Immunoglobulins and cytochromes are also abundantly represented by 288 (8.4%) and 261 (7.6%) sequences, respectively. This phenomenon not only adversely affects the statistics of sequence comparison but also means that, for newly discovered sequences, chances are small that database searching will reveal anything of biological significance in any individual case (11).

We have previously discussed, in some detail, a number of programs that are available for searching data banks of DNA and protein sequences (7). Since that time, Lipman and Pearson (13) have devised two programs that significantly improve the speed and sensitivity of database searching. FASTP is used for protein similarity searches and its companion program, FASTN, is used for nucleic acid database comparisons. We will use these programs in conjunction with the cDNA and protein sequences of human apoA-IV to illustrate several important aspects of database searching. But first we need to address the issue of sensitivity and specificity as it relates to the type of sequence data used for analysis and the selection of search/alignment parameters.

In general, DNA comparisons are less sensitive and subject to more background "noise" than protein comparison (11, 14, 15). This is primarily due to codon degeneracy and to the fact that DNA has only a 4-character alphabet.[5] The protein alphabet, of course, has 20 characters and each of these can be assigned a large number of "character states" representing different physical, chemical, or evolutionary properties of the amino acids or even particular post-translational modifications. In rapidly evolving proteins (or very distantly related ones) there may even be no statistically significant relationships among DNA sequences that diverged from a common ancestral gene, yet the protein products may have the same three-dimensional structure, active site geometry, etc. (15). In a direct test of the efficacy of DNA versus amino acid sequences for analyzing gene duplication in the myosin rod, McLachlan (14) showed that DNA comparisons, even at the codon level, were less informative than protein comparisons and did not yield any significant new observations. Still, it is important to analyze both protein and DNA sequences when they are available because certain evolutionary issues can only be addressed at the nucleotide level, particularly in the analysis of non-coding sequences.

Prior to conducting a database search, one must specify one or a number of comparison parameters that will influence the speed, sensitivity, and specificity of the search. For example, programs FASTP and FASTN (13) require a user-selected *k-tuple* defined as a group of k consecutive sequence elements (amino acid residues or nucleotides) where k is a small, positive integer. In practical terms, the smaller the k-tuple, the more sensitive a search will be (although possibly at the expense of specificity). Larger k-tuple values can significantly accelerate a search because fewer comparisons are required. Larger k-tuples also increase specificity due to the fact that matches between *groups* of similar elements are less likely to occur than single-element matches in unrelated sequences.

Some type of scoring system must be used for quantitating sequence matches. The simplest (and most specific) system is to score for sequence *identities* alone where identical residues or nucleotides receive a score of 1 and non-identical elements are assigned a score of 0. Alternate scoring systems, based on the genetic code or a scale of "conservative" substitutions are much more sensitive at detecting distant relationships. FASTP employs the *mutation data* (PAM250) *matrix* originally devised by Dayhoff and co-workers (16). This system is based on observed frequencies of amino acid substitutions in protein families. For example, tryptophan residues tend to be highly conserved in evolution and this fact is reflected by a high score (+17) in the PAM250 matrix. Substitution of cysteine for tryptophan, on the other hand, is an unlikely event and this is reflected in a replacement score of −8.

Gap penalty is another type of comparison parameter often used in database searches and sequence alignments (see below). Often, similarity between sequences can be maximized when hypothetical insertion or deletion mutations (i.e., gaps) are permitted in the alignment. However, nucleotide and amino substitutions occur far more frequently in evolution than insertions or deletions (16) and thus a gap penalty (that decrements the matching score by some user-specified value) is imposed to control for this phenomenon. Gap penalties are not used in FASTP and FASTN because these programs were designed to rapidly detect relatively short, localized similarities in which the occurrence of an insertion or deletion would be unlikely.

Fig. 1 displays the results of searching the NBRF database with the sequence of human apoA-IV (9, 17) using FASTP. This program computes similarity scores for all sequences in the database and then calculates the mean and standard deviation for this score distribution (Fig. 1A). The user can then inspect this distribution for highly significant scores and select a number of sequences for display, ranked according to their initial scores. (Optimized scores are subsequently calculated, allowing for gaps.) The top 20 scores for the human apoA-IV database search are listed in Fig. 1B. It can be seen that the highest scores are achieved for other apolipoproteins, specifically apoA-I and apoE. Apolipoproteins A-II, C-I, C-II and C-III did *not* appear in the top 80 scores despite their presence in the database (data not shown). A curious similarity was detected between apoA-IV and nematode

[5] On average, random DNA sequences of similar base composition should have 25% identity (11).

A

```
<  2     0 :
   4     0 :
   6     1 :=
   8     1 :=
  10    16 :========
  12    28 :==============
  14    40 :====================
  16   112 :=========================================================
  18   115 :=========================================================
  20   227 :=========================================================
  22   306 :=========================================================
  24   338 :=========================================================
  26   346 :=========================================================
  28   366 :=========================================================
  30   335 :=========================================================
  32   285 :=========================================================
  34   233 :=========================================================
  36   209 :=========================================================
  38   131 :=========================================================
  40   103 :=========================================================
  42    70 :===================================
  44    57 :=============================
  46    35 :==================
  48    25 :=============
  50    19 :==========
  52    14 :=======
  54    12 :======
  56     7 :====
  58     6 :===
  60     2 :=
  62     0 :
  64     1 :=
  66     1 :=
  68     1 :=
  70     0 :
  72     0 :
  74     0 :
  76     0 :
  78     0 :
  80     0 :
> 80     5 :===
 778594 residues in  3447 sequences, mean score:  27.7 (7.33)
  836 scores better than 33 saved, ktup: 1, fact: 4  scan time:  0:10:45.75
```

B

```
The best scores are:                                          init,  opt
>P1;LPRTA4   :  Apolipoprotein A-IV precursor - Rat          1193, 1193
>P1;LPRTE    :  Apolipoprotein E precursor - Rat              275,  325
>P1;LPHUA1   :  Proapolipoprotein A-I precursor - Human       269,  305
>P1;LPHUE    :  Apolipoprotein E precursor - Human            263,  314
>P1;LPDGA1   :  Apolipoprotein A-I - Dog                      214,  277
>P1;MWKW     :  Myosin heavy chain - Nematode                  67,   93
>P1;TDRT     :  Thy-1 membrane glycoprotein precursor - Rat    65,   66
>F1;VEPG     :  Vimentin - Pig (fragments)                     64,  100
>P1;FRRTL    :  Ferritin light chain - Rat                     59,  101
>P1;FRHOL    :  Ferritin light chain - Horse                   59,   93
>P1;WJFFEN   :  Specific body pattern development protein - F  58,   67
>P1;ITHUC    :  Alpha-1-antichymotrypsin precursor - Human     58,   82
>P1;FRHUL    :  Ferritin light chain - Human                   57,   87
>P1;STHO     :  Somatotropin - Horse                           57,   69
>P1;MZXR4S   :  Nonstructural protein NCVP4 - Simian 11 rotav  57,   88
>P1;VCVWM2   :  env polyprotein - Friend mink cell focus-indu  57,   66
>P1;VCVWEM   :  env polyprotein Pr80 - Moloney murine leukemi  56,   67
>P1;P2IV68   :  P2 protein - Influenza A virus (strain A/NT/6  56,   72
>P1;KIBYG    :  Phosphoglycerate kinase (EC 2.7.2.3) - Baker'  56,   59
>F1;DMCH     :  Desmin - Chicken (fragments)                   56,   84
```

C

```
The best scores are:                                          init,  opt
>F1;FPRT     :  Alpha-fetoprotein - Rat (fragment)             57,   62
>P1;HARB     :  Hemoglobin alpha chain - Rabbit                55,   55
>F1;TMCHA    :  Tropomyosin alpha chain, skeletal muscle - Ch  55,   55
>P1;GNMV1M   :  pol polyprotein - Moloney murine leukemia vir  54,   56
>P1;RNECS    :  DNA-directed RNA polymerase (EC 2.7.7.6) sigm  54,   58
>P1;SUBSN    :  Subtilisin (EC 3.4.21.14) precursor - Bacillu  54,   54
>P1;CBUTB    :  Cytochrome b - Trypanosoma brucei mitochondri  53,   58
>P1;EQXL     :  Proenkephalin precursor - South African toad   53,   53
>P1;TMHOBP   :  Tropomyosin beta chain, platelet - Horse       52,   85
>P1;VHIHMJ   :  Nucleocapsid protein - Murine coronavirus MHV  52,   57
>P1;CSBO     :  Catalase (EC 1.11.1.6) - Bovine                52,   52
>P1;KIHUG    :  Phosphoglycerate kinase (EC 2.7.2.3) - Human   52,   56
>P1;GNMVGV   :  pol polyprotein - AKV murine leukemia virus    51,   53
>P1;QXBY31   :  Hypothetical oxi 3 intron 1 protein - Baker's  51,   57
>P1;ATFF7    :  Actin 7 - Fruit fly                            50,   59
>P1;GNLJGB   :  pol polyprotein - Bovine leukemia virus        50,   50
>P1;TLBPHL   :  Minor tail protein precursor H - Bacteriophag  50,   69
>P1;ATFY     :  Actin - Physarum polycephalum, Dictyostelium   49,   55
>P1;RLCSA    :  Ricin D, A chain - Castor bean                 49,   55
>P1;RZCS     :  Seed storage protein, small and large chains   49,   49
```

Fig. 1. Comparison of human plasma apoA-IV with the entries in the NBRF protein sequence database. The FASTP program, developed by Lipman and Pearson (13), was used to compare human plasma apoA-IV with each of the 3712 sequences present in release 7.0 of the National Biomedical Research Foundation (NBRF) Protein Sequence Database. A MicroVAX II (Digital Equipment Corp.) running on version 4.2 of the VMS operating system was used for this and all other computations described in the present work. Panel A: Histogram displaying the distribution of scores for all sequences in the database when plasma apoA-IV was used as the query sequence. Examination of the distribution shows that there are five sequences in the database which yielded initial scores >80. Panel B: A list of the twenty sequences present in the NBRF database that most resembled human plasma apoA-IV when the search was conducted using a *k-tuple* of 1. Individual database entries are identified by their retrieval key codes. Panel C: The results of searching the NBRF database using a randomized apoA-IV sequence. Randomization was achieved using the shuffling algorithm of Knuth (18).

myosin, although even the optimized score, 93, is much less significant than that for the lowest scoring apolipoprotein (dog apoA-I, score = 277).

What is the real significance of the non-apolipoprotein matches identified in this database search? Database search programs are designed always to generate "answers" and it is entirely up to the user to assess the results for evolutionary, structural, functional, or statistical significance. It is instructive to repeat our previous database search using a *shuffled* version of apoA-IV, i.e., a sequence of the same length and amino acid composition as apoA-IV but with a *random* sequence of residues. The results of this search (top 20 scores) are shown in Fig. 1C. The mean and standard deviation for this score distribution are very similar to those obtained with the real apoA-IV sequence (Fig. 1B). However, the initial and optimized score values for matches obtained with the *random* sequences are in the same range as those of the non-apolipoprotein sequence matches for the *real* sequence comparisons. In effect, this indicates the level of background "noise" for a search under these conditions.

The use of randomized sequences to test hypotheses about sequence relationships deserves further comment. To assess the statistical significance of a computed similarity score for two sequences (e.g., sequence A and sequence B), the standard approach is to employ Monte Carlo methods to generate a large sample of reference sequences that are randomly permuted variations of one of the two real sequences under study (19, 20). Sequence A is then compared with all of the random variations of sequence B (or vice versa), thus creating a distribution of similarity scores against which the similarity score for the real sequences can be compared. (For example, see the use of the RDF program in reference 8.) However, the statistical patterns inherent in biological sequences make the definition of randomness problematic (20) and this may account for cases of statistically significant similarities that are biologically meaningless. For example, note the results for the Protein Sequence Database search with apoA-IV summarized in Fig. 1B. The mean matching score was 27.7 with a standard deviation of 7.33. The matching score for apoA-IV with human ferritin light chain was 57, nearly 4 standard deviations above the mean score. Indeed, all of the non-apolipoprotein matches were significant at approximately this level. Does this mean that all of these varied sequences are in some way related to apoA-IV? Clearly they are not. The fundamental issue is this: just what does one mean by a random sequence? Or, restated in statistical parlance: just what is the null hypothesis? Does one simply shuffle the sequence maintaining the amino acid (or base) composition? Or should one also preserve nearest neighbor frequencies, codon usage frequencies in nucleotide sequences, etc.? Depending upon how one defines the null hypothesis, many different statistical conclusions are possible.

