Nomenclature Committee of the International Union of Biochemistry (NC-IUB)

Nomenclature for Incompletely Specified Bases in Nucleic Acid Sequences

Recommendations 1984

CONTENTS

1. Introduction
2. Applications of a Standard Nomenclature
   2.1. Recognition Sequences in DNA for Restriction Enzymes
   2.2. Recognition Sequences in DNA for Other Enzymes
   2.3. Recognition Sequences in DNA for Enzymes Involved in Translation
   2.4. Codon Degeneracy
   2.5. Construction of Ancestral Sequences by Parsimony Procedures
   2.6. Other Uses
3. Allocation of Symbols
   3.1. Guanine, Adenine, Thymine, Cytosine: G, A, T, C
   3.2. Purine (adenine or guanine): R
   3.3. Pyrimidine (thymine or cytosine): Y
   3.4. Adenine or Thymine: W
   3.5. Guanine or Cytosine: S
   3.6. Adenine or Cytosine: M
   3.7. Guanine or Thymine: K
   3.8. Adenine or Thymine or Cytosine: H
   3.9. Guanine or Cytosine or Thymine: B
   3.10. Guanine or Adenine or Cytosine: V
   3.11. Guanine or Adenine or Thymine: D
   3.12. Guanine or Adenine or Thymine or Cytosine: N
4. Other Accessory Symbols
5. Discussion
   5.1. Interpretation of Primary Sequencing Data
   5.2. Modified Nucleotides
References

1. Introduction

With the introduction of methods of rapid nucleic acid sequence determination, synthesis of mixed oligonucleotide probes and computer-assisted analysis of nucleic acid sequences, the need for a standard nomenclature for incompletely specified bases in nucleic acid sequences has become apparent. This document provides recommendations for such a nomenclature.

1. Document of the Nomenclature Committee of the International Union of Biochemistry (NC-IUB) whose members are H. B. F. Dixon (Chairman), H. Bielka, C. R. Cantor, C. Liébecq (representing the IUB Committee of Editors of Biochemical Journals), N. Sharon, S. F. Velick and J. F. G. Vliegenthart. NC-IUB thanks the panel, whose members were F. Blattner (USA), N. L. Brown (UK), D. L. Brutlag (USA), W. M. Fitch (USA), W. Goad (USA), R. Grantham (France), G. Hamm (Federal Republic of Germany), L. H. Kedes (USA), R. Lathe (France, convener), D. W. Mount (USA), J. Schroeder (USA), R. Staden (UK), P. A. Stockwell (New Zealand), for drafting these recommendations. IUB also thanks other members of the IUPAC-IUB Joint Commission on Biochemical Nomenclature (K. L. Loening, G. P. Moss and J. Reedijk) for consultation. Comments may be sent to any member of the NC-IUB, or to its secretary, A. Cornish-Bowden, Department of Biochemistry, University of Birmingham, P.O. Box 363, Birmingham B15 2TT, England, or to the convener of the panel, R. Lathe, whose present address is: A.F.R.C. Animal Breeding Research Organisation, West Mains Road, Edinburgh EH9 3JQ, Scotland.

Copyright 1985 by the Federation of European Biological Societies (published in Eur. J. Biochem. 150:1-5). All rights reserved. Reprinted with permission.
acid sequences, the use of a single symbol to designate a variety of possible nucleotides at a single position has become widespread over the last few years. Whereas the use of, for example, the symbols R and Y to designate purine (A or G) and pyrimidine (C or T) ribonucleotides respectively [1] is generally accepted, no agreed symbols exist for the other possible combinations. Indeed, a plethora of diverse systems has proliferated in the last few years [2-11]. It is striking that, in one extreme case, the combination (C or G) has been represented by at least five different symbols [2-4, 8, 11]. A standardized set of symbols is thus required to prevent confusion.

The symbols are intended to be applicable to both deoxyribonucleic and ribonucleic acids. Thus it is important to note from the outset that the recommended symbols will not discriminate between DNA and RNA, and the symbol T will be employed at all positions where U might appear in the RNA. Similarly, no distinction will be made in the symbols between base, nucleoside and nucleotide. Sequences may be assumed to have a deoxyribose backbone (DNA) unless specified otherwise. These changes from earlier recommendations [1] reflect great advances in techniques for sequencing DNA so that RNA sequences are now commonly deduced from the corresponding DNA sequences. Since the standard representation of a DNA sequence may be converted to the corresponding RNA sequence by the simple expedient of substituting T by U, it is not envisaged that data banks based on computer storage facilities will inevitably contain entries for both DNA and its RNA equivalent. Authors should always, however, make it clear which strand of DNA or RNA a given sequence refers to, and in circumstances where confusion between DNA and RNA is likely the sequence may be prefixed with the lower-case letter d or r, as in the previous recommendations [1].

As the present recommendations present unique alphabetic symbols for each nucleotide combination, the use of upper- and lower-case letters as equivalent does not lead to confusion. However, such use may cause confusion between r (ribo-) and R (purine), and care must be taken in those rare cases where the various symbols are used in combination. In general, it should be emphasised (i) that upper-case symbols are advocated, and (ii) that the present recommendations are not intended to prejudice any possible future use of contrasting upper- and lower-case letters for specific purposes.

It was previously [1] recommended that hyphens should be used to represent phosphodiester linkages in known nucleotide sequences. As there is now little danger of confusion between codon triplets and nucleotide sequences this recommendation is no longer considered necessary. Hyphens may therefore be omitted from sequences, and are omitted from all sequences in this document. In addition it may be assumed that all sequences are presented 5' to 3' unless otherwise specified, although specific mention of this fact is not discouraged.

Although several diverse systems of symbols for incompletely specified bases already exist in the literature, this presentation makes no systematic review. Details of the previous recommendations may be found in [1], and of systems that have been used in the literature in [2-11].

2. Applications of a Standard Nomenclature

2.1. Recognition Sequences in DNA for Restriction Enzymes

Most restriction enzymes and their corresponding methylases recognise simple unique nucleotide sequences in DNA. For example, EcoRI and BamHI recognise the sequences 5'GAATTC-3' and 5'GGATCC-3' respectively. Nevertheless, a growing
class of enzymes includes those that recognise series of derivative sequences, where two or more bases may be present at a particular position in the recognition sequence (for a complete listing see [12]). For instance, the enzyme *AvaI* recognises four different sequences 5'-CCCGGG-3', 5'-CCCGAG-3', 5'-CTCGGG-3' and 5'-CTCGAG-3'. The recognition sequence for *AvaI* may thus be represented as 5'-CYCGRC-3', where Y represents a pyrimidine and R represents a purine, as recommended previously [1]. However, several newer enzymes recognise combinations that are not covered by the existing symbols. For instance, *AccI* recognises the sequence 5'-GT(A or C)(G or T)AC-3' [13]. *SduI* recognises the sequence 5'-G(A or G or T)GC(A or C or T)C-3' [14]. The present symbols are intended to cover these possibilities.