Another way to evaluate the significance of putative relationships is to determine whether they persist under different circumstances. For instance, when the search summarized in Fig. 1B was repeated using a k-tuple of 2 instead of 1, the pattern of non-apolipoprotein sequence matches is almost completely different (data not shown). Additional searches, using nucleotide sequence data, can also be helpful in distinguishing authentic relationships from spurious ones. **Fig. 2** shows the results of searching the primate and rodent libraries of GenBank using the human apoA-IV cDNA sequence (9). As in the protein database searches, related apolipoprotein sequences occupy the highest scoring positions. However, except for another species of myosin, the non-apolipoprotein matches are all different (although this may be in part due to

```
The best scores are:                                          init,  opt

>RATAPOAIV   RAT APOLIPOPROTEIN A-IV MRNA, COMPLETE CDS.     2676, 3020
>HUMAPOA1    HUMAN LIPOPROTEIN APOAI GENE, COMPLETE CODING SE 420,  570
>HUMAPOAIA   HUMAN LIVER APOLIPOPROTEIN A-I (APOA-I) MRNA.    420,  572
>HUMAPOAI1   HUMAN APOLIPOPROTEIN A-I GENE, COMPLETE CODING S 420,  708
>HUMAPOE     HUMAN APOLIPOPROTEIN E MRNA.                     288,  752
>HUMAPOEF    HUMAN APOLIPOPROTEIN E (APOE) MRNA FRAGMENT.     288,  488
>RATAPOAI    RAT APOLIPOPROTEIN A-I (APOA-I) MRNA, COMPLETE C 232,  526
>RATAPOE     RAT APOLIPOPROTEIN E MRNA.                       208,  450
>RATMYHCA    RAT CARDIAC MYOSIN HEAVY CHAIN INSERT 21/26, MRN 194,  426
>RATMYHCB    RAT CARDIAC MYOSIN HEAVY CHAIN INSERT 5, MRNA.   188,  424
>HUMEPKER    HUMAN EPIDERMAL KERATIN MRNA.                    180,  536
>HUMKER56K   HUMAN 56K CYTOSKELETAL TYPE II KERATIN MRNA.     160,  516
>MUSIL2T     MOUSE INTERLEUKIN-2 MRNA, COMPLETE CDS.          152,  206
>MUSIL21     MOUSE INTERLEUKIN-2 GENE: EXON 1 AND EXON 2.     152,  228
>HUMIFNAH2   HUMAN LEUKOCYTE INTERFERON (IFN-ALPHA) ALPHA-H2  150,  234
>HUMIFNAH    HUMAN LEUKOCYTE INTERFERON (IFN-ALPHA) ALPHA-H M 150,  222
>HUMIGCB7    HUMAN IG GERMLINE H-CHAIN J-MU-DELTA REGION: MU  136,  204
>HUMIFNA1    HUMAN LEUKOCYTE INTERFERON ALPHA N GENE, PARTIAL 136,  158
>MUSIFNG     MOUSE IMMUNE INTERFERON (IFN-GAMMA) CDNA TO MRNA 126,  206
>MUSTRB3     MOUSE T(15:12) TRANSLOCATION REGION: C-MYC EXON  126,  286
```

Fig. 2. Comparison of the human apoA-IV cDNA sequence with the 1711 entries present in the Primate plus Rodent files of the Genetic Sequence (DNA) Data Bank (Release 35.0). FASTN was used to conduct this search with a *k-tuple* of 3. The top 20 matches are displayed.

differences in sequence content between the protein and DNA databases). It is interesting to note that the optimized aligment score for keratin and apoA-IV actually exceeds that for human apoA-IV and rat apoE. Again, this may be an example of a spurious "homology" due to similarities in base composition. Alternatively, sequences may be similar by *analogy* rather than *homology.* The former term implies convergent evolution based on structural constraints. It would be erroneous to conclude that keratin is a member of the apolipoprotein gene family based on such a result for a database search. Serendipitous discoveries are, however, not uncommon and such a result would merit further investigation.

Returning now to the issue of false negative results, in neither the protein nor DNA database searches were all members of the apolipoprotein gene family identified. This may reflect the simple fact that statistically significant similarities may not be demonstrable among all members of a sequence family (11). This phenomenon is especially relevant for the apolipoproteins due to the nature of the amphipathic helix as a structural and functional unit. The properties of the amphipathic helices (e.g., lipid binding, LCAT activation) do not appear to depend upon a unique sequence of amino acids but rather a particular spatial distribution of interchangeable residues whose side chains have similar chemical properties (see section on protein engineering below). This may permit a rate of sequence divergence considerably more rapid than in other protein families. Indeed, the apoA-IV gene appears to have diverged considerably even between rats and mice based on hybridization analysis (8). These two rodent species shared a common ancestor only 8–14 million years ago (21).

Optimized sequence alignments

Once two similar sequences have been identified, the next step is to generate an optimized alignment between them (although such an alignment does not prove that they are ancestrally related, ref. 11). An alignment is simply an arrangement of two sequences so that all or most of their matching elements correspond to one another. The problem of sequence alignment is subsumed under the more general problem of the *analysis of differences.*[6] Two sequences that have shared a common ancestor can acquire differences in four ways: *i)* substitutions, *ii)* deletions and insertions (indels), *iii)* compressions and expansions, and *iv)* transpositions. Only the first two types of differences are commonly dealt with in molecular sequence alignment techniques. This limitation is unfortunate because the latter two are probably more common than realized in genes that have evolved by unequal crossing-over (e.g., the apolipoproteins). The analysis of differences is accomplished by *sequence comparison.* Sequence comparison yields two types of information: *distances* and *optimum analyses* (which may consist of alignments, tracings, or listings). For some purposes, the emphasis is on distances (e.g., inferring molecular phylogenies) and this application is described in a subsequent section.

A number of alignment programs are currently available for proteins (ALIGN, PRTALN) and nucleic acids (ALIGN, NUCALN). While PRTALN (and NUCALN) only score identities (10), ALIGN can use either the unitary or mutation data matrices (16). As with the database search algorithms, the user must select a number of alignment parameters. Different "optimal" alignments can be generated with different sets of parameters. Thus, strictly speaking, an alignment is optimal only in the context of the defined parameters. For example, sequence alignments may differ depending upon the scoring matrix selected, whether nucleotide or amino acids are used, or whether a high or low penalty is imposed for gaps (indels).

Problems can occur when computing an alignment and when interpreting the results. For the apolipoproteins, difficulties arise when trying to align sequences that differ in length (e.g., apoA-IV and apoA-II) and that contain multiple internal duplications. A series of unequal cross-events can scramble the order of individual repeats in a tandem array of repeating sequences. Unfortunately, none of the alignment algorithms permit transposition of an element or elements in a linear sequence.

As if these difficulties weren't enough, sequence comparisons can produce topologically incorrect alignments even when the overall sequence homology is high. Structural elements in homologous sequences that are coincident in three dimensions are defined as being topologically equivalent (15). The dataset of homologous proteins whose atomic coordinates are known to high resolution is small. Nonetheless there are *many* examples where alignments of these proteins, based on their primary sequences, had to be subsequently readjusted once topologically equivalent residues had been identified (see ref. 15 for a review).

Sequence comparisons and the definition of repeating units

It is now well established that the apolipoproteins are largely composed of "repeating sequences" that are some multiple of 11 amino acids or 11 nucleotides. Barker and Dayhoff (3), Fitch (1), and McLachlan (2) were first to discover the existence of repeating peptides in apolipoprotein A-I. The results of Barker and Dayhoff (3) indicated that there were also repeating peptides in apolipoproteins A-II, C-I, and C-III, but these repeats were not so obvious as the more highly conserved structures in apoA-I. Karathanasis, Zannis, and Breslow (22) confirmed the presence of repeating units in apoA-I from an analysis of the nucleotide sequence. Subsequently, analysis of the cDNA and protein sequence of rat apoA-IV disclosed that it had a

[6] An excellent overview of sequence comparison may be found in Chapter 1 of reference 19.

large number of highly conserved repeats closely related to those in apoA-I (5, 6). The periodic structure of apoA-IV was later confirmed by analysis of the human sequence (9, 17). The repeating nature of the structure of apoE was not appreciated until sensitive analytical methods were applied (6, 8, 23, 24).

Computational techniques generally known as *comparison matrix methods* have been enormously useful in identifying and studying repeated sequences in the apolipoproteins (2, 6–8, 24). Because these techniques have been so important, we will describe in detail how comparison matrices are computed and how one interprets the results.

In its most simple form, a comparison matrix represents a sliding comparison of two sequences.[7] At increments of one residue (base), one sequence is compared with another (which need not be of the same length) and all matching sequence elements are noted. This process is continued until the beginning of the first sequence reaches the end of the second sequence. To provide a graphic representation of all matching sequence elements, a two-dimensional array is constructed with one sequence represented on the ordinate and the second sequence on the abscissa. Each element of this array may be thought of as containing either a value of 1 or 0 (null), the former indicating a sequence identity and the latter the absence of a match. When represented as a graph, sequence identities, then, appear as points plotted in this coordinate space. Long stretches of sequence identity appear as diagonal lines extending from upper left (amino terminus or 5' end) to lower right (carboxy terminus or 3' end). A slightly more sophisticated form of this method uses non-overlapping groups of sequence elements (e.g., codons) for scoring or allows matches other than identities (e.g., conservative amino acid substitutions).

The simple comparison matrix method described above provides no information on the statistical significance of the observed matches. To aid in the interpretation of sequence similarities (especially in cases of weak or biologically unlikely matches), knowledge of the probability that a match or string of matches could have occurred by chance is required. Both McLachlan (2, 14, 25) and Dayhoff's group (16, 26) have provided solutions to this problem and their methods are implemented in the programs CMPSEQ84 and DOTMATRIX, respectively. A general description of their approach is illustrated in **Fig. 3** using apoA-I and apoA-IV as examples.

First, a span length (or window size) must be chosen.[8] This is because every (overlapping) subsequence equal to the span length in apoA-I will be compared with every overlapping subsequence of the same length in apoA-IV. For example, suppose that we select a span length of 23 residues (see Fig. 3A). Then what occurs in overlapping subsequence comparisons is that residues 1–23 of apoA-I are compared with residues 1–23 of apoA-IV, then residues 2–24 of apoA-I are compared with residues 1–23 of apoA-IV, then residues 3–25 of apoA-I are compared with residues 1–23 of apoA-IV, and so on until the carboxy terminus of apoA-I is reached. This entire procedure is then repeated for residues 2–24, 3–25, etc. of apoA-IV until the carboxy terminus of this sequence is reached. The total number of span comparisons performed (elements in the comparison matrix) is equal to the product of the sequence lengths. The choice of a particular span length is somewhat arbitrary but should be small relative to the length of the sequences under study. Otherwise, short homologous segments will be missed. Generally speaking, however, the shorter the span, the more likely it is that a matching span will occur by chance. Empirical testing has shown that a reasonable "default" span length is 25 residues (see ref. 16) with the lower limit being 5–6 residues (27). One can choose a span length more rationally in cases where repeating sequences are suspected and there is some idea of the length of the repeated unit. In this case, the span length should be equal to or slightly exceed the repeat length for preliminary analysis. The choice of different span lengths can influence the outcome of a comparison (cf. panels B and C of Fig. 4 as well as Table 1).

For every comparison of two spans, a matching score is calculated. This matching score is simply the sum of similarity scores for each aligned pair of elements in the span. One may elect to score for exact matches (identities) or use some other criteria based upon codons or physical, chemical, or evolutionary properties of the amino acids. The matching score is then stored in an array at the position representing the *central* element of the aligned spans and this is why the span length must be an odd integer (Fig. 3A). For example, for a span of 23 residues, residue number 12 would be the center of the span with 11 residues on either side (see Fig. 3A). If the matching score for this span exceeded a predetermined threshold value (described below), a point corresponding to the *center* of the span would be plotted. In two sequences that are closely related, points representing the highest scoring spans would fall along a single main diagonal. Internally repeated sequences reveal themselves as shorter diagonals offset from but parallel to the main diagonal (Fig. 3D).

There are several ways of estimating the probability that a span has achieved a particular score by chance. McLachlan uses a probability generating function to determine the frequencies of scores in infinite random sequences with the same amino acid compositions as the real sequences under study. Dayhoff's approach is more

[7] Expressed in a more algorithmic fashion, a sliding comparison consists of sequentially offsetting one sequence with the other until the offset equals the length of the second sequence.

[8] In computer science terms, a span can be defined as a character string or a one-dimensional array. The comparison algorithm requires that the span be an odd integer.

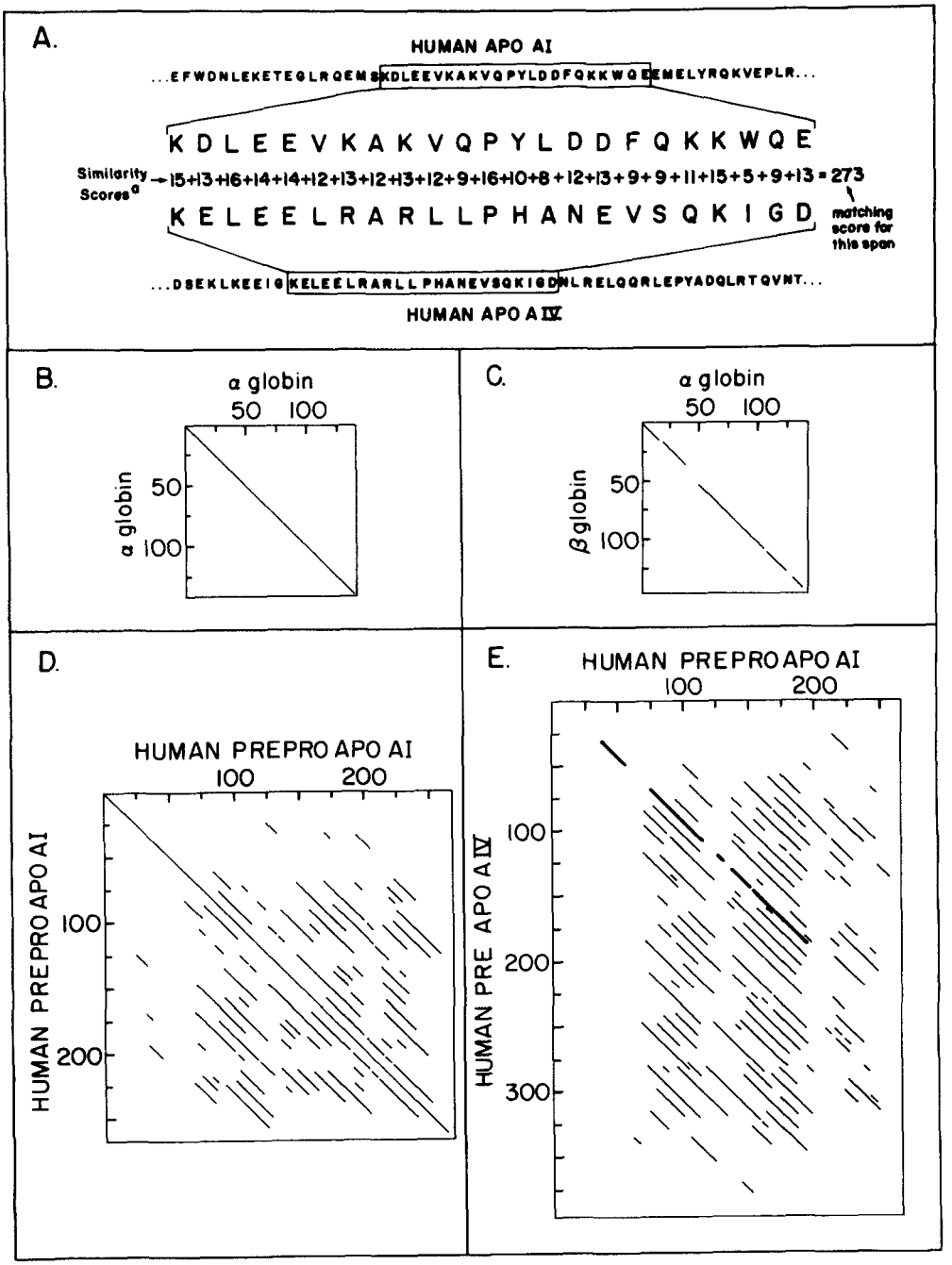

Fig. 3. Comparison matrix analysis. Panel A: Illustration of the method used to construct a comparison matrix. To generate a matrix, matching scores for every pair of overlapping spans in the two sequences are calculated. The comparison score for a span is the sum of similarity scores across the span. Span length is defined by the user. The CMPSEQ84 program (14) uses the PAM250 scale with one modification — the number 10 is added to each value in the mutation data matrix so that only positive integers are obtained. The comparison matrix method can be used to compare a protein sequence against itself (an intrasequence comparison) or it can be employed to identify similarities between two different proteins (intersequence comparison matrix). The number of spans in an intrasequence matrix equals the square of the sequence length minus a correction factor described by McLachlan (14) for incomplete spans located at the edges of the matrix. For every matching score that exceeds a threshold of statistical significance (again defined by the user), a point is plotted representing the CENTER of that span (the proline residue in this example). Panel B: Intrasequence comparison matrix of human alpha globin. A span of 23 was selected for the comparison. The threshold for plotting was set so that the probability of achieving that score by chance alone was less than 1 in 1000 (10^{-3}). Panel C: Intersequence comparison matrix of human alpha versus beta globin. Span = 23, threshold = 10^{-3}. These sequences are colinearly related. Discontinuities in the main diagonal reflect regions of relative sequence divergence; displacements of the main diagonal represent insertion or deletion mutations. Panel D: Intrasequence comparison matrix of human preproapoA-I. The shorter diagonals, which are parallel to the main diagonal, represent regions of internal homology (span = 23 and the threshold for plotting was set at 10^{-3}). Panel E: Intersequence comparison matrix of human preproapoA-I versus human preapoA-IV using a span of 23 and a threshold of 10^{-3}. The long main diagonal shared by the two sequences (indicated by the double-thickness line) reflects the colinear relationship of these two paralogous proteins.

empirical and simply uses a large number of random shuffles of the input sequences to generate a probability distribution. In both cases, one inspects the distribution to find, for example, a score that occurs with a chance frequency of only 10^{-3} and then uses this score as a threshold for plotting. This threshold seems to provide a reasonable balance of sensitivity and specificity. We illustrate below comparisons performed at different threshold levels.

An alternative approach is to begin the analysis at a very low threshold (with a high probability of chance occurrence) and continue to raise the cut-off score until the "background" disappears. This approach is analogous to washing a Southern blot at higher and higher stringencies (increasing temperature or decreasing salt concentration).[9] We prefer to begin with a threshold of 10^{-3} and then increase or decrease this value based upon preliminary results. In this manner, we know from the beginning whether there are any highly significant homologies present, and subsequent analyses can be directed toward more precise definition of these sequences. However, if a matrix plot at the initial threshold level appears to have a great deal of background "noise," then the threshold score may be increased to a value that has a lower chance probability. Alternatively, if plotting at the initial threshold reveals an absence of similar spans, then one could decrease the threshold score. The use of well-chosen comparison parameters is, as with almost every other program we describe, essential for a meaningful interpretation of the results. For example, using comparison matrices, Cheung and Chan (28) were initially unable to demonstrate repeating units in the amino acid or cDNA sequences of apoA-I. Fitch, Smith, and Breslow (29) have pointed out that the repeats in apoA-I would have been apparent if only a different set of parameters had been chosen.

Comparison matrix methods may be used to analyze the relationships between two different sequences (an *inter*sequence comparison) or to compare a sequence with itself (an *intra*sequence comparison). Intersequence comparisons are very useful in aligning two related sequences, especially in the presence of insertion and deletion mutations. Intrasequence comparisons are exceedingly helpful in identifying and visualizing internally repeated sequences. We will illustrate these two types of comparisons with specific examples, described below. But first we need to define some terms that aid greatly in the conceptualization of sequence relationships.

When two sequences have diverged from a common ancestor, they are said to be *homologous* (30). However, the term "homology" is often less rigorously applied to mean "similarity" between sequences for which an ancestral relationship has not been proved or even postulated. To avoid semantic confusion, we will use the term homology in its strictest sense. Homologous sequences can be classified more precisely according to their evolutionary history (30).

Orthologous sequences reflect the phylogenetic branching order of the species in which they are found and have identical functions. Some examples would be human α-globin and mouse α-globin or human apoA-I and rat apoA-I. *Paralogous* sequences are products of gene duplication events that were fixed prior to speciation resulting in the formation of a gene family. For example, human α-globin, β-globin, and myoglobin are paralogous sequences as are human apolipoproteins A-I, A-IV, and E. Divergence of paralogous sequences often results in evolution of new functions and/or patterns of regulation.

Fig. 3 shows two *inter*sequence comparison matrices of paralogous proteins. The comparison of α- and β-globin (panel C of the figure) illustrates that these two sequences are colinearly related but that an insertion or deletion mutation has occurred as evidenced by displacement of the main diagonal near residue 50. This matrix also shows no evidence for any internally repeated sequences at this statistical threshold (10^{-3}).

Fig. 3E is also an intersequence comparison of paralogous proteins. This comparison of human apoA-I and A-IV discloses a long main diagonal signifying that these sequences are colinearly related. Interruptions within, and displacements of, the main diagonal arise from regions of relative sequence divergence and insertion/deletion mutations, respectively. In addition, many shorter diagonals, offset from the main diagonal, are present. These represent the internally repeated sequences of apoA-I and A-IV.

Fig. 4 illustrates the effects of different threshold levels and span lengths on *intra*sequence comparisons of human pre-apoE. For the matrices in panels A, B, and D, the span was constant at 45 residues and the threshold levels for plotting were varied. The calculated cumulative probability distributions (**Table 1A**) for all three matrices are exactly the same; the plotting threshold only controls the number of spans displayed according to a chosen level of statistical significance. The matrix in Fig. 4A displays all spans with expected frequencies (chance probabilities) of $<10^{-2}$. Likewise, the matrices in panels B and D display spans with expected frequencies of $<10^{-3}$ and $<10^{-4}$, respectively. For the matrices in panels B and C, the threshold was constant at 10^{-3} and the span length was varied (panel B, span = 45 residues; panel C, span = 23 residues). In this case the probability distributions are different (cf. Table 1A and B).

The pattern of repeated sequences in human pre-apoE contains a blank zone (located within the final exon of its gene). This area marks a region of considerable sequence divergence. Nevertheless, when the repeats fade away, they generally reappear on the same diagonal. This is even more apparent when longer spans or lower thresholds

[9] We thank T. B. Rajavashisth for suggesting this analogy to us.

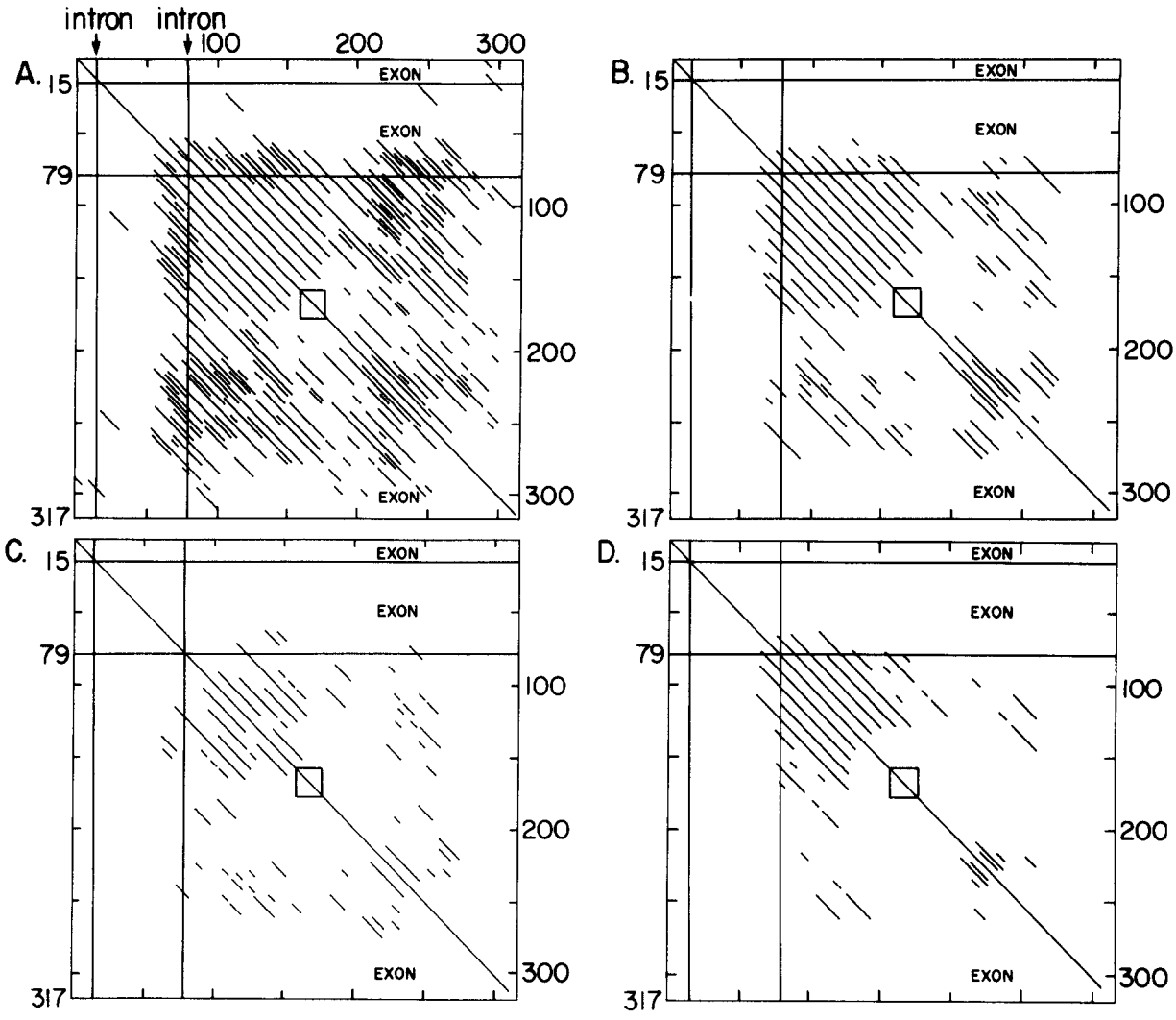

Fig. 4. Comparison matrix analysis of human preapoE. The amino acid sequence of the primary translation product of apoE mRNA was compared against itself and using two different span lengths and three separate plotting thresholds. Numbers along the upper abscissas and the right ordinates represent residue numbers for the preprotein, while numbers along the left upper ordinates represent the residues interrupted by introns (23, 31). The number at the lower left corner of each plot represents the COOH-terminus residue of human preapoE. The LDL receptor binding domain, as defined by Innerarity et al. (32) and Weisgraber et al. (33), is enclosed by a box. Panel A: Matrix generated using a span = 45 and a threshold score for plotting that, in infinite random sequences with the same amino acid compositions as preapoE, would be achieved by chance alone less than 1 in 100 times (10^{-2}). Panel B: Span = 45, threshold = 10^{-3}. Panel C: Span = 23, threshold = 10^{-3}. Panel D: Span = 45, threshold = 10^{-4}. These comparison matrices show that there are repeated sequences encoded by both the third and fourth exons of the human apoE gene. Furthermore, by altering the threshold values it is apparent that the repeat system is continuous even though at the higher threshold values a central blank zone can be appreciated. This zone is indicative of sequence divergence within the repeat system. The presence of insertions and deletions is further illustrated by the alignments shown in Fig. 5.

are used (compare panels A, B, and D of Fig. 4). This observation supports the overall continuity of the repeat system in apoE.

The figure also illustrates two additional points. First, the repeated sequences in apoE span intron–exon boundaries, just as they do in the apoA-I and A-IV genes (8, 20). The implications of this phenomenon for apolipoprotein evolution have been discussed elsewhere (8). Second, the apoB,E (LDL) receptor-binding region of human apoE (residues 140–160 in the mature plasma protein, or residues 158–178 in pre-apoE) is located near the junction between a strong repeated sequence domain and the central divergent zone of exon 4 (see the boxed area in Fig. 4). We have previously noted that the residues that constitute the binding domain have the potential to assume an amphipathic α-helical conformation (6–8). The comparison matrices in Fig. 4 demonstrate that this *specialized* domain in human apoE is, nevertheless, a component of the periodic structure of the protein.

Besides comparison matrices, there are other ways to

TABLE 1. Self-comparison statistics for apolipoprotein E

Score (S)	No. of Spans of Score S or Higher	Observed Frequency	Expected Frequency	Ratio Observed/Expected
A. Span 45, mean 426, SD 22				
426	49,378	0.493	0.514	.959
448	16,514	0.165	0.915×10^{-1}	1.80
→ 464	4,384	0.438×10^{-1}	0.104×10^{-1}	4.21
470	2,898	0.289×10^{-1}	0.388×10^{-2}	7.45
→ 477	1,698	0.170×10^{-1}	0.111×10^{-2}	15.32
→ 489	572	0.571×10^{-2}	0.102×10^{-3}	55.98
492	440	0.439×10^{-2}	0.536×10^{-4}	81.90
514	62	0.619×10^{-3}	0.320×10^{-6}	1934.0
539	2	0.200×10^{-4}	0.431×10^{-9}	46403.0
B. Span 23, mean 217, SD 14				
217	50,410	0.503	0.493	1.02
231	15,710	0.157	0.118	1.33
245	3,234	0.323×10^{-1}	0.110×10^{-1}	2.94
→ 256	742	0.741×10^{-2}	0.995×10^{-3}	7.45
264	148	0.148×10^{-2}	0.137×10^{-3}	10.8
273	40	0.399×10^{-3}	0.123×10^{-4}	32.4
285	2	0.200×10^{-4}	0.378×10^{-6}	52.9

Comparison parameters were the same as in Fig. 4. Expected frequencies were calculated from the double matching probability with edge corrections (14, 25). Score values of 464, 477, and 489 (indicated by arrows in panel A) have expected frequencies (chance probabilities) of $< 10^{-2}$, $< 10^{-3}$, and $< 10^{-4}$, respectively. These scores were used as threshold levels for plotting in panels A, B and D of Fig. 4. The score value of 256 indicated by the arrow in panel B of the table has a probability of $< 10^{-3}$ and was the threshold selected for plotting the data displayed in panel C of Fig. 4. As stated in the text, shorter spans have a higher probability of chance similarity than longer spans. This is particularly evident in the ratios of observed to expected frequencies.

identify and define repeated sequences. We have used (7) the NBRF RELATE program (26) to study the repeated sequences in rat apoA-IV. This program can be applied to either protein or DNA sequences. The SEQ program (34), marketed by IntelliGenetics Inc.,[10] can be used to search for internal similarities in nucleotide sequences. Karathanasis et al. (22) and Das et al. (23) have used the latter approach in analyzing repeated sequences in two human apolipoprotein genes. They described six 66-nucleotide repeats in exon 4 of the apoA-I gene (22) and eight 66-nucleotide repeats in exon 4 of the apoE gene (23). Other analytic techniques have shown that the repeat systems in these genes are actually much larger than this (6, 8, 35). Thus the SEQ program appears to underestimate the true extent of the repeat pattern in these sequences, possibly due to a suboptimal set of program parameters[11] or the use of DNA rather than protein sequence data.