2.2. Recognition Sequences in DNA for Other Enzymes

Restriction enzymes are highly specific for particular nucleic acid sequences. For many other enzymes the specificity is rather more lax, however, and the symbols are intended to meet in part the need for presenting a schematic summary of the sequence features. For instance, sequences recognised by the RNA polymerase of *Escherichia coli* may be presented as the juxtaposition of two sequences 5'-AA(A or T)NTNNN(C or G)TTGACA-3' and 5'-(T or G)NNTATAAT-3' separated by 13 to 16 nucleotides (adapted from [15, 16]), where N represents any nucleotide. A similar treatment may be applied to the recognition sequences for other DNA binding proteins such as repressor molecules.

2.3. Recognition Sequences in RNA for Enzymes Involved in Translation

RNA sequences are, as mentioned above, most conveniently represented as their DNA counterparts. Thus the basic elements of a translation initiation site in *Escherichia coli* may be represented by 5'-(G or A)(G or A)GGG-NNNNAN(C or T)ATGNN(A or T)NNNNN(C, T or G) (adapted from [17]). Similarly, translation initiation sites in eukaryotic mRNAs tend to conform to the sequence 5'-ANNATG(G or A)-3' [18].

2.4. Codon Degeneracy

Although there are 64 possible triplet codons, there are only 20 different amino acids coded by them. Thus most amino acids are inserted into a growing polypeptide chain in response to two or more different triplets in the mRNA ([19] for a general review). For example, proline is coded by 5'-CCN-3' and alanine by 5'-GCN-3'. In other cases the pattern may be more complex, such as for isoleucine, which is coded by 5'-AT(T, C or A)-3'. Note that certain amino acids (e.g. serine) may be coded by two distinct groups of triplets [here 5'-TCN-3' and 5'-AG(T or C)-3'], which cannot be adequately represented as 5'-(T or A)(C or G)N-3' (see Table 4). It is to be noted that synthetic oligonucleotide probes for detecting protein-coding sequences often involve the preparation of 'mixed probes'. Here a mixture of two (or more) nucleotides is incorporated at a single position in the oligonucleotide to take account of the redundancy of the genetic code (for instance [20]). It is anticipated that a single-letter code might be used to designate such mixtures.

2.5. Construction of Ancestral Sequences by Parsimony Procedures [21]

Where two descendants differ in nucleic acid sequence at a particular position (for instance A in one and G in the other), the putative ancestral sequence can be represented [10] using a single-letter code, in this case R.
2.6. Other Uses

The symbols are intended to be useful for all purposes in which the exact identity of a nucleotide may vary. Thus uncertainties encountered with primary nucleic acid sequence data may, in some cases, be represented using standard symbols.

3. Allocation of Symbols

In the choice of symbols the following considerations have been taken into account: (i) conformity to previous IUPAC-IUB nomenclature [1]; (ii) logical derivation; (iii) ease of memorisation; (iv) availability of symbols on a standard typewriter keyboard; (v) historical precedence.

3.1. Guanine, Adenine, Thymine, Cytosine: G, A, T, C

These one-letter symbols have previously been established [1] and are generally used. There is, however, a problem of discriminating between the upper-case letters G and C on poorly copied sequences. Nevertheless, the use of alternative symbols for G (such as a barred-G, G) is not recommended. Discrimination between the lower-case letters is much clearer. Note that T and U may, in general, be considered as being synonyms, though care should be taken to avoid ambiguity in circumstances where it is likely, e.g. in discussing artificial hybrids of DNA and RNA and in cases where specific distinction between T and U is advisable.

3.2. Purine (adenine or guanine): R

R is the symbol previously recommended [1].

3.3. Pyrimidine (thymine or cytosine): Y

Y is the symbol previously recommended [1].

3.4. Adenine or Thymine: W

Although several diverse symbols have been used for this pair, (and for the reciprocal pair G + C), only two symbols have a rational basis, L and W: L derives from DNA density (light: G + C—heavy—would thus be H); W derives from the strength of the hydrogen bonding interaction between the base pairs (weak for A + T: G + C—strong—would thus be S). However, the system recommended for the three-base series (not-A = B, etc., see below, section 3.8) rules out H as this would be not-G. W is thus recommended.

3.5. Guanine or Cytosine: S

The choice of this symbol is discussed above in section 3.4.

3.6. Adenine or Cytosine: M

There are few common features between A and C. The presence of an NH₂ group in similar positions on both bases (fig. 1) makes possible a logically derived symbol. A and N being ruled out, M (from aMino) is recommended.

3.7. Guanine or Thymine: K

By analogy with A and C (section 3.6), both G and T have Keto groups in similar positions (fig. 1).
FIG. 1.—Origin of the symbols M and K. The four bases are drawn so as to show the relationship between adenine and cytosine on the one hand, which both have amino groups at the ring position most distant from the point of attachment to the sugar, and between guanine and thymine on the other, which both have keto groups at the corresponding position. The ring atoms are numbered as recommended [24–26], although for the present purpose this has the disadvantage of giving discordant numbers to the corresponding positions.

3.8. Adenine or Thymine or Cytosine: H

Not-G is the most simple means of memorising this combination and symbols logically related to G were examined. F and H would both be suitable, as the letters before and after G in the alphabet, but A would have no equivalent to F. The use of H has historical precedence [2].

3.9. Guanine or Cytosine or Thymine: B

Not-A as above (section 3.8).

3.10. Guanine or Adenine or Cytosine: V

Not-T by analogy with not-G (section 3.8) would be U but this is ruled out to eliminate confusion with uracil. V is the next logical choice. Note that T and U may in some cases be considered to be synonyms.

3.11. Guanine or Adenine or Thymine: D

Not-C as above (section 3.8).

3.12. Guanine or Adenine or Thymine or Cytosine: N

This symbol is suggested by the sound of the word ‘aNy’. The use of X to represent an unknown base is acknowledged, but this is not recommended as the symbol refers to xanthine [1]. Occasionally it may be desirable to distinguish between unspecified (N) and unknown (X), but if X is used for this purpose it should be explicitly defined.
4. Other Accessory Symbols

There are a number of instances in which additional symbols may be required for routine work. Although this section provides a number of suggestions, these do not form part of the present recommendations.

First, we consider the uncertainty as to whether a base exists at a certain position or not. A symbol denoting 'G or A or T or C or no nucleotide', for example '?' or '+', might be used to define regions of uncertainty of limited variable size in a recognition sequence (see for instance [22]). Alternatively, one of these symbols might be used as a modifier to denote uncertainty: '?A' might, for instance, denote 'A or no nucleotide at this position'. Second, the unambiguous absence of a nucleotide introduced into a sequence for alignment or comparison purposes alone could be represented by '·'; though a simple space has much to recommend itself. Third, a specified number of unknown nucleotides might be represented by a symbol such as '=' in conjunction with numerals, so that, for example, '='300=' might denote the presence of 300 unknown nucleotides. Fourth, the symbol 'N' (unknown or unspecified) may be replaced by the hyphen '-' in circumstances where rapid visual discrimination between 'known' (essential) and 'unknown' (non-essential) sequences is desirable. The value of this may be judged by comparing 'NNNNNCNNGNTNN' with '-----C--G-T- -', for example. Note that the use of the lower-case letter n may avoid the necessity for an additional symbol, as in 'nnnnnCnnGnnTnn'.