[10] IntelliGenetics, Inc., 124 University Ave., Palo Alto, CA 94301.

[11] The intrasequence comparison option of SEQ requires the selection of nine parameters: AfterMismatch, Expect, GUPair, LoopOut, MaxLoop, MinLoop, MaxDist, MinMatch, and PercentMatch. The settings of these various parameters can greatly modify the results obtained as well as their statistical significance. Values of the parameters used for the apoA-I (22) and apoE (23) work were not reported.

Consensus sequences and correlation analysis

A *consensus sequence* is defined as an *idealized* sequence in which each position represents the nucleotide base or amino acid residue most often found when many *real* sequences are compared (12). Expressed more formally, each element of a consensus sequence is the statistical *mode* of a set of observed elements in corresponding positions of aligned sequences. With a small number of sequences, there may be no mode. Alternatively, there may be more than one mode, such as when dealing with a bimodal distribution. A mode is rarely used as a basis for performing statistical inferences (36). Consensus sequences may also be a poor basis for experimental studies (see below).

Several common problems are encountered in applying the concept of consensus sequences to apolipoproteins. First, consensus sequences are imperfect *approximations* of real sequences. Comparisons of approximations can generate results which are difficult to interpret or are of uncertain validity. Quantitative comparisons (fractional "homologies") among consensus nucleotide sequences for apoA-I, A-IV, and E have been performed (17). These analyses implied that apoA-IV and E were more closely related than apoA-IV and A-I. However, statistical analyses of the *actual* sequences resulted in just the opposite conclusion: the repeat systems in apoA-I and A-IV are

much more highly conserved with respect to each other than either is to that of apoE (6–8, 35).

Second, reducing a series of repeats to a single consensus sequence sometimes has the effect of obscuring important variations among individual repeat units. Unique characteristics of particular repeats tend to get "averaged out" of a consensus and such unique features can be of prime biological importance. This phenomenon is no more evident than in the case of apoE in which one of the repeated docosapeptides represents the LDL receptor-binding domain (see Fig. 4 and **Fig. 5**). Highly divergent repeats may not even be recognizable as repeats if the criterion for their identification is a high degree of similarity to a consensus sequence.

Third, some investigators have inferred consensus amino acid sequences from consensus nucleotide sequences. In our experience, using amino acid sequences directly (to arrive at a consensus) results in consensus sequences that are demonstrably more accurate representations of conserved physical–chemical properties (see below). Consensus sequences derived from protein sequence data tend to have fewer ambiguous elements (multimodal characteristics). Even so, one should not expect consensus sequences, a priori, to mimic properties of real sequences. Based upon calculations of averaged hydrophobicities, for example, a peptide based on the consensus sequence for rat apoA-IV (5) would *not* be expected to form stable complexes with lipid (as defined by their behavior in the ultracentrifuge (J. T. Sparrow, personal communication).

One final problem with consensus sequences is that they are often equated with ancestral sequences. This equation is generally not valid. As previously stated, a consensus sequence is simply a set of those elements most frequently found in corresponding positions among a collection of *extant* sequences. To properly reconstruct an ancestral sequence, one first needs to construct a tree depicting the phylogenetic relationships among the extant descendents (30). Such a tree is usually based upon some type of distance metric (19). Branch points (nodes) of the tree represent hypothetical ancestral sequences usually arrived at by applying the *parsimony criterion* which directs us to prefer the phylogeny that requires the fewest nucleotide substitutions (39). However, Felsenstein (40, 41) has pointed out that maximum parsimony is not necessarily equivalent to maximum likelihood. Indeed, to reconstruct a "true" ancestral sequence, one would need to know the base composition and/or codon usage frequencies in the (extinct) common ancestor. Fundamentally, natural selection operates on the phenotype (proteins), not the genotype (DNA). Thus it is possible that present-day DNA sequences could be quite different from their ancestors even though both might code for similar amino acid sequences. Therefore, the derivation of a hypothetical ancestral sequence is not so simple as merely generating a consensus sequence.

Despite their many limitations, consensus sequences can occasionally be quite useful. For example, we have used consensus sequences to characterize the magnitude and extent of periodic structure in *rat* apoA-IV (5) by correlation analysis (42). With this technique, relationships between protein sequences can be expressed by correlation coefficients. First a peptide (e.g., consensus sequence) and protein for comparison are converted to series of numerical values based on one or a number of physical, chemical, structural, and evolutionary properties of the amino acids. Kubota et al. (42) originally defined 10 parameter sets which we have extended to 20 (**Table 2**).[12] The shorter sequence is then sequentially compared with the longer sequence at increments of one residue and a correlation coefficient is computed for each of these overlapping spans. When the peptide is aligned with a similar subsequence of the protein, this results in a positive correlation. When all of the values are plotted as a function of sequence position, a *correlogram* is obtained. Periodicity in protein structure is indicated by the presence of multiple correlations with spacing a function of the repeat length (see below).

Correlograms can be considered "spectral" analyses of protein properties (43) with individual parameter sets being analogous to wavelengths in an electromagnetic spectrum. It is often of interest to *decompose* a spectrum into its component wavelengths. One can "decompose" a multidimensional protein correlogram by performing the analysis with individual parameter sets rather than using combined values. In this manner, properties that provide the greatest contribution to a correlation can be identified. For example, a particular functional characteristic of a sequence might depend more on the bulkiness and polarity of residues than on their potential to assume α-helical conformations. Relative mutability, however, would be the index characteristic most suitable for evolutionary studies. Of course, not all of the parameter sets are independent with respect to each other. For instance, one would expect a strong connection between hydrophobicity and solvent-accessible surface area and an inverse correlation between α-helix and β-sheet potential.

The complete human apoA-IV sequence (9, 17) now provides us with a very large data set of repeat units. From this data set, we have derived a unified consensus sequence for apoA-IV (**Table 3**) and used it to study the periodic structure of apoA-IV as a function of various amino acid properties.

The periodicity in human apoA-IV is illustrated by the series of correlograms displayed in **Fig. 6** and **Fig. 7.** The unified apoA-IV consensus sequence defined in Table 3

[12] Copies of the program and parameter sets for computing Kubota correlograms are available from Mark Freeman with the complete FORTRAN source code.

was used to scan the mature plasma protein. The combined parameters of α-helical potential, relative mutability, and hydropathy (as defined by Kyte and Doolittle, ref. 52) were selected to generate the correlogram shown in panel A. These parameter sets represent three different properties of amino acids, a structural property, an evolutionary property, and a physico-chemical property. An 11- as well as a 22-residue periodicity can be discerned in the plot. Threshold levels of significance for these correlations can be estimated by repeating the analysis using randomized versions of the sequences. In this case, the highest correlation coefficient obtained was 0.345.

By comparison, panels B and C of Fig. 6 show the results obtained when only α-helical potential and relative mutability, respectively, were used. The periodicity is much more difficult to appreciate, a large number of negative correlations are seen, and the baselines are quite wide. When the Kyte-Doolittle hydropathy scale was used (panel D of Fig. 6), the 11-residue periodicity was clearly evident and could be seen to extend throughout almost the entire length of the sequence. The conclusion from these studies is that combined values often improve the signal-to-noise ratio in correlation analyses, but at the same time may obscure important features of the periodic structure. In addition, the magnitude of the correlations may be affected (compare panels A and D). When we decomposed the spectrum shown in panel A, it became clear that the component that contributes most to apoA-IV periodicity is the relative hydropathy of amino acid residues (panel D).

Correlation analysis can also be used to compare and contrast the properties of different types of parameter sets. We did this for three different hydrophobicity scales[13] using human apoA-IV and the unified apoA-IV consensus sequence (Fig. 7). The Nozaki-Tanford scale is based upon free energy changes associated with the transfer of amino acid side chains from a polar to nonpolar environment (54, 56, 57). The Bull and Breese (53) scale derives from surface tension measurements. The scale described by Rose et al. (55) is based upon the solvent-accessible surface area of residues in proteins of known structure.

The Bull and Breese scale is somewhat better than the Nozaki-Tanford scale at revealing sequence periodicity in apoA-IV (compare Fig. 7A with 7B). However, in the former case the baseline is quite wide. The correlogram computed using solvent-accessible surface area as the index characteristic is the best of the three in detailing the basic 11-residue periodicity with an excellent signal-to-noise ratio (Fig. 7C). Both the Rose and Kyte-Doolittle scales (Fig. 6D) illustrate the organization of undecapeptide repeat units with greater clarity than the other two systems.[14] The significance of variations in the *magnitudes* of the correlation coefficients computed using the different scales remains to be determined. Repeating these analyses using *actual* repeat sequences rather than a consensus sequence may shed further light on this issue.

Computer-assisted prediction of protein structure and function

The rate of discovery of new protein sequences (often inferred from DNA sequence data) has long since surpassed the abilities of biochemists to purify and characterize every newly determined sequence using classical techniques. In some cases, the proteins themselves may not even exist as expressed gene products; a particular primary structure may only be an hypothesis deduced from an open reading frame in a DNA sequence. The ultimate goal of computer-assisted analysis is to predict the structure and function of proteins and enzymes from first principles applied to sequence data alone. Although this ambitious goal is many years away, some progress has been made. We can at least generate experimentally testable hypotheses from analogies with sequences of known structure or function.

For example, structural characteristics of rat apoA-IV, and its homology with human apoA-I, indicated that apoA-IV should be capable of activating lecithin:cholesterol acyltransferase (LCAT) and also that the apoA-IV and A-I genes would be linked (5). The LCAT-activating ability of apoA-IV was demonstrated by Steinmetz and Uterman (59). Karathanasis (17) and Elshourbagy et al. (9) have shown that the A-I and A-IV genes are indeed closely linked. Deletion mutations and nonconservative amino acid substitutions in the sequence of rat apoA-I suggested that this protein might be functionally deficient (relative to human apoA-I) in its ability to bind lipids and activate LCAT (6). Competition studies by Rifici, Eder, and Swaney (60) showed that human apoA-I is able to displace rat apoA-I from lipid vesicles, indicating decreased lipid-binding potential for the rat protein. Pownall, Pao, and Massey (61) showed that rat apoA-I is only about 50% as effective as human apoA-I in activating either human or rat LCAT.

A single mammalian apolipoprotein has yet to be crystallized. In fact, the available evidence seems to indicate that the apolipoproteins do not possess stable tertiary structures at all and that most of their functional properties depend on secondary structure (although there may be weak, cooperative interactions between different

[13] We thank F. J. Kezdy and J. T. Sparrow for their suggestions regarding alternate hydrophobicity scales.

[14] Recent results of Lipman, Pastor, and Lee (58) indicate that fractional accessibilities of amino acid residues may be more reliable than their hydrophobicities for predicting structural and antigenic features of protein subsequences.

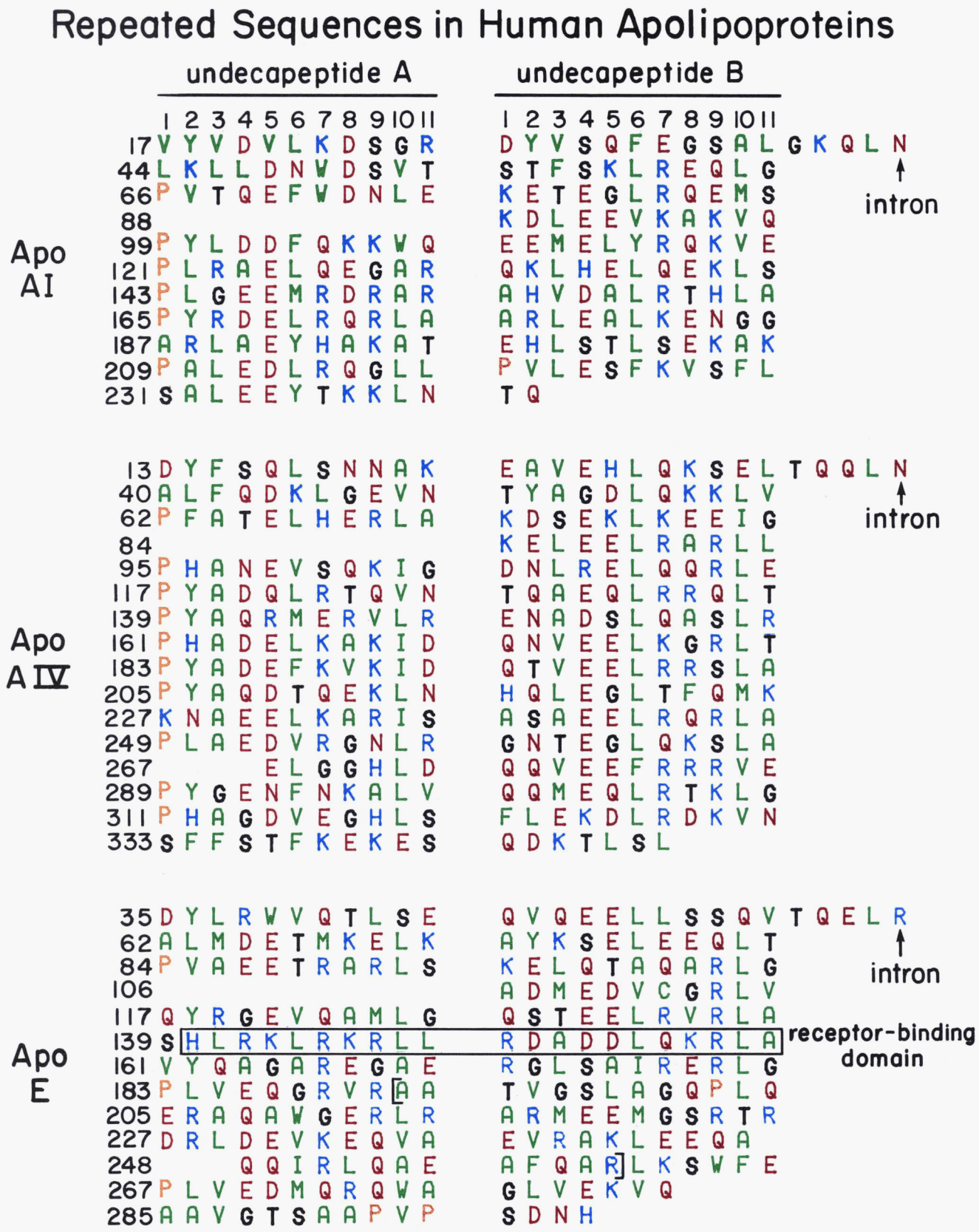

Fig. 5. Organization of repeated sequences in human apoA-I, apoA-IV, and apoE. Sequence positions along the fundamental undecapeptide repeat unit are numbered 1 to 11. Numbers along the left margin are residue numbers for the mature plasma proteins. Amino acids are color-coded by hydropathy index and charge as previously described (5). Hydrophobic amino acids are indicated in green. Acidic amino acids as well as their amide

TABLE 2. Parameter sets used in correlation analysis

Parameter	Reference
Relative mutability	Dayhoff (16)
Bulkiness	Zimmerman, Eliezer, and Simha (43)
Polarity	Zimmerman, Eliezer, and Simha (43)
Hydrophobicity	Jones (44)
pK, α-amino group	Sober (45)
pK, carboxyl group	Sober (45)
α-Helix potential	Chou and Fasman (46)
β-Sheet potential	Chou and Fasman (46)
Non-bonded energy per residue	Oobatake and Ooi (47)
Non-bonded energy per atom	Oobatake and Ooi (47)
Hydrophilicity	Hopp and Woods (48)
Hydrophilic score	Levitt (49)
β-Turn score	Chou and Fasman (46)
Free energy	Wolfenden et al. (50)
Fraction buried	Chothia (51)
Hydropathy index[a]	Kyte and Doolittle (52)
Hydrophobicity	Bull and Breese (53)
Hydrophobicity[b]	Edelstein et al. (54)
Standard state accessibility	Rose et al. (55)
Solvent-accessible surface area	Rose et al. (55)

[a]The Kyte-Doolittle hydropathy scale is based on the combined values of Wolfenden et al. (50) and Chothia (51).

[b]These parameters represent values combined from Nozaki and Tanford (56) and Steinberg and Thornton (57).

regions of secondary structure (ref. 62). The apolipoproteins have free energies of denaturation of less than 4 kcal/mole (62–65). ApoA-IV seems to possess the greatest thermodynamic instability of all (65). Calculations based on circular dichroism (CD) spectra in the presence of increasing concentrations of guanidine hydrochloride resulted in a value for the free energy of denaturation of only 0.2 kcal/mole (65). In contrast, corresponding values for some representative globular proteins are: myoglobin and lysozyme (9 kcal/mole), alpha-chymotrypsin (11.4 kcal/mole), ribonuclease (16.3 kcal/mole), and β-lactoglobulin (22.3 kcal/mole) (66).

derivatives are represented by red (glutamic acid and aspartic acid have the same hydropathy index (−3.5) in the Kyte Doolittle scale as their uncharged amide derivatives, glutamine and asparagine). Basic amino acids are indicated in blue. Small, neutral amino acids are uncolored. Proline is indicated in yellow to emphasize the regularity of its occurrence in the first position of many of the repeat units. The organization of repeated sequence blocks in apoA-I and apoA-IV have been defined from analyses reported in earlier publications (8, 9). These sequences illustrate the point that the docosapeptide repeat units are tandem arrays of two related undecapeptides (labeled A and B). Each 11-mer is more similar to the 11-mer once removed than to the adjacent 11-mer (e.g., undecapeptides B do not begin with proline). ApoE lacks the unusually even distribution of landmark proline residues in apoA-I and A-IV. However, because all three of these proteins diverged from a common ancestor (6–8, 19, 35), and because apoA-IV has the most highly conserved repeats, it is possible to delineate the repeats in apoE by alignment with apoA-IV. In two sequences of unequal length whose repeats have diverged, there are many possible alignments that occur as multiples of the repeat length (eleven residues in this case). Thus stringent alignment criteria are necessary to define the relationship that truly reflects the phylogenetic history of the repeats. With this caveat in mind, and matching only identical residues between the sequences, we defined the periodic structure of apoE by alignment with apoA-IV. The precise placement of gaps between residues 192 and 260 (enclosed by brackets) varied with different alignment algorithms and parameters. The first repeat of apoE is found in its third exon and has undergone length divergence so that it is 27 residues long. An intron interrupts the 27th codon of this repeat unit just as in apoA-IV (and A-I). Following two exact 22-mers, there is a "displaced" 11-mer (again just as in apoA-IV and A-I) which illustrates that each docosapeptide unit is a tandem array of two related undecapeptides. Unlike the repeated docosapeptides in apoA-I and A-IV, most of which begin with a proline residue, only three of the approximately eleven repeats in apoE begin with this amino acid. By arranging the apoE sequence as shown, it is apparent that the LDL receptor-binding domain is a derivative of a docosapeptide repeat unit. Several positions along this sequence share some of those residues that are most highly conserved among the repeats (columns A7, A10, B4–5–6, B9, and B10). It is interesting to note that the positively charged residues in columns A7 and B9 (corresponding to Arg 145 and Arg 158 in the mature plasma protein) are thought to be the critical ones for receptor binding (32, 33, 37). The receptor-binding domain of apoE has the highest density of positively charged residues among all of the repeat units in the protein with a total of nine arginines, lysines, or hisitidines in its 22-residue span. The number of these basic residues in the other docosapeptide repeat units ranges only from two to six. The LDL receptor itself contains a repeated sequence domain thought to represent the binding site for apolipoproteins E and B (38). This domain consists of eight repetitions of a 40 amino acid segment with a high content of negatively charged residues. Using methods described in this paper, we determined that these repeats in the LDL receptor bear no discernible structural or ancestral relationships to the repeated sequences in apoE. This is a case of two unrelated families of repeated sequences that have evolved to interact with one another as receptor and ligand.

TABLE 3. Unified consensus sequence for apoA-IV docosapeptide repeats

Undecapeptide A											Undecapeptide B										
1	2	3	4	5	6	7	8	9	10	11	1	2	3	4	5	6	7	8	9	10	11
*		*												*		*				*	
P	Y	A	E	E	L	K	G	K	L	N	Q	N	V	E	E	L	R	R	R	L	A
68	32	61	25	25	46	18	21	29	46	25	27	33	27	70	20	83	43	17	20	53	13
				D		Q									D				Q		T
				25		18									20				20		13
																					V
																					13

All of the docosapeptide repeats in human and rat apolipoproteins A-IV are aligned according to Boguski et al. (5) assuming a one-to-one correspondence between residues 13-332 of the mature plasma proteins (9). In all, there are 28 undecapeptides A and 30 undecapeptides B. The frequences of residues in corresponding positions of the repeat units were determined and are expressed as percentages. Residues that are highly conserved (>50%) are indicated by an asterisk (see also Fig. 5). Residues in the consensus sequence above represent modal elements as described in the text. In cases where there were two or more modes, all of the equally predominant residues are shown, but with the residue that is most common in the human sequence shown first. (In human apoA-IV, the repeat units are more highly conserved with respect to each other than in the rat sequence (ref. 9, and unpublished observations).)

Researchers have thus turned to noncrystallographic methods of structure identification, primarily circular dichroism and empirical prediction. (Synthetic peptides have also been extensively used to test structural hypothesis and some of these studies will be discussed below.) Unfortunately, the accuracy of methods for the empirical prediction of secondary structure is discouragingly low. Even the best predictive schemes are accurate, at most, approximately 50% of the time and the most widely used method (Chou-Fasman[15]) is not the most accurate (67). Two characteristics of predictive methods would appear to further limit their accuracy when applied to the apolipoproteins. First, conformational parameters for structure prediction are biased in favor of *globular* proteins and may not be directly applicable to non-globular proteins. In addition, predictive methods generally do not take into account the influences of ligand binding. Specifically with regard to apolipoproteins, it is well known that helix formation (as defined spectroscopically) can be induced by lipid binding. Thus empirical prediction of protein conformation in the presence of bound ligand may be relatively meaningless. The method of Garnier, Osguthorpe, and Robson (68) may be preferable for apolipoprotein studies because of its ability to incorporate decision constants, based upon CD spectra, into the prediction.

In the absence of supporting experimental evidence, great caution must be exercised in interpreting the results of a structural prediction. For example, our studies on rat apoA-IV (5, 7) suggested that about 56% and 15% of the residues existed in α-helix and β-sheet conformations, respectively. These results appeared to correlate well with the observed content of α-helix (52-54%) and β-sheet (11%) estimated from CD spectra (65, 69). Even so, one cannot draw any firm conclusions about the precise locations of residues that may exist in these conformations. Indeed, regions of predicted α-helix in rat apoA-IV do not display a precise one-to-one correspondence with the proline-punctuated docosapeptide repeat units (5, 7). It should also be noted that other CD studies on rat and human apoA-IV resulted in quite different estimates of the α-helical content (33-37%) of these proteins (64). Thus, even under the best of circumstances, a wide gulf may exist between prediction and reality.

The helical hydrophobic moment (70) was originally devised (71) to quantify the concept of the amphipathic (amphiphilic) helix which had previously been qualitatively represented by "helical wheel" diagrams (72). Calculation of hydrophobic moments has now been used extensively to analyze apolipoprotein sequences (73-76). However, the hydrophobic moment is a *vector* sum and its calculation presumes a knowledge of the underlying protein structure. In cases where the atomic coordinates are *not* known, the hydrophobic moment can be estimated based upon empirical predictions of secondary structure (71), and most of the studies on apolipoproteins have been based on structural predictions or assumptions. Thus the same caveats applied to the interpretation of secondary structure predictions must also be applied to hydrophobic moment analyses, except in such cases where independent (preferably direct experimental) evidence of helical conformation exists. It is interesting to note that Eisenberg, Weiss, and Terwilliger (71) specifically excluded the apo-

[15] Kabsch and Sander (27) have noted that ambiguities in the Chou-Fasman method often give different results in the hands of different people and thus the method is not programmable without extension or modification. An algorithm is not necessarily the same as its implementation!

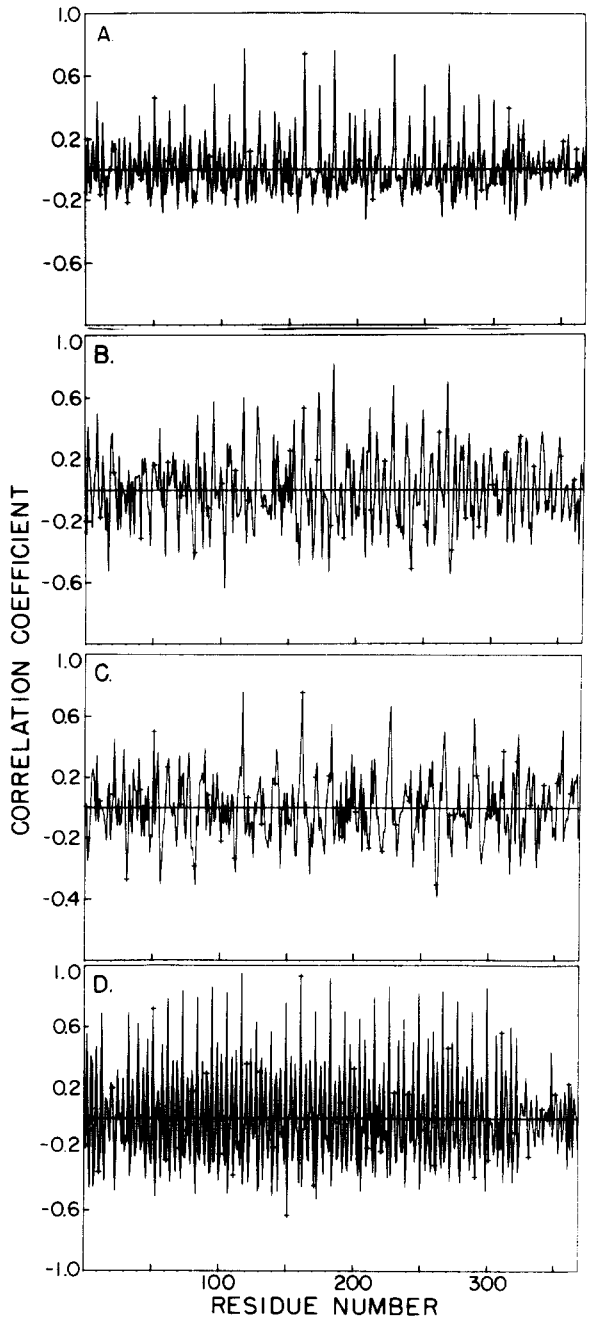

Fig. 6. Correlation analysis of repeated sequences in human plasma apoA-IV. This technique is based on an algorithm described by Kubota and coworkers (42). Protein sequences are converted to sequences of numerical values based on a number of quantitative properties of amino acids. The shorter test sequence is compared in a sliding fashion with the longer sequence at increments of one residue. A strong positive correlation indicates that the shorter sequence is aligned with a highly similar region (similarity is defined in the context of the parameter(s) selected for study). The correlation coefficient is plotted at a position corresponding to the NH_2-terminal residue of the segment of the larger sequence that is being compared to the shorter test span. In this figure, the unified apoA-IV consensus sequence (Table 3) was used to scan the sequence of human apoA-IV (taken from ref. 9). Correlation coefficients were computed for each of the overlapping spans based on individual as well as combined parameters. The small horizontal lines perpendicular to and intersecting some peaks represent spacing marks that occur every five residues. Panel A: Correlogram generated using a combination of α-helical potential (46), relative mutability (16), and hydropathy (defined by Kyte and Doolittle, ref. 52) as parameters. Panel B: Correlogram produced when alpha helical potential alone was used. Panel C: Results obtained when relative mutability was the only parameter. Panel D: Pattern of repeats defined using the hydropathy index of Kyte and Doolittle.

lipoproteins from their original study due to the lack of direct information on their three-dimensional structures.