In addition, the use of the oblique or slash '/' may present advantages in the definition of the precise cleavage sites of restriction endonucleases. For instance, the cleavage specificity of the common enzyme EcoRI might be represented by G/AATTC, where cleavage occurs in both strands of the self-symmetrical sequence between the G and A residues.

It is emphasised that the symbols appearing in this section do not form an integral part of the recommendations and must therefore be defined explicitly in the context in which they are used.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
<th>Origin of Designation</th>
</tr>
</thead>
<tbody>
<tr>
<td>G ......</td>
<td>G</td>
<td>Guanine</td>
</tr>
<tr>
<td>A ......</td>
<td>A</td>
<td>Adenine</td>
</tr>
<tr>
<td>T ......</td>
<td>T</td>
<td>Thymine</td>
</tr>
<tr>
<td>C ......</td>
<td>C</td>
<td>Cytosine</td>
</tr>
<tr>
<td>R ......</td>
<td>G or A</td>
<td>puRine</td>
</tr>
<tr>
<td>Y ......</td>
<td>T or C</td>
<td>pYrimidine</td>
</tr>
<tr>
<td>M ......</td>
<td>A or C</td>
<td>aMino</td>
</tr>
<tr>
<td>K ......</td>
<td>G or T</td>
<td>Keto</td>
</tr>
<tr>
<td>S ......</td>
<td>G or C</td>
<td>Strong interaction (3 H bonds)</td>
</tr>
<tr>
<td>W ......</td>
<td>A or T</td>
<td>Weak interaction (2 H bonds)</td>
</tr>
<tr>
<td>H ......</td>
<td>A or C or T</td>
<td>not-G, H follows G in the alphabet</td>
</tr>
<tr>
<td>B ......</td>
<td>G or T or C</td>
<td>not-A, B follows A</td>
</tr>
<tr>
<td>V ......</td>
<td>G or C or A</td>
<td>not-T (not-U), V follows U</td>
</tr>
<tr>
<td>D ......</td>
<td>G or A or T</td>
<td>not-C, D follows C</td>
</tr>
<tr>
<td>N ......</td>
<td>G or A or T or C</td>
<td>aNy</td>
</tr>
</tbody>
</table>
Table 2
Definition of Complementary Symbols

<table>
<thead>
<tr>
<th>Symbol</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>G</th>
<th>H</th>
<th>K</th>
<th>M</th>
<th>S</th>
<th>T</th>
<th>V</th>
<th>W</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complement</td>
<td>T</td>
<td>V</td>
<td>G</td>
<td>H</td>
<td>C</td>
<td>D</td>
<td>M</td>
<td>K</td>
<td>S*</td>
<td>A</td>
<td>R</td>
<td>W*</td>
<td>N*</td>
</tr>
</tbody>
</table>

* In certain cases the symbol and its complement are identical.

5. Discussion

The present nomenclature, summarised in table 1, has been formulated to deal with incomplete specification of bases in nucleic acid sequences. In cases where two or more bases are permitted at a particular position the nomenclature permits the allocation of a single-letter symbol. The nomenclature may also be applied where uncertainty exists as to extent and/or identity. For double-stranded nucleic acids table 2 permits the allocation of symbols to the complementary strand. Examples are given whereby the nomenclature is applied to sequences recognised by certain type II restriction endonucleases (table 3) and to uncertainties in deriving a nucleic acid sequence from the corresponding amino acid sequence (table 4).

Two applications fall outside the scope of the nomenclature and these are considered separately below.

Table 3
Single-Letter Code Recognition Sequences for Several Type II Restriction Endonucleases
Recognising Multiple Sequences

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>Recognition Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accl</td>
<td>G T/ M K A C</td>
</tr>
<tr>
<td>Accl (HglDI)</td>
<td>G R/ C G Y C</td>
</tr>
<tr>
<td>AlIII</td>
<td>A/ C R Y G T</td>
</tr>
<tr>
<td>Aval</td>
<td>C/ Y C G R G</td>
</tr>
<tr>
<td>AvalII</td>
<td>G/ G W C C</td>
</tr>
<tr>
<td>Banl</td>
<td>G/ G Y R C C</td>
</tr>
<tr>
<td>BanII (HgiIII)</td>
<td>G R G C Y/ C</td>
</tr>
<tr>
<td>CfrI (GdiII)</td>
<td>Y/ G G C C R</td>
</tr>
<tr>
<td>EcoRII</td>
<td>/ C C W G G</td>
</tr>
<tr>
<td>Hael</td>
<td>W G G/ C C W</td>
</tr>
<tr>
<td>Haell</td>
<td>R G C G C/ Y</td>
</tr>
<tr>
<td>HgiAI</td>
<td>G W G C W/ C</td>
</tr>
<tr>
<td>HindII (HincII)</td>
<td>G T Y/ R A C</td>
</tr>
<tr>
<td>Ncil (CauII)</td>
<td>C C/ S G G</td>
</tr>
<tr>
<td>NspI (Nsp-7524-I)</td>
<td>R C A T G/ Y</td>
</tr>
<tr>
<td>NspBII</td>
<td>C M G/ C K G</td>
</tr>
<tr>
<td>Sdul (Nsp-7524-II)</td>
<td>G D G C H/ C</td>
</tr>
<tr>
<td>XhoII</td>
<td>R/ G A T C Y</td>
</tr>
</tbody>
</table>