Despite these caveats, when the vector addition method proposed by Eisenberg and colleagues has been applied to define the amphipathicity of α-helical segments in apolipoproteins, the α-helix amphipathicities, as reflected in the mean α-helical hydrophobic moments (u_H), are similar to those in water-soluble, globular proteins which do not interact with lipids (74). This suggests that the high surface activity and affinity for the lipid–water interface arises not from unusual amphipathic properties of individual α-helical segments in apolipoproteins, but rather from the *cooperative* effect of several α-helices with moderate u_H values. Apolipoproteins contain many helical segments per molecule. When amphipathicity is averaged across all helices in the molecule to give $\bar{u}_H$, the product (fraction of α-helix $\times$ $\bar{u}_H$) gives a good correlation with their surface activity (74, 75). This result emphasizes the need to consider the cooperative effects of several helices in treating the functions of apolipoproteins.

We now turn to two important aspects of apolipoprotein structure that often appear to be incompletely understood: 11-residue repeat units and the role of proline residues. The length and significance of a repeating motif may differ according to the context in which it is defined. The distribution of observed α-helix lengths in globular proteins ranges from about 5 to 25 residues (77) with the average length corresponding to 11 residues or three helix turns (78). Thus repeated sequences that are multiples of 11 residues would be favorably accommodated in helical domains (2, 5). However, as discussed below, undecapeptides are probably too short to be the *functional* units of apolipoproteins.

Many of the docosapeptide repeat units in apolipoproteins A-I, A-IV, and E are "punctuated" by proline residues (Fig. 5). The nature of this punctuation is the assumption that proline residues are incompatible with helical conformations. Proline residues can and do occur as components of helices in proteins of known crystal structure (e.g., ref. 15)[16]. Nonetheless, they necessarily disrupt an α-helix. This is because the conformation in an α-helix is precisely defined with residue i to i + 3 hydrogen-bonding. Because proline is an imino acid it cannot participate in such hydrogen-bonding.

[16] We thank L. J. Banaszak for first pointing this out to us.

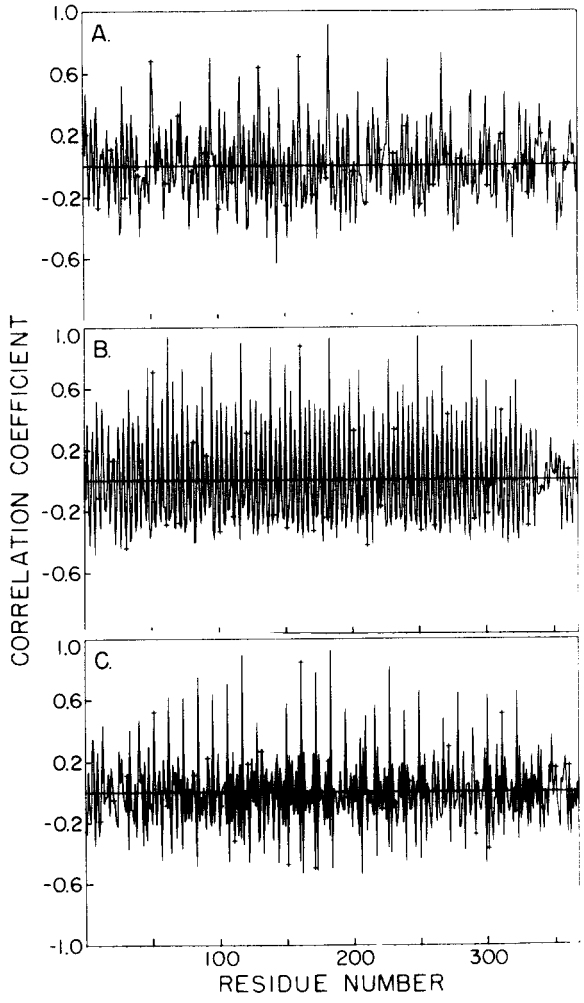

Fig. 7. Comparison of correlograms obtained using the unified apoA-IV consensus sequence with a variety of different hydrophobicity scales. As in Fig. 6, the unified apoA-IV consensus sequence derived from the data presented in Table 3 was utilized to scan the sequence of human plasma apoA-IV at increments of one residue. Panel A: Correlogram obtained using the Nozaki-Tanford hydrophobicity scale (56). Panel B: Correlogram obtained with values taken from the hydropathy scale of Bull and Breese (53). Panel C: Correlogram obtained using solvent-accessible surface area (55) as the index characteristic.

Glycine and asparagine residues are even stronger "helix-breakers" than proline, according to Chou–Fasman rules (46). In apolipoproteins A-I, A-IV, and E, glycine and asparagine residues are more than twice as abundant as proline residues (Fig. 5), yet this fact is rarely addressed in discussions of apolipoprotein structure. Furthermore, we are not aware of any work with synthetic peptides that has investigated the potential influences of glycine or asparagine residues on the properties of amphipathic helices.

The assumption of proline as helix-breaker is probably valid for the apolipoproteins. There is experimental evidence that supports this view (discussed in the following section). Also, in those proteins where prolines *do* occur as components of α-helices, it is likely that necessary and stable *tertiary* interactions override steric constraints at the secondary structure level (15). As discussed previously, the apolipoproteins do not seem to possess stable tertiary structures, as evidenced by their low free energies of denaturation (63–65), which can be an order of magnitude less than those of globular proteins (64). Thus one might expect the influence of tertiary interactions on apolipoprotein secondary structure to be minimal. Lastly, the strong evolutionary conservation of proline residues in apolipoprotein repeat units argues for an important structural role.

PROTEIN ENGINEERING AND THE FUNCTIONAL ANALYSIS OF APOLIPOPROTEIN REPEATS

With the implementation of methods for efficient peptide synthesis, it has been possible to directly analyze the properties of model oligopeptides with amphipathic (or amphiphilic) secondary structures. These studies provide a strategy for formally assessing the physical and functional characteristics of "real" repeats as well as idealized (consensus) sequences in apolipoproteins. In other words, elements that were initially defined by computational techniques can be "audited" by the peptide engineer.

Recently, correlations between repeat length and function have been critically assessed for human apoA-I. A 22-amino acid oligopeptide containing the exact sequence of residues 144–165 is capable of activating LCAT but does so to a lesser extent than the intact A-I polypeptide (79). A synthetic docosapeptide:

$$NH_2\text{-}\overset{1}{\text{Pro}}\text{-Lys-Leu-Glu-Glu-Leu-Lys-Glu-Lys-Leu-Lys-Glu-Leu-Leu-Glu-Lys-Leu-Lys-Glu-Lys-Leu-}\overset{20}{\text{Ala}}\text{-COOH}$$

engineered for "optimal" amphipathic helical secondary structure, but having no sequence identities with any domain in apoA-I, had a higher affinity for phospholipid surfaces than did the naturally occurring apoA-I segment (i.e., the peptide representing residues 144–165, ref. 79–81). The physical properties of this "natural" 22-residue peptide, as well as the docosapeptide with epitomized amphipathicity, were compared to the physical properties of a 44-amino acid oligopeptide encompassing residues 121–164 of apoA-I (62). The surface behavior of the tetratetracontapeptide more closely resembled the surface behavior of plasma apoA-I than did either of the 22-residue peptides (see **Table 4**). However, it was unclear whether this "improvement" in surface behavior was due to the increase in chain length per se or whether it reflected the influence of the second, adjacent docosapeptide. Nakagawa et al. (82), therefore, synthesized a

44-residue peptide composed of two identical copies of the idealized docosapeptide. Table 4 demonstrates that linkage of the two segments had a variety of physical–chemical consequences, each one of which made this oligopeptide more similar to intact plasma apoA-I. For example, its α-helicity in 50% trifluorethanol was greater than any of the 22-residue peptides and was very similar to that of apoA-I. In addition, it had a higher tendency to form peptide micelles in aqueous solution than either of the docosapeptides. Finally, the limiting molecular area of the peptide adsorbed at amphipathic interfaces closely resembled that of apoA-I. The limiting molecular area for the *model* docosapeptide at an air–water interface was 22 Å per amino acid (80). This indicated that the peptide was not fully helical and that it contained segments with random coil structure. The 44-residue model peptide, on the other hand, occupies 14–16 Å per amino acid at this interface which is precisely the value obtained with intact plasma apoA-I (83).

Analogous experiments were performed by Anantharamaiah et al. (84) and Chung et al. (85). These workers constructed a model octadecapeptide with amphipathic characteristics:

NH_2 ^{1}Asp-Trp-Leu-Lys-Ala-Phe-Tyr-Asp-Lys-Val-Ala-
Glu-Lys-Leu-Lys-Glu-Ala-^{18}Phe-COOH

as well as a proline-punctuated dimer (octadecamer-Pro-octadecamer). Both peptides formed complexes with dimyristoyl phosphatidylcholine (DMPC) but the covalently linked dimer had greater lipid affinity than the monomer. This was presumed to reflect a cooperative effect of the two linked prototypic amphipathic helices. When incubated with unilamellar phosphatidylcholine egg vesicles, the proline-linked dimer proved to be a potent LCAT activator, more so than the octadecapeptide or, for that matter, apoA-I (the 37-residue synthetic peptide had 140% of the activity of plasma A-I at a 1:1.75 peptide/egg PC ratio). Chung et al. (85) also observed that the LCAT activating capacity of the proline-punctuated dimer correlates with its ability to convert preformed egg phosphatidylcholine vesicles to protein annulus bilayer discs, an effect which neither A-I nor the octadecapeptide alone had.

When considered together, the peptide studies summarized above suggest that the functional repeat could be operationally defined as being 44 residues long. This is not inconsistent with the results obtained from the computational analyses presented earlier, namely that a primordial evolutionary unit specifying an 11- or 22-residue segment underwent intraexonic expansion to generate large segments of the modern apolipoproteins. Rather, these results underscore the need to consider a repeat unit in several contexts (i.e., as a structural, functional, or evolutionary unit).

What conclusions can be made from these studies about the significance of proline punctuation? The break introduced in the 44-residue peptide by proline allows the hydrophobic faces of the α-helices to remain in phase (82). Moreover, the central location of the proline residue results in a mildly concave hydrophobic surface that is well suited to fit the geometry of the HDL surface. The surface of human HDL_3 is (relatively) highly curved reflecting its radius of 40–50 Å. Calculations of the concavity of the proline-punctuated 44-residue repeat indicate that it closely matches the curvature of the HDL surface (83). The proline-punctuated amphiphilic docosapeptides found in apoA-I are not unique to this polypeptide. Identical structures have been described in mellitin, a strongly amphipathic peptide which produces its cytotoxic effects after interaction with the cell membrane (86, 87). Nakagawa et al. (82) have suggested that the 44-residue span is not only the paradigmatic functional unit of apoA-I but a previously unappreciated structure found in proteins, a structure that exhibits properties that are distinct from a single α-helix.

Formal analysis of the positional importance of this proline residue has not been reported. Such an analysis would involve systematic displacement of this residue from its central position in the dimer and/or replacement with residues that do or do not have helix-breaking capabilities. A preliminary report of the effects of introducing proline residues into (as opposed to between) a model amphiphilic eicosapeptide has appeared (88). When helix propagation was restricted by proline substitution such that the helical segment was less than 15 residues long, stable phospholipid–peptide complexes were not formed.[17] Such a study exemplifies an approach that can be used to test the validity of structural hypotheses.

MOLECULAR EVOLUTION AND SEQUENCE PHYLOGENIES

In several ways, the mammalian apolipoproteins are biochemically unique macromolecules and this creates some difficulties for the analysis of their evolution. As noted above, selective pressures may conserve a particular pattern of generic lipophilic and hydrophilic side chains rather than a precise sequence of amino acids. (One notable exception to this phenomenon is the LDL receptor-binding domain of apoE.) This degree of interchangeability seems to go well beyond the extent of conservative substitution observed in the evolution of most proteins. Consider, for example, albumin which has many superficial similarities to the apolipoproteins in that it is a

[17] H. J. Pownall, personal communication.

TABLE 4. Comparative analysis of the physicochemical properties of apoA-I and several model oligopeptides

Property	ApoA-I (144–165)	ApoA-I (121–164)	Model A-I (1)	Model A-I (2)	ApoA-I
Peptide length	22	44	22	44	243
Air–water interface					
Collapse pressure (dyn/cm^2)	8	12	22	24	22
Area (Å^2/amino acid)	20.8	18.8	23	16.1	16.3
α-Helicity (%) in 50% trifluorethanol	40	60	61	65	73
Binding to lecithin-coated beads					
Saturating area (Å^2/amino acid)			22.0	13.8	14.8
$K_d \times 10^8$ (M)			280 (± 170)	44 (± 9)	8.5 (± 2)

Taken from Nakagawa et al. (ref. 82).

serum protein that binds lipids and evolved by multiplication and divergence of an ancestral repeat unit (89). Albumin is a "dispensable" protein and thus evolves very rapidly (90). Human and rat albumins are 74% identical in their amino acid sequences, whereas human and rat apolipoproteins A-I are only 64% homologous globally and share only 58% identity in their lipid-binding domains (6). Thus the codons that underlie apolipoprotein amphipathic sequences have evolved rapidly, with the effect that much information on their evolutionary history may have been lost. Indeed, Luo et al. (35) have recently shown that the rate of nucleotide substitution among apolipoprotein genes is considerably higher than the average rate for mammalian genes.