**NOTE.**—Recognition sequences are presented 5' to 3': the exact position of cleavage is indicated here by /.

**SOURCE.**—[12].
Table 4

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Single-Letter Code</th>
<th>Triplet (5'-3')</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glycine</td>
<td>G</td>
<td>GGN</td>
</tr>
<tr>
<td>Alanine</td>
<td>A</td>
<td>GCN</td>
</tr>
<tr>
<td>Valine</td>
<td>V</td>
<td>GTN</td>
</tr>
<tr>
<td>Leucine</td>
<td>L</td>
<td>YTN (CTN and TTR)*</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>I</td>
<td>ATH</td>
</tr>
<tr>
<td>Proline</td>
<td>P</td>
<td>CCN</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>F</td>
<td>TTY</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>Y</td>
<td>TAY</td>
</tr>
<tr>
<td>Cysteine</td>
<td>C</td>
<td>TGY</td>
</tr>
<tr>
<td>Methionine</td>
<td>M</td>
<td>ATG</td>
</tr>
<tr>
<td>Histidine</td>
<td>H</td>
<td>CAY</td>
</tr>
<tr>
<td>Lysine</td>
<td>K</td>
<td>AAR</td>
</tr>
<tr>
<td>Arginine</td>
<td>R</td>
<td>MGN (CGN and AGR)*</td>
</tr>
<tr>
<td>Tryptophan</td>
<td>W</td>
<td>TGG</td>
</tr>
<tr>
<td>Serine</td>
<td>S</td>
<td>WSN (TCN and AGY)*</td>
</tr>
<tr>
<td>Threonine</td>
<td>T</td>
<td>ACN</td>
</tr>
<tr>
<td>Aspartic acid</td>
<td>D</td>
<td>GAY</td>
</tr>
<tr>
<td>Glutamic acid</td>
<td>E</td>
<td>GAR</td>
</tr>
<tr>
<td>Asparagine</td>
<td>N</td>
<td>AAY</td>
</tr>
<tr>
<td>Glutamine</td>
<td>Q</td>
<td>CAR</td>
</tr>
<tr>
<td>Aspartic acid or asparagine</td>
<td>B</td>
<td>RAY</td>
</tr>
<tr>
<td>Glutamic acid or glutamine</td>
<td>Z</td>
<td>SAR</td>
</tr>
<tr>
<td>Terminator</td>
<td></td>
<td>TTR (TAR and TGA)*</td>
</tr>
<tr>
<td>Unknown</td>
<td>X</td>
<td>NNN</td>
</tr>
</tbody>
</table>

* The sequence of amino acids is uniquely specified by the nucleotide sequence. Similarly, it is possible to convert an amino-acid sequence to a linear order of base uncertainties, but this raises problems with the codons for leucine, arginine, serine and termination. With leucine, for example, the coding triplets are precisely specified by GIN and TTR, but combining these gives YTN, which also includes two phenylalanine codons, TTT and TTC. Thus information may be lost when an amino-acid sequence is converted into a single sequence of base-uncertainty symbols. To avoid ambiguity, therefore, it is important to make it clear whenever the triplet YTN, for example, occurs in a sequence deduced from the occurrence of a leucine residue in the corresponding amino acid sequence that it does not include TTT or TTC as possibilities, etc. To emphasise this, it may be helpful to print such triplets in italics.

5.1. Interpretation of Primary Sequencing Data

In certain cases the nomenclature permits uncertainty encountered in nucleic acid sequencing to be represented using a single letter code. At least two applications, namely 'either X or XY' and 'probably X', are not adequately handled. Specialised nomenclatures have been developed specifically for such purposes (for instance [3]).

5.2. Modified Nucleotides

In a number of organisms DNA and RNA are modified at certain positions. For instance, the DNA of *Escherichia coli* is usually methylated on N-6 of the adenine residue in the sequence 5'-GATC-3' [23]. The present nomenclature does not allocate any specific symbol to these modified nucleotides for the following reasons. (1) The presence or absence of a given modification depends upon the location of the DNA.
Sequences modified in one organism may not be modified in another. (2) Modification is usually statistical, in that only a proportion of possible sites for modification may actually be utilised in vivo. Modification of a nucleotide or base in a given polynucleotide is not a function of the sequence per se. Although it is recognised that stable RNA species (tRNA and rRNA) often carry a constant pattern of post-translational modification, the present nomenclature is not intended to overlap with or supplant existing systems. It would probably be impossible to devise a simple and logical system that avoided all conflict with previous usage. One should therefore recognise that such conflict is possible and take steps to prevent it from generating confusion, for example in relation to the symbols B, D and S, which have been recommended for 5-bromouridine, 5,6-dihydrouridine and thiouridine respectively [1], or W, which is sometimes used for wyosine [27].

REFERENCES


10. FITCH, W. M. 1983. unpublished manuscript.


An Analysis of Replacement and Synonymous Changes in the Rodent L1 Repeat Family

Stephen C. Hardies,2 Sandra L. Martin,3 Charles F. Voliva,4 Clyde A. Hutchison III, and Marshall H. Edgell

Department of Microbiology and Immunology, University of North Carolina

L1 is a family of long interspersed repetitive sequences in mammals that includes the BamHI family in rodents and the KpnI family in primates. Previous studies have shown that L1 repeats contain a long open reading frame and that the family evolves in concert. Working with 32 rodent elements for which DNA sequence is available, we used the distribution of replacement and synonymous changes to determine which L1 lineages had been expressing their reading frame. The evidence obtained is consistent with there having been a small number of L1 genes that have been expressing a functional protein. Much of the concerted evolution in L1 is accounted for by the tendency of these functioning L1 genes to continually create nonfunctional pseudogenes by reinsertion into the genome of sequences derived from their transcripts. The gain of new pseudogenes is balanced by the loss of old pseudogenes with a half-life of 2 Myr. Therefore, most of the observed L1 repeats are at a dead end with respect to either the expression of the L1 protein or the potential to elaborate further copies of themselves. However, the turnover of L1 pseudogenes is sufficient to constitute a vast flux of sequences into and then out of the flanking regions of all cellular genes. If the presence of flanking L1 pseudogenes affects the expression of other genes in an even a subtle fashion, this process should represent a major source of genetic variation. A second level of concerted evolution occurs within the functional L1 sequences in a pattern that did not meet our expectations for selfish DNA. Also, in spite of the marked suppression of replacement relative to synonymous changes in functioning L1 genes, they evolve at an overall rate accelerated to the level of their own pseudogenes.

Introduction

L1 or LINES-1 is a family of mammalian long interspersed repetitive sequences that includes the KpnI family of humans and the BamHI family in mice (Singer 1982a, 1982b; Singer et al. 1983; Singer and Skowronski 1985; Burton et al. 1986). L1 is characterized as a 6-7-kb-long sequence, one end of which is missing from most copies (Fanning 1983; Voliva et al. 1983; for review, see Rogers 1985; Singer and Skowronski 1985). Many elements have poly(A) tracts at one end and are surrounded by short direct repeats. These features have been taken to suggest that most L1 sequences

1. Key words: concerted evolution, molecular drive, L1 repeat family, replacement and synonymous changes.

Address for correspondence and reprints: Dr. Stephen C. Hardies, Department of Biochemistry, University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Drive, San Antonio, Texas 78284.

2. Current address: Department of Biochemistry, University of Texas Health Science Center at San Antonio, San Antonio, Texas 78284.

3. Current address: Synergen, 1885 33d Street, Boulder, Colorado 80301.

4. Current address: Department of Genetics, University of California, Berkeley, California 94720.

© 1986 by The University of Chicago. All rights reserved.
0737-4038/86/0302-3205$02.00
are insertions derived from a reverse-transcribed L1 transcript (for review, see Rogers 1985).

A protein-encoding function for the L1 family has been inferred in several ways. Sequenced L1 members often have open segments much greater than the 20-codon average expected for a random sequence (Manuelidis 1982; Thayer and Singer 1983; Martin et al. 1984; Potter 1984). Jagadeeswaran et al. (1982) noted that an individual KpnI sequence had a triplet periodicity like that of coding sequences. However individual L1 members usually carry defects in the consensus reading frame. Martin et al. (1984) compared a 300-bp region between mouse and primate and showed a conserved frame with a higher proportion of changes at synonymous (S) positions than at replacement (R) positions. Similar results are found over at least a 1400-bp region (Burton et al. 1986). This marked perturbation of the ratio of R to S changes (R/S) is taken as strong confirmation that the long open reading frames have acted as the template for a protein, at least historically. So, although some L1 members carry defects, their ancestors descended under selective pressure to express a protein. This paper extends that result to recently diverged rodent L1 sequences, showing that functional L1 sequences have persisted to within at least the past few million years, if not until the present.

The L1 families undergo concerted evolution (Dover 1982; Martin et al. 1985). The analysis of Martin et al. (1985) was based on 32 300-bp sequences from three different species of mice. A tree was derived representing the history of sequence exchanges leading to these elements. In this paper we have further examined the concerted spread of changes into the population of repeats by localizing each change on the tree. Each change was cataloged according to whether it was an R or S change. This allowed us to determine which parts of the concerted evolution process were subject to conservative selection for the reading frame. The results of that analysis support the model that most L1 sequences are processed pseudogenes derived from the mRNA of functional L1 genes. We present an observation on the mechanism of the concerted evolution among the L1 pseudogenes that distinguishes between continuous dispersal of new pseudogenes and gene conversion. We show that there are multiple functional L1 genes and that they also undergo concerted evolution among themselves. The mechanism of concerted evolution among functional L1 genes appears to differ from that affecting L1 pseudogenes.

The theory of molecular drive (Dover 1982; Ohta and Dover 1983, 1984) is based partly on evaluation of concerted evolution of the rodent L1 families, under the name MIF-1. One conclusion of this work was that concerted evolution is accelerated in the L1 family because some members are more efficient donors in sequence exchanges than others. By a donor we mean any sequence that makes a second copy of itself. During gene conversion, the sequence that ends up in two copies will be considered the donor. During any duplicative process, regardless of the mechanism, the original template will be considered the donor.

The accelerating effect of preferred donors was illustrated by the elaboration of a model for concerted evolution by gene conversion in which some of the sequence elements had a conversion bias (Dover 1982). As noted by Dover, other mechanisms of concerted evolution will be similarly accelerated if a bias is built into the mechanism. For example, in the processed-pseudogene model for concerted evolution of L1, functional L1 genes would be preferred over pseudogenes to be the template for the creation of new pseudogenes. This is because functional genes would presumably be more likely to be transcribed.
The results of this study support this mechanism as a major contributor to the acceleration of concerted evolution of L1 pseudogenes. That is, we find that L1 pseudogenes are preferentially created from functional genes. Conversely, we found no evidence for the further elaboration of any pseudogenes into multiple copies.

We also report evidence for preferred sequence donors during concerted evolution of the functional L1 genes among themselves. Since the mechanism of concerted evolution among the functional L1 genes is not established, the reason why some are preferred sequence donors cannot presently be determined. We coin the term “molecular drivers” to refer to these special sequences. Base changes that arise in the molecular drivers will preferentially spread through the repeat family as described by the theory of molecular drive (Dover 1982).

Methods

The sequences and the tree relating them were taken from Martin et al. (1985). A donor is defined as being any ancestral sequence at a node. A lineage of sequences that includes multiple donors is defined as representing a molecular driver.

Each base change was positioned on the tree by the method of Fitch (1977) with assistance from the program MPN (Czelusniak et al. 1982). Because of the light mutation density in these data, there is little ambiguity arising from parallel and back mutations. There are 30 parallel or back mutations detected out of a total of 174 changes mapped on the tree. Of these, 15 are such well-separated parallel mutations that their resolution is unambiguous (e.g., a variant occurring in only one sequence from Mus domesticus and one sequence from M. platythrix). The few base changes that were not discretely localized were distributed evenly over their possible positions according to the method of Fitch (1971). The same general results were obtained when these changes were excluded from the analysis. A program named RSBRANCH was written in FORTRAN and run on an IBM mainframe to perform the above bookkeeping and to classify changes as R or S.

The absolute time scale was calibrated by assuming that the M. domesticus/M. platythrix split occurred 11 Myr ago (V. Sarich, personal communication) and that the node separating the platythrix repeats from those of the other two rodents represents that split (Martin et al. 1985). Nodes were timed by their relative position on the driver lineage. The results can be read from figure 2. (Fig. 1 is not to scale.) Figure 2 was scaled as follows: Branch lengths for the thickened lineages were taken from figure 1. These lineages include the molecular drivers plus the two direct descendants from each penultimate driver node. The latter are not formally part of the driver lineages; however, they are included because they provide the only way to estimate the height of the lowest driver nodes. An average-node-height tree was then constructed from these selected lineages; that is, each node was assigned a height equal to the sum of (1) one-half of the sum of the heights of its two descendants and (2) the length of the branches leading to these descendants. In this way, the position of each node representing the joining of two drivers was fixed and scaled on figure 2. The nondriver branches were then joined as in the following example: P21 joins its driver 2 changes above a node fixed by the above procedure and 3.9 changes below another fixed node (from fig. 1). It was scaled between these two nodes on figure 2 according to these proportions.

The topology-independent categorization (see Results) of changes as private or nonprivate was done with a program (TREELESS) written in FORTRAN for execution on an IBM mainframe. A private change occurs when a given nucleotide appears in
only a single sequence at a particular position. A nonprivate change occurs at a given position if there are multiple nucleotides each shared by more than one sequence. Mutations to private bases were classified as being of the R or S type on the assumption that no other changes had occurred in the codon. If more than one nonprivate base appeared at the same position as a private base and if the nonprivate bases did not all give the same R or S change, the change was distributed among the two types in the proportion of the nonprivate bases that gave rise to those two types. Changes from one nonprivate base to another were proportioned between R and S types according to the average of all possible combinations of such changes per codon weighted by the relative frequency of the respective bases.

The R/S value expected for a pseudogene was calculated with the aid of a program (RS) written in FORTRAN for execution on an IBM mainframe. This program divided the potential changes to an L1 sequence into R and S sites and subdivided those classes into transitions and transversions. An unselected value of R/S = 3.2 was then found after correction for the fourfold excess of transitions over transversions observed in these sequences.

Martin et al. (1984) calculated an S/R ratio for a mouse/primate comparison that indicated selection on the reading frame. We recalculated this value in a form that would be directly comparable to the observed R/S values in this paper. R and S differences were subdivided into transitions and transversions and individually corrected for parallel and back mutations as described by Brown et al. (1982). In order to handle codons with multiple changes we used the procedure of Miyata and Yasunaga (1981) for averaging over the possible intermediate codons. Then, instead of using the averaging procedure of Brown et al. we back-calculated the number of changes in each category by multiplying the corrected divergence by the number of sites. Summing transitions and transversions we found R and S values that estimate the actual number of changes that occurred before the obscuring effect of parallel and back mutation. An SE that includes a contribution from the multiple-hit correction was calculated according to the method of Tajima and Nei (1984). The primate/rodent R/S ratio was calculated using the last 312 bp of the M. domesticus reading frame and the latest primate consensus sequence available from Singer and Skowronski (1985). Actual values were as follows: total R differences, 54.3; total S differences, 48.7; corrected R changes, 60.3; corrected S changes, 77.6; corrected R/S, 0.8; and SE, 0.3. A program for performing this calculation (RSTVS) was written in FORTRAN for execution on an IBM mainframe.