When sequences have diverged from a common ancestor, it is often of interest to construct a "tree" representing their evolutionary relationships. Attempts at "chemical paleogenetics" have a history extending back more than 20 years (91) and many numerical methods have been devised to infer phylogenies from molecular sequence data (reviewed in ref. 40). These methods are primarily designed to obtain *species* phylogenies from analysis of representative *orthologous* protein or nucleotide sequences. The application of these techniques to *paralogous* sequences is somewhat problematic (see below). The goal in this case is to determine the history of sequence duplications from which individual members of a multi-gene family emerged.

The most popular approach to estimating sequence phylogenies is based upon the criterion of *maximum parsimony,* which we have discussed previously. However, just as for secondary structure predictions, the most popular method is not necessarily the best method. Felsenstein (41) has criticized the statistical underpinnings of parsimony methods and discussed alternatives such as maximum likelihood and compatibility methods.[18] The latter method is particularly helpful when there is much heterogeneity in the data set (see below). The compatibility approach builds a phylogeny using only that subset of the data which comprises the largest set of mutually compatible features (e.g., aligned segments of protein or gene sequences).

No matter what approach is taken, however, the group of sequences to be analyzed must first be aligned in such a way that each element (base or residue) of one sequence corresponds to its counterparts in all of the other sequences. For sets of conserved orthologous sequences, finding the correct alignment is a relatively straightforward matter. The lengths of these sequences are usually very similar, if not identical, making the placement of any gaps (insertions or deletions) rather obvious. In contrast, there can be great difficulties in aligning a pair of divergent, paralogous sequences that differ greatly in length. The problem grows exponentially as one attempts to align more than two sequences.[19] Of course, the number of possible alignments can be limited by making simplifying assumptions (e.g., align by identities only, use a very high gap penalty, etc.). However, there is always some doubt about the biological reality of these assumptions.

For the apolipoproteins, one solution to the alignment problem has been to use intron/exon junctions as reference points for the alignment (35, 92). In other words, when computing an alignment, one considers only those matches that occur between sequences derived from corresponding exons. For example, homologies among repeated sequences in exons 3 and 4 would be ignored. This approach works fairly well in practice except in those cases where corresponding exons are highly polymorphic in length (exons 4 of the apolipoprotein genes). One is then again faced with the difficulty of making assumptions about the relative significance of substitutions versus indels. Furthermore, as previously noted, the issue of compressions, expansions, and intrasequence transpositions is not even addressed by present alignment techniques. All of these mechanisms are theoretically possible

[18] Maximum parsimony, maximum likelihood, and compatibility methods for calculating molecular sequence phylogenies have been implemented as a series of portable Pascal computer programs. This Phylogeny Interference Package (PHYLIP) may be obtained from Joseph Felsenstein, Department of Genetics SK-50, University of Washington, Seattle, WA 98195.

[19] Algorithms that perform multiple, simultaneous alignments (rather than simple pairwise comparisons) are currently under development (D. Lipman, personal communication).

in sequences that evolve via multiple rounds of nonreciprocal recombination.

We have recently identified yet another obstacle (potentially the most serious one) to phylogenetic analysis of families of paralogous sequences (8). *Gene conversion* is the nonreciprocal transfer of DNA sequences from one homologous gene to another and is a relatively common occurrence resulting in the *concerted* evolution of multi-gene families (93, 94). Gene conversion events that take place long after the evolutionary separation of duplicate, divergent genes would completely invalidate phylogeny reconstructions based on distance metrics. The acquisition of a larger amount of comparative evolutionary data (i.e., apolipoprotein sequences from nonmammalian species) would aid greatly in resolving this problem.

The seven apolipoproteins (A-I, A-II, A-IV, C-I, C-II, C-III, E) that are thought to have diverged from a common ancestral gene range in size from 83 to 396 amino acid residues (the primary translation products) and the homologies among some of these sequences are certainly less than striking (6, 7, 35). There is even some evidence for the concerted evolution of apolipoprotein genes (8). Nevertheless, attempts have been made to construct partial apolipoprotein sequence phylogenies (3, 6, 8, 35), and undoubtedly more attempts will be made now that the database of human apolipoprotein sequences is nearly complete. One other method for estimating phylogenies does not use sequence data directly but rather uses restriction endonuclease site maps instead (for discussion, see ref. 95). Indeed this approach would be best for rapidly obtaining comparative data from a number of species and this method is being applied to the analysis of human apolipoprotein genes (A. R. Templeton, personal communication).

Certain inconsistencies of apolipoprotein sequence relationships become apparent when *global* analyses are refined by optimized alignments of individual exons (8). Molecular phylogenies based upon the complete sequences (35) are not entirely consistent with exon-by-exon analyses (ref. 8 and unpublished observations). There are several possible explanations for this phenomenon. As described previously, limited gene conversion may obscure the precise evolutionary history of a group of sequences. Alternatively, individual exons may have undergone differential rates of mutation, fixation, and recombination (6, 8). It is certainly conceivable that the stringency of selective pressures might vary in different regions of the sequences (see below).

In an interesting variation on phylogeny estimation, Fitch et al. (29) have attempted to infer the evolutionary history of repeated sequences *within* the human apoA-I gene. However, only the COOH-terminal 146 residues of apoA-I were considered in this study and these residues represent only 13 of 16 or more repeated undecapeptide units (6, 35). It would be very interesting to repeat such a study using a complete data set and the apoA-IV and E sequences as well. In this way, one could test the validity of the proposed apoA-I repeat phylogeny (29). If all of the *intra*genic amplification occurred in the common ancestor of the apoA-I, A-IV, and E genes, then the inferred repeat phylogenies for the three individual, *extant* sequences should be the same or very similar.

In our analyses, we have tended to focus on repeated sequences at the amino acid level for reasons of sensitivity (as previously discussed) but also in the interest of relating protein structure to function. However, Rajavashisth et al. (24) have obtained some very intersting results by applying the seminal work of Ohno (96) to the mouse apoE cDNA sequence. Using comparison matrices and other types of computational analyses, they found that the most primitive repeat unit is an *11-nucleotide* segment of which higher order 22- and 33-nucleotide repeats are composed. This finding is of structural and evolutionary significance for the entire apolipoprotein family.

One final interesting aspect of apolipoprotein evolution is that portions of the sequences appear to have evolved at different rates (6, 9, 35). This phenomenon is well known to occur in other gene families, such as globin (97). The most highly conserved regions among orthologous apolipoproteins A-I and A-IV appear to be the amino terminal domains (6, 9). Although the significance of this finding remains to be determined, *the most slowly changing domains of proteins are those that interact with other proteins,* be they cofactors, receptors, other subunits in a multimeric enzyme, or other members of multienzyme systems (97).[20] Conversely, the most rapidly changing regions are those that interact with smaller molecules or have less specific functions. Protein cross-linking experiments might be profitably directed toward testing these possibilities.

SUMMARY AND FUTURE NEEDS

Various computational methods have already contributed substantially to our understanding of the biochemistry and molecular biology of the apolipoproteins. Computer programs have been used to detect subtle sequence homologies, to elucidate evolutionary relationships, and to generate experimentally testable predictions of protein structure and function. The pitfalls of computer-assisted analysis can be avoided by the careful selection of programs, data, and parameters, and also by the rigorous interpretation of results. Probability and statistics can be very helpful but some unresolved questions about the

[20] Using highly conserved regions as reference points is another solution to the alignment problem.

statistical properties of biological sequences limit their usefulness in certain cases.

Several techniques in widespread use may actually be suboptimal for apolipoprotein sequence analysis. For example, Chou-Fasman rules for secondary structure prediction may be inferior to other methods. Consensus sequences are of very limited usefulness and are inappropriate for most types of quantitative analysis. Compatibility methods may be better than those based on maximum parsimony for estimating sequence phylogenies. The use of several popular hydrophobicity scales requires reevaluation.

The continuing development of new algorithms is necessary to insure the robustness and reliability of computational techniques. For example, a program for multiple, simultaneous sequence alignments would be of great utility in evolutionary studies. Alignment techniques that allow for compressions, expansions, and transpositions would be quite helpful for analyzing families of sequences that arose by nonreciprocal recombination. New approaches for studying the information content of nucleotide sequences would complement experimental analysis of gene regulation. Finally, artificial intelligence techniques and knowledge-based, expert systems could make computer-assisted analysis more accessible and easier to perform.

We thank David Eisenberg, Joseph Felsenstein, Walter Fitch, Dean Goddette, David Lipman, A. D. McLachlan, William Pearson, and David Swofford for providing copies of various computer programs. We are grateful to Leonard Banaszak, Luis Glaser, Alan Templeton, and Richard Wrenn for critical readings of the manuscript. We are indebted to Diane Merritt for allowing M. S. B. time away from clinical responsibilities to complete this work. This work was supported by Grants AM 30292, HL 18577, and AM 31615 from the National Institutes of Health. Grant DMB-8520320 from the National Science Foundation provided partial support for computation. M. S. B. is supported by a Medical Scientist Training Program Grant GM 07200 and the Gerty T. Cori Predoctoral Fellowship from Sigma Chemical Company. J. I. G. is an Established Investigator of the American Heart Association.

Manuscript received 7 May 1986.

REFERENCES

1. Fitch, W. M. 1977. Phylogenies constrained by the crossover process as illustrated by human hemoglobins and a thirteen-cycle, eleven-amino-acid repeat in human apolipoprotein A-I. *Genetics.* **86:** 623-644.
2. McLachlan, A. D. 1977. Repeating helical pattern in apolipoprotein A-I. *Nature.* **267:** 465-466.
3. Barker, W. C., and M. O. Dayhoff. 1977. Evolution of lipoproteins deduced from protein sequence data. *Comp. Biochem. Physiol.* **57b:** 309-315.
4. Segrest, J. P., and R. J. Feldman. 1977. Amphipathic helices and plasma lipoproteins: a computer study. *Biopolymers.* **16:** 2053-2065.
5. Boguski, M. S., N. Elshourbagy, J. M. Taylor, and J. I. Gordon. 1984. Rat apolipoprotein A-IV contains 13 tandem repetitions of a 22-amino acid segment with amphipathic helical potential. *Proc. Natl. Acad. Sci. USA.* **81:** 5021-5025.
6. Boguski, M. S., N. Elshourbagy, J. M. Taylor, and J. I. Gordon. 1985. Comparative analysis of repeated sequences in rat apolipoproteins A-I, A-IV, and E. *Proc. Natl. Acad. Sci. USA.* **82:** 992-996.
7. Boguski, M. S., N. Elshourbagy, J. M. Taylor, and J. I. Gordon. 1986. Rat apolipoprotein A-IV: application of computational methods for studying the structure, function and evolution of a protein. *Methods Enzymol.* **128:** 753-773.
8. Boguski, M. S., E. H. Birkenmeier, N. A. Elshourbagy, J. M. Taylor, and J. I. Gordon. 1986. Evolution of the apolipoproteins: structure of the rat apoA-IV gene and its relationship to the human genes for apoA-I, C-III and E. *J. Biol. Chem.* **261:** 6398-6407.
9. Elshourbagy, N. A., D. W. Walker, M. S. Boguski, J. I. Gordon, and J. M. Taylor. 1986. The nucleotide and derived amino acid sequence of human apolipoprotein A-IV mRNA and the close linkage of its gene to the genes of apolipoproteins A-I and C-III. *J. Biol. Chem.* **261:** 1998-2002.
10. Wilbur, W. J., and. D. J. Lipman. 1983. Rapid similarity searches of nucleic acid and protein data banks. *Proc. Natl. Acad. Sci. USA.* **80:** 726-730.
11. Doolittle, R. F. 1981. Similar amino acid sequences: chance or common ancestry? *Science.* **214:** 149-159.
12. Lewin, B. 1985. Genes. 2nd edition. John Wiley & Sons, New York. 295.
13. Lipman, D. J., and W. R. Pearson. 1985. Rapid and sensitive protein similarity searches. *Science.* **227:** 1435-1441.
14. McLachlan, A. D. 1983. Analysis of gene duplication repeats in the myosin rod. *J. Mol. Biol.* **169:** 15-30.
15. Bajaj, M., and T. Blundell. 1984. Evolution and the tertiary structure of proteins. *Annu. Rev. Biophys. Bioeng.* **13:** 453-492.
16. Dayhoff, M. O. 1978. Atlas of Protein Sequence and Structure. Vol. 5, suppl. 3. National Biomedical Research Foundation, Washington, DC. 1-8, 345-375.
17. Karathanasis, S. K. 1985. Apolipoprotein multigene family: tandem organization of human apolipoprotein A-I, C-III, and A-IV genes. *Proc. Natl. Acad. Sci. USA.* **82:** 6374-6378.
18. Knuth, D. E. 1969. The Art of Computer Programming. Vol. 2, Seminumerical Algorithms. Addison-Wesley, Reading, MA. 125.
19. Sankoff, D., and J. B. Kruskal, editors. 1983. Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison. Addison-Wesley, Reading, MA. 1-44.
20. Lipman, D. J., W. J. Wilbur, T. F. Smith, and M. S. Waterman. 1984. On the statistical significance of nucleic acid similarities. *Nucleic Acids Res.* **12:** 215-226.
21. Jacobs, L. L., and D. Pilbeam. 1980. Of mice and men: fossil-based divergence dates and molecular "clocks." *J. Hum. Evol.* **9:** 551-555.
22. Karathanasis, S. K., V. I. Zannis, and J. L. Breslow. 1983. Isolation and characterization of the human apolipoprotein A-I gene. *Proc. Natl. Acad. Sci. USA.* **80:** 6147-6151.
23. Das, H. K., J. McPherson, G. A. Bruns, S. K. Karathanasis, and J. L. Breslow. 1985. Isolation, characterization, and mapping to chromosome 19 of the apolipoprotein E gene. *J. Biol. Chem.* **260:** 6240-6247.