For R/S measurements on various groups of rodent sequences, SEs for the number of R (or the number of S) were taken to be the square root of R (or S) in accordance with the Poisson distribution. SEs for R/S values were calculated according to \( \sqrt{\frac{R}{S}} \sqrt{\frac{1}{R} + \frac{1}{S}} \). This equation is derived from \( \Delta(R/S)/(R/S) \). A value of \( \sim 1 \) Myr under selection before inactivation was calculated by fitting points from figure 3 to the equation \( R/S = \frac{[V_{ar}(x) + V_{ps}(t - x)]/[V_{as}(x) + V_{ps}(t - x)]}{[V_{ar} + V_{ps}(t - x)]/[V_{as} + V_{ps}(t - x)]} \), where \( V_{ar} \) and \( V_{pr} \) are rates of R for active and pseudogenes, \( V_{as} \), and \( V_{ps} \) are rates of S changes for active and pseudogenes, \( x \) is time under selection, and \( t \) is the average age of members in each category. The equation was simplified using \( V_{ar} = 0.8V_{as} \) and \( V_{pr} = 3.2V_{ps} \) (see above) and \( V_{ar} + V_{as} = V_{pr} + V_{ps} \) (found empirically for these data; see Results). Solving for \( x \) gives \( x = \frac{[t(3.2 - R/S)]}{1.33(1 + R/S)} \). The middle two points in figure 3B were used because greatest accuracy should be obtained where the branches have a nearly equal mixture of selected and unselected
Replacement/Synonymous Analysis of Rodent L1

descent. Actual values obtained were \( t = 2.35 \) Myr, \( R/S = 1.3 \), and \( x = 1.46 \) Myr for the second point; and \( t = 3.97 \) Myr, \( R/S = 2.7 \), and \( x = 0.4 \) Myr for the third point. Taken collectively, the data from these two points were \( t = 3.16 \) Myr, \( R/S = 1.94 \), and \( x = 1.0 \) Myr. The estimated range covered by one SE was 0.4–1.9 Myr.

Results

Replacement/Synonymous Analysis

In a functional coding sequence, \( R \) changes are more likely to be detrimental—and therefore selected against—than are \( S \) changes. As a consequence, the \( R/S \) ratio, as measured in a molecular phylogenetic reconstruction, discriminates between functional genes and pseudogenes. An example of the use of this method can be seen in the analysis of \( \beta \)-globin sequences (Czelusniak et al. 1982). Functional \( \beta \)-globin genes on average showed an \( R/S \) ratio of \( \sim 1.0 \), whereas the average for \( \beta \)-globin pseudogenes was 2.2. Pseudogenes are expected to have 2.5–3.0 times as many \( R \) as \( S \) changes because there are more potential sites for \( R \) changes. The analysis of the mouse \( \beta \)-globin pseudogene, \( \beta 2 \), further illustrates the sensitivity of this method. Its \( R/S \) value of 1.8 is intermediate between that of a functional gene and that of a pseudogene. Subsequent analysis (Phillips et al. 1984) showed that \( \beta 2 \) had a history that included some time as a functional gene and some time as a pseudogene. A systematic approach to timing the inactivation of pseudogenes based on this principle has been published (Miyata and Yasunaga 1981).

Throughout this paper the term “functional,” when applied to L1 sequences, refers to expression of the reading frame as evaluated by \( R/S \) values indicating selection. It does not refer to the capacity to transpose or to participate in other genetic or biological activities.

L1 Repeats Include Both Genes and Pseudogenes

The distribution of \( R \) and \( S \) changes in the tree relating 32 rodent L1 sequences is shown in figure 1. There is a low frequency (average of 4%; Martin et al. 1985) of base changes among these sequences; therefore, we expect a low incidence of parallel and back mutation. So the tree in figure 1 should be a good estimate of the historical sequence of base changes. The overall \( R/S \) ratio for these sequences is 1.6 ± 0.3. This is intermediate between the value expected for pseudogenes (3.2; see Methods) and that observed for the divergence of functional L1 sequences between rodents and primates (0.8; see Methods). So the lines of descent in figure 1 include a mixture of sequences under selection and pseudogenes.

The Molecular Drivers are Real Genes

Dover’s (1982) treatment of molecular drive in this family predicts that a subset of the elements should be the predominant sequence donors regardless of the mechanism that supports its concerted evolution. On an evolutionary tree, any sequence above a node is a donor, because it has given rise to a second copy of itself. We find that a small number of lineages can be chosen that include all of the nodes (thick branches, fig. 1). Each of these lineages defines the descent of a sequence that is frequently a donor. The small number of such sequences observed is in agreement with Dover’s analysis. Consideration of these lineages in isolation yields an \( R/S \) value of 0.9 ± 0.3, showing that they are functional genes under selection to make a protein. We refer to these sequences as molecular drivers. They are expected to drive their sequences into the genome in such a way that the long-term evolution of the entire
family will reflect the short-term behavior of the drivers. This expectation is fulfilled in this case, in that the $R/S$ value between rodents and primates matches that of the rodent drivers considered in isolation.

When constructing a scaled tree (fig. 2), we had to recognize the possibility that the molecular drivers and the other sequences might evolve at different rates. This tree was scaled based on driver lineages only (see Methods). In this way the estimate of the time when the different sequences arise from the driver would not be altered if, for example, the nondrivers evolve at a faster rate.

Other Functional L1 Genes Are Also Present

The average $R/S$ value among the branches that remain after the molecular drivers (those that end in a contemporary sequence) are factored out is $2.2 \pm 0.4$, still indicating
a mixture of functional genes and pseudogenes. When we categorized those branches by length, we discovered that short branches of one or two mutations showed the same R/S value as the molecular drivers and that longer branches showed an increasing R/S, until the pseudogene value was approached (fig. 3A). Furthermore, when the age of each branch was calculated from its position within the tree (fig. 2) rather than from its own length, the same result was obtained (fig. 3B). Thus we conclude that, on average, a nondonor sequence spends enough time under continued selective pressure to accumulate one or two mutations before being inactivated to become a pseudogene (see thin solid and dashed lines, fig. 1). Using the middle two points in figure 3B and an adaptation of the method of Miyata and Yasunaga (1981), we calculated that the average period of time spent under selection is ~1 Myr (see Methods). Note that this value is an average and that the actual behavior of any individual sequence cannot currently be measured.
Each L1 Gene Generates Multiple Pseudogenes

At this point it is important to consider the effect of mobile sequences on the structure of the tree. When a sequence moves, it takes the divergence of its parent with it and then accumulates further divergence at the new location. This should have no effect at all on the topology of the reconstructed tree. Each branch should still represent the history of a sequence, although the sequence may not have spent its entire past in the place where it was found. However, if the character of the sequence is different in the new location, there may be an abrupt shift in its rate of evolution in the middle of a branch.