24. Rajavashisth, T. B., J. S. Kaptein, K. L. Reue, and A. J. Lusis. 1985. Evolution of apolipoprotein E: mouse sequence and evidence for an 11-nucleotide ancestral repeat unit. *Proc. Natl. Acad. Sci. USA.* **82:** 8085–8089.
25. McLachlan, A. D. 1972. Repeating sequences and gene duplication in proteins. *J. Mol. Biol.* **72:** 417–437.
26. Dayhoff, M. O., W. C. Barker, and L. T. Hunt. 1983. Establishing homologies in protein sequences. *Methods Enzymol.* **91:** 524–545.
27. Kabsch, W., and C. Sander. 1984. On the use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations. *Proc. Natl. Acad. Sci. USA.* **81:** 1075–1078.
28. Cheung, P., and L. Chan. 1983. Nucleotide sequence of a cloned cDNA of human apolipoprotein AI. *Nucleic Acids Res.* **11:** 3703–3715.
29. Fitch, W. M., T. Smith, and J. L. Breslow. 1986. Detecting internally repeated sequences and inferring the history of a duplication. *Methods. Enzymol.* **128:** 773–788.
30. Fitch, W. M., and E. Margoliash. 1970. The usefulness of amino acid and nucleotide sequences in evolutionary studies. *Evol. Biol.* **4:** 67–109.
31. Paik, Y-K., D. J. Chang, C. A. Reardon, G. E. Davies, R. W. Mahley, and J. M. Taylor. 1985. Nucleotide sequence and structure of the human apolipoprotein E gene. *Proc. Natl. Acad. Sci. USA.* **82:** 3445–3449.
32. Innerarity, T. L., E. J. Friedlander, S. C. Rall, K. H. Weisgraber, and R. W. Mahley. 1983. The receptor-binding domain of human apolipoprotein E: binding of apolipoprotein E fragments. *J. Biol. Chem.* **258:** 12341–12347.
33. Weisgraber, K. H., T. L. Innerarity, K. J. Harder, R. W. Mahley, R. W. Milne, Y. L. Marcel, and J. T. Sparrow. 1983. The receptor-binding domain of human apolipoprotein E: monoclonal antibody inhibition of binding. *J. Biol. Chem.* **258:** 12348–12354.
34. Brutlag, D. L., J. Clayton, P. Friedland, and L. H. Kedes. 1982. SEQ: a nucleotide sequence analysis and recombination system. *Nucleic Acids Res.* **10:** 279–294.
35. Luo, D-C., W-H. Li, M. N. Moore, and L. Chan. 1986. Structure and evolution of the apolipoprotein multigene family. *J. Mol. Biol.* **187:** 325–340.
36. Colton, T. 1974. Statistics in Medicine. Little, Brown and Co., Boston. 29–30.
37. Mahley, R. W., and T. L. Innerarity. 1983. Lipoprotein receptors and cholesterol homeostasis. *Biochim. Biophys. Acta.* **737:** 197–222.
38. Yamamoto, T., C. G. Davies, M. S. Brown, W. J. Schneider, M. L. Casey, J. L. Goldstein, and D. W. Russell. 1984. The human LDL receptor: a cysteine-rich protein with multiple Alu sequences in its mRNA. *Cell.* **39:** 27–38.
39. Eck, R. V., and M. O. Dayhoff. 1966. Atlas of Protein Sequence and Structure. National Biomedical Research Foundation, Washington, DC.
40. Felsenstein, J. 1982. Numerical methods for inferring evolutionary trees. *Q. Rev. Biol.* **57:** 379–404.
41. Felsenstein, J. 1983. Inferring evolutionary trees from DNA sequences. *In* Statistical Analysis of DNA Sequence Data. B. S. Weir, editor. Marcel Dekker, Inc., New York and Basel. 133–150.
42. Kubota, Y., S. Takahashi, K. Nishikawa, and T. Ooi. 1981. Homology in protein sequences expressed by correlation coefficients. *J. Theor. Biol.* **91:** 347–361.
43. Zimmerman, J. M., N. Eliezer, and R. Simha. 1968. The characterization of amino acid sequences in proteins by statistical methods. *J. Theor. Biol.* **21:** 170–201.
44. Jones, D. D. 1975. Amino acid properties and side-chain orientations in proteins: a cross correlation approach. *J. Theor. Biol.* **50:** 167–183.
45. Sober, H. A., editor. 1970. Handbook of Biochemistry, Selected Data for Molecular Biology. 2nd edition. The Chemical Rubber Co., Cleveland, OH.
46. Chou, P. Y., and G. D. Fasman. 1978. Empirical predictions of protein conformation. *Annu. Rev. Biochem.* **47:** 251–276.
47. Oobatake, M., and T. Ooi. 1977. An analysis of nonbonded energy of proteins. *J. Theor. Biol.* **67:** 567–584.
48. Hopp, T. P., and K. R. Woods. 1981. Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci. USA.* **78:** 3824–3828.
49. Levitt, M. 1976. A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.* **104:** 59–107.
50. Wolfenden, R., L. Anderson, P. M. Cullis, and C. C. B. Southgate. 1981. Affinities of amino acid side chains for solvent water. *Biochemistry.* **20:** 849–855.
51. Chothia, C. 1976. The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.* **105:** 1–14.
52. Kyte, J., and R. F. Doolittle. 1982. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157:** 105–132.
53. Bull, H. D., and K. Breese. 1974. Surface tension of amino acid solutions: a hydropathy scale of amino acid residues. *Arch. Biochem. Biophys.* **161:** 665–670.
54. Edelstein, C., F. J. Kézdy, A. M. Scanu, and B. W. Shen. 1979. Apolipoproteins and the structural organization of plasma lipoproteins: human plasma high density lipoprotein-3. *J. Lipid Res.* **20:** 143–153.
55. Rose, G. D., A. R. Geselowitz, G. S. Lesser, R. H. Lee, and M. H. Zehfus. 1985. Hydrophobicity of amino acid residues in globular proteins. *Science.* **229:** 834–838.
56. Nozaki, Y., and C. Tanford. 1971. The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions. Establishment of a hydropathy scale. *J. Biol. Chem.* **246:** 2211–2217.
57. Steinberg, M. J. E. and J. M. Thornton. 1979. On the conformation of proteins: hydrophobic ordering of strands in β sheets. *J. Mol. Biol.* **115:** 1–17.
58. Lipman, D. J., R. W. Pastor, and B. Lee. 1986. Local sequence patterns of hydrophobicity and solvent accessibility in soluble globular proteins. *Biopolymers.* In press.
59. Steinmetz, A., and G. Utermann. 1985. Activation of lecithin:cholesterol acyltransferase by human apolipoprotein A-IV. *J. Biol. Chem.* **260:** 2258–2264.
60. Rifici, V. A., H. A. Eder, and J. B. Swaney. 1985. Isolation and lipid-binding properties of rat apolipoprotein A-IV. *Biochim. Biophys. Acta.* **834:** 205–214.
61. Pownall, H. J., Q. P. Pao, and J. B. Massey. 1985. Isolation and specificity of rat lecithin:cholesterol acyltransferase: comparison with the human enzyme using reassembled high density lipoproteins containing ether analogs of phosphatidylcholine. *Biochim. Biophys. Acta.* **833:** 456–462.
62. Fukushima, D., S. Yokoyama, D. J. Kroon, F. J. Kezdy, and E. T. Kaiser. 1980. Chain length-function correlation of amphiphilic peptides. *J. Biol. Chem.* **255:** 10651–10657.
63. Edelstein, C., and A. M. Scanu. 1980. Effect of guanidine hydrochloride on the hydrodynamic and thermodynamic properties of human apolipoprotein A-I in solution. *J. Biol. Chem.* **255:** 5747–5754.
64. Dvorin, E., W. W. Mantulin, M. F. Rohde, A. M. Gotto, H. J. Pownall, and B. C. Sherrill. 1985. Conformational properties of human and rat apolipoprotein A-IV. *J. Lipid Res.* **26:** 38–46.

65. Weinberg, R., and M. S. Spector. 1985. Structural properties and lipid binding of human apolipoprotein A-IV. *J. Biol. Chem.* **260:** 4914–4921.
66. Pace, C. N., and K. E. Vanderburg. 1979. Determining globular protein stability: guanidine hydrochloride denaturation of myoglobin. *Biochemistry.* **19:** 288–292.
67. Kabsch, W., and C. Sander. 1983. How good are predictions of protein secondary structure? *FEBS Lett.* **155:** 179–182.
68. Garnier, J., D. J. Osguthorpe, and B. Robson. 1978. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **120:** 97–120.
69. Swaney, J. B., F. Braithwaite, and H. A. Eder. 1977. Characterization of the apolipoproteins of rat plasma lipoproteins. *Biochemistry.* **16:** 271–278.
70. Eisenberg, D. 1984. Three-dimensional structure of membrane and surface proteins. *Annu. Rev. Biochem.* **53:** 595–623.
71. Eisenberg, D., R. M. Weiss, and T. C. Terwilliger. 1982. The helical hydrophobic moment: a measure of the amphiphilicity of a helix. *Nature.* **299:** 371–374.
72. Schiffer, M., and A. B. Edmundson. 1967. Use of helical wheels to represent the structure of proteins and to identify segments with helical potential. *Biophys. J.* **7:** 121–135.
73. Pownall, H. J., R. D. Knapp, A. M. Gotto, and J. B. Massey. 1983. Helical amphipathic moment: application to plasma apolipoproteins. *FEBS Lett.* **159:** 17–23.
74. Krebs, K. E., and M. C. Phillips. 1983. The helical hydrophobic moments and surface activities of serum apolipoproteins. *Biochim. Biophys. Acta.* **754:** 227–230.
75. Krebs, K. E., and M. C. Phillips. 1984. The contribution of α-helices to the surface activities of proteins. *FEBS Lett.* **175:** 263–266.
76. Ponsin, G., J. T. Sparrow, A. M. Gotto, Jr., and H. J. Pownall. 1986. In vivo interaction of synthetic acylated apopeptides with high density lipoproteins in rat. *J. Clin. Invest.* **77:** 559–567.
77. Srinivasan, R. 1976. Helical length distribution from protein crystallographic data. *Indian J. Biochem. Biophys.* **13:** 192–193.
78. Schulz, G. E., and R. H. Schirmer. 1978. Protein evolution. *In* Principles of Protein Structure. Chapter 9, Springer-Verlag, New York. 166–205.
79. Kroon, D. J., J. P. Kupferberg, E. T. Kaiser, and F. J. Kezdy. 1978. Mechanism of lipid-protein interaction in lipoproteins: a synthetic peptide-lecithin vesicle model. *J. Am. Chem. Soc.* **100:** 5975–5977.
80. Fukushima, D., J. P. Kupferberg, S. Yokoyama, D. J. Kroon, E. T. Kaiser, and F. J. Kezdy. 1979. A synthetic amphiphilic helical docosapeptide with the surface properties of plasma apolipoprotein A-I. *J. Am. Chem. Soc.* **101:** 3703–3704.
81. Yokoyama, S., D. Fukushima, J. P. Kupferberg, F. J. Kezdy, and E. T. Kaiser. 1980. The mechanism of activation of lecithin:cholesterol acyltransferase by apolipoprotein A-I and an amphiphilic peptide. *J. Biol. Chem.* **255:** 7333–7339.
82. Nakagawa, S. H., H. S. H. Lau, F. J. Kezdy, and E. T. Kaiser. 1985. The use of polymer-bound oximes for the synthesis of large peptides usable in segment condensation: synthesis of a 44 amino acid amphiphilic peptide model of apolipoprotein A-I. *J. Am. Chem. Soc.* **107:** 7087–7092.
83. Camejo, G. 1969. The structure of human high density lipoprotein: a study of the effect of phospholipase A and trypsin on its components and of the behavior of the lipid and protein moieties at the air–water interphase. *Biochim. Biophys. Acta.* **175:** 290–300.
84. Anatharamaiah, G. M., J. L. Jones, C. G. Brouillette, C. F. Schmidt, B. H. Chung, T. A. Hughes, A. S. Bhown, and J. P. Segrest. 1985. Studies of synthetic peptide analogs of the amphipathic helix—structure of complexes with dimyristoyl phosphatidylcholine. *J. Biol. Chem.* **260:** 10248–10255.
85. Chung, B. H., G. M. Anatharamaiah, C. G. Brouillette, T. Nishida, and J. P. Segrest. 1985. Studies of synthetic peptide analogs of the amphipathic helix—correlation of structure with function. *J. Biol. Chem.* **260:** 10256–10262.
86. Kaiser, E. T., and F. J. Kezdy. 1983. Secondary structures of proteins and peptides in amphiphilic environments. *Proc. Natl. Acad. Sci. USA.* **80:** 1137–1143.
87. Kaiser, E. T., and F. J. Kezdy. 1984. Amphiphilic secondary structure: design of peptide hormones. *Science.* **223:** 249–255.
88. Ponsin, G., L. Hester, A. M. Gotto, Jr., H. J. Pownall, and J. T. Sparrow. 1986. Lipid–peptide association and activation of lecithin:cholesterol acyltransferase: effect of α-helicity. *J. Biol. Chem.* **261:** 9202–9205.
89. Alexander, F., P. R. Young, and S. M. Tilghman. 1984. Evolution of the albumin:α-fetoprotein ancestral gene from the amplification of a 27 nucleotide sequence. *J. Mol. Biol.* **173:** 159–176.
90. Wilson, A. C., S. S. Carlson, and T. J. White. 1977. Biochemical evolution. *Annu. Rev. Biochem.* **46:** 573–639.
91. Pauling, L., and E. Zuckerkandl. 1963. Chemical paleogenetics: molecular "restoration studies" of extinct forms of life. *Acta Chem. Scand.* **17:** S9–S16.
92. Shelley, C. S., C. R. Sharpe, F. E. Baralle, and C. C. Shoulders. 1985. Comparison of the human apolipoprotein genes: apoA-II presents a unique functional intron–exon junction. *J. Mol. Biol.* **186:** 43–51.
93. Arnheim, N. 1983. Concerted evolution of multigene families. *In* Evolution of Genes and Proteins. M. Nei and R. K. Koehn, editors. Sinauer Associates, Sunderland, MA. 38–61.
94. Fink, G. R., and T. D. Petes. 1984. Gene conversion in the absence of reciprocal recombination. *Nature.* **310:** 728–729.
95. Templeton, A. R. 1983. Convergent evolution and nonparametric inferences from restriction data and DNA sequences. *In* Statistical Analysis of DNA Sequence Data. B. S. Weir, editor. Marcel Dekker, New York. 151–179.
96. Ohno, S. 1984. Repeats of base oligomers as the primordial coding sequences of the primeval earth and their vestiges in modern genes. *J. Mol. Evol.* **20:** 313–321.
97. Dickerson, R. E., and I. Geis. 1983. Hemoglobin: Structure, Function, Evolution and Pathology. Benjamin/Cummings, Menlo Park, CA. 66–111.