In order to reconcile the indication of function on each branch with the frequent truncation of L1 sequences, we interpret the inactivation occurring on each branch to represent movement of the sequence to a new genomic location. That is, ~1 Myr after derivation of its sequence from the molecular driver, a functional gene generated a truncated processed pseudogene at another location, which was subsequently sampled. This means that there must be multiple functional L1 genes that are sequence donors for the creation of pseudogenes. The difference between the above donors that are not molecular drivers and the donors that are molecular drivers is in the nature of the progeny that they produce. Donors that are not molecular drivers produce progeny that have no further capacity to donate copies of themselves and therefore are at a dead end with respect to the process of concerted evolution. If the number of sequences sampled approaches the number of these secondary donors, then two progeny from the same donor may be sampled and a node defining that ancestor will appear on the tree. As these events recede into the past, they leave no lasting effect on the evolution of the family. On the other hand, the molecular drivers are genes that preferentially donate in exchanges among functional L1 genes. This guarantees that their sequences will dominate the family of functional genes—and subsequently the pseudogenes.
Thus the molecular drivers emerge as masters over a two-tiered system of sequence exchanges. The first tier is concerted evolution among functional L1 genes, the second is the production of pseudogenes. From this position, the behavior of the molecular drivers determines the major trends to be found in the tree—and the long-term tempo and mode of evolution of the entire L1 family.

Depending on the number of pseudogenes that are derived from each functional gene, there will be an amplification of sequences that are separated from the molecular driver by >1 Myr. Such an amplification is observed in the frequency distribution of repeats categorized by age (fig. 4). We see that repeats of between 1 and 2 Myr (containing one change from the molecular driver) are the most prevalent, while older repeats drop off in frequency, indicating the operation of a clearance mechanism (see below). However, the very youngest repeats (those with no differences from the molecular driver) are vastly underrepresented. The observed amplification occurs at ~1 Myr, corresponding to the average time of inactivation.

A Control against Computationally Introduced Artifacts

We then devised a method to show that these results will reflect the true behavior of the L1 family even if the tree from which they were calculated should turn out to be incorrect. Any reasonable tree, regardless of how it was inferred, will have the vast majority of private changes (variants found in one sequence only) on the nondonor lineages. Similarly, all changes that produced a base shared by more than one sequence (traditionally called phylogenetically informative changes) will mostly be placed on the donor lineages of any reasonable tree. This is because the alternative would require
such positions to be changed twice (i.e., would require a high incidence of parallel and back mutation). However, it is unlikely that many positions would be changed twice in a data set in which a large majority of positions are not changed at all.

We then tabulated the $R/S$ values for these two classes of base changes (see table 1), finding them to be 3.4 for private changes and 1.0 for the other changes. The private base changes were grouped according to total number of private changes per sequence and plotted against the group $R/S$ value (not shown) to derive a result similar to that in figure 3A. Similarly, the frequency distribution of private changes per sequence (not shown) mirrors the result in figure 4. So, our key observations are supported by a method that does not depend on any particular tree structure. The tree describing the true history must lead to similar results unless it is extremely unparsimonious.

A Model of the L1 Family

Our complete model for the history and population structure of the L1 family (fig. 5) specifies that all of the function is confined to a subset of presumably fully intact repeats. Concerted evolution within this subset is preferentially driven by a small number of its members, here referred to as molecular drivers. The functional L1 genes derive their sequences on an average of once every 1 Myr from the molecular drivers and frequently donate truncated copies to the genome, copies that thereafter evolve as pseudogenes.

The number of functional L1 genes apparently sampled in our tree is greater than 10 per species of rodent. Otherwise we would have sampled multiple truncated copies from the same parent, causing the nodes to appear at the time of inactivation

### Table 1
Categorization of Private Changes in L1 Tree as Replacements ($R$) or Synonymous ($S$) Changes

<table>
<thead>
<tr>
<th>SEQUENCE</th>
<th>TYPE OF CHANGE</th>
<th>TYPE OF CHANGE</th>
<th>TYPE OF CHANGE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SEQUENCE</td>
<td>$R$</td>
<td>$S$</td>
</tr>
<tr>
<td>P25</td>
<td>D7</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>P31</td>
<td>D2G</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>P1</td>
<td>D5</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>P27</td>
<td>D1</td>
<td>2.9</td>
<td>0.1</td>
</tr>
<tr>
<td>P22</td>
<td>D13</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>P22</td>
<td>D6</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>P26</td>
<td>D10</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>P21</td>
<td>D11</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>P23</td>
<td>D16</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>P29</td>
<td>D2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>D3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>D4G</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sums</th>
<th>$R$</th>
<th>$S$</th>
<th>$R/S$</th>
</tr>
</thead>
<tbody>
<tr>
<td>All private changes</td>
<td>64.1</td>
<td>18.9</td>
<td>3.4 ± 0.9</td>
</tr>
<tr>
<td>All nonprivate changes</td>
<td>25.6</td>
<td>26.4</td>
<td>1.0 ± 0.3</td>
</tr>
</tbody>
</table>
FIG. 5.—Model for the history of the L1 gene family. The upper part of the figure illustrates a family of functional L1 genes undergoing concerted evolution with a preferential donor. The lower part illustrates the associated family of processed L1 pseudogenes being distributed throughout the genome and subsequently being removed.

rather than before. There are some cases in our sample of sequences where this may have happened, e.g., D1 and D5. However, since there are no unambiguous cases, we suspect that even repeats like D1 and D5, which have identical parents, are probably paralogous. Since the sampling process involves a long period of time and the entire population of each species, models can be built with fewer functional L1 genes in any given mouse genome. In fact, a model can be built in which all functional L1 genes reside in a virus and never in a mouse genome. Assuming that all functional elements are intact over the entire 6–7 kb, their numbers would have to be small enough to be consistent with the number of observed intact elements, i.e., on the order of 10,000 (M. B. Comer, personal communication). Thus the possible range for the copy number of functional L1 genes per genome is rather large at this time.

The concerted evolution of the L1 family can be seen as the interplay of several continuous processes operating in a steady state. For the sake of the following illustration only, we will consider the sequence exchanges among the functional L1 genes as arising from a biased gene conversion mechanism, in the manner of Dover (1982). Then the limit of 1 Myr of functional divergence away from the molecular driver would simply represent the mean time required for any member of the functional L1 gene family to evolve away from the molecular driver before it was converted back to the driver sequence. All functional members of the L1 family would throw defective copies of themselves out into the genome at a constant rate. The amplification process, then, would not be a sudden burst but a smooth accumulation of copies throughout the period between successive conversions of the parent sequence.

Alternatively, consider a concerted evolution process for functional L1 genes in which they are continually created from the driver by a dispersal process similar to that which creates the pseudogenes. Then the 1-Myr limit for the divergence of functional genes from the driver would have to represent the average time until inactivation of functional genes. During this time they would be creating truncated pseudogenes, and at the time of inactivation they might persist as full-length pseudogenes themselves.
However, the loss of the capacity to make truncated copies would have to coincide with the loss of function.

This latter model uses similar mechanisms to explain the concerted evolution of the functional genes and the concerted evolution of the pseudogenes. However, to fit the data, some mechanism would have to cause the molecular drivers to repeatedly emerge from among the functional genes as the preferred sequence donor for future generations. This would have to be done without making the molecular drivers preferred direct donors of pseudogenes, or else the other functional genes would not have shown up in our analysis as intermediates in the flow of sequence information from driver to pseudogenes.

The Rate of Evolution of L1 Genes Is Unusually High

The divergence of mouse from primate L1 genes has been observed to be unusually high, with S sites changing faster than is normally expected even for pseudogenes (F. H. Burton, personal communication). For example, in the primate/rodent comparison we found 70% (48.7/69.7) of S sites to differ before correction for parallel and back mutation. By comparison, the large intervening sequences of the β-globins diverged by only 45% between these same species (Hardies et al. 1984). This trend is also apparent among the sequences examined in this paper (table 2). On average, from each node the subsequent length of the driver branch is similar to that of the pseudogene branch. Though there are fewer R changes to the drivers as a result of selection, there is a balancing increase in S changes. Therefore, on inactivation the rate of variation within the L1 pseudogenes accelerates at R sites but decelerates at S sites. Thus the rate of change of the L1 molecular drivers, before selection against unacceptable replacements, is high, even by comparison to that of their own pseudogenes. Furthermore, since the rate of evolution of the pseudogenes \(4.1 \times 10^{-9}\) changes/site-year; Martin et al. 1985) is typical of that for pseudogenes \(4.9 \times 10^{-9}\) changes/site-year; Li et al. 1981), we can say that it is the molecular drivers that are behaving oddly.

The Rate of L1 Pseudogene Turnover

The rate calculation of Martin et al. (1985) was revised upward for the pseudogenes by using a frequency distribution of only the pseudogenes (fig. 4). The curve, which describes the relative number of repeats of different ages, decays by half approximately every 2 Myr. Consistent with the conclusion of Martin et al. (1985), we conclude that

### Table 2
Total Length of Driver versus Nondriver Lineages

<table>
<thead>
<tr>
<th>DRIVER/NONDRIEVER PAIR</th>
<th>(R)</th>
<th>(S)</th>
<th>Total</th>
<th>(R)</th>
<th>(S)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>D4G/D16</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>D13/C11</td>
<td>5</td>
<td>3</td>
<td>8</td>
<td>7</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>C8/D7</td>
<td>3.6</td>
<td>11</td>
<td>14.6</td>
<td>9.6</td>
<td>1</td>
<td>10.6</td>
</tr>
<tr>
<td>P27/P26</td>
<td>0.5</td>
<td>3</td>
<td>3.5</td>
<td>0.5</td>
<td>1</td>
<td>1.5</td>
</tr>
<tr>
<td>P23/P1</td>
<td>6.9</td>
<td>4</td>
<td>10.9</td>
<td>7.1</td>
<td>1</td>
<td>8.1</td>
</tr>
<tr>
<td>Total</td>
<td>17</td>
<td>21</td>
<td>38 ± 6.2</td>
<td>25.2</td>
<td>5</td>
<td>30.2 ± 5.5</td>
</tr>
</tbody>
</table>

* Pairs were selected to avoid redundancy in the driver lineages.
one-half of the pseudogenes are lost and replaced by new copies every 2 Myr. So a clearance mechanism operates at about the same rate as the pseudogene generation mechanism, keeping the total copy number about the same. Also, we note a possible excess of older branches, which may represent sequences that have escaped the clearance mechanism. These conclusions about a clearance mechanism are based on the assumption that the rate of insertion is roughly constant with respect to time. Alternatively, one could explain the curve in figure 4 by proposing that the insertion process had been rapidly accelerating during the last few million years.

Martin et al. (1985) calculated a 3.3-Myr half-life for turnover of L1 sequences. Here we have resolved that figure into 1 Myr for sequence information to pass from the driver through functional genes into the pseudogene population and 2 Myr for the pseudogenes to turn over.

Discussion

To summarize our findings (see fig. 5), the L1 repeats consist of a set of functional L1 genes and a large associated family of pseudogenes. The term “functional” refers to the expression of the reading frame as evaluated by the suppression of R changes, not to the capacity to transpose or to participate in other genetic events. The pseudogenes undergo concerted evolution by continually being dispersed and turned over. A separate genetic mechanism causes concerted evolution among the functional L1 genes.

L1 sequences that generate additional copies of themselves are referred to as “donors” in this paper. Donor capacity and the capacity to function as a gene were evaluated separately with no assumption as to their relation. We found that donor capacity was correlated with evidence of functional-gene expression. We find no support for the idea that unexpressed L1 repeats ever propagate more copies of themselves. Our results are derived from averages of a number of repeats and so do not exclude other mechanisms operating at low frequency. However, the mechanisms that we identified are the ones that shape the overall dynamics of concerted evolution in the rodent L1 family.

Our results are consistent with the model suggesting that L1 repeats are mainly reinserted cDNAs that were reverse transcribed from RNA. This model explains the correlation between gene expression and donor capacity, since an mRNA would be intermediate for both. However, there may be other models that explain the correlation between expression and donor capacity.

The L1 Gene Family

All that we know about the functional L1 genes has been deduced from the reconstruction of their sequence from the sequences of the L1 pseudogenes. It is possible that none of the sequences actually characterized were functional L1 genes. Furthermore, it may be difficult to recognize a functional L1 gene even if it were cloned and sequenced. Much of what we know about the L1 repeats is really characteristics of the pseudogenes. For example, while the pseudogenes are interspersed among other genes, the functional L1 genes might just as well be clustered. The functional genes might have introns—and hence not conform to the consensus pseudogene restriction map. They might have homology with one another beyond the traditional boundaries of the repeat—and conceivably could contain structures not found in the pseudogenes, such as long terminal repeats.
A pseudogene generated from an L1 gene on a different chromosome will become separated from its parent during subsequent breeding cycles. In this way, the collection of pseudogenes in an extant animal will represent all of the L1 alleles in the previous interbreeding population. Consequently, each of our results contains an influence from population genetics as well as from intragenomic exchanges. We have not attempted to resolve these influences in this presentation.

There are several circumstances proposed under which selfish DNA can persist without benefiting the host organism (Doolittle and Sapienza 1980; Orgel and Crick 1980). We point out that the pattern of evolution in L1 is not consistent with at least the most prevalent paradigm for evolution of selfish DNA. In this case, a competent selfish sequence would have to constantly amplify itself at a sufficient rate to balance the loss of members resulting from inactivating mutations. Since the properties of the entire L1 family depend on the sequence of the molecular drivers, it is the rate of duplication and loss of molecular drivers that are relevant to the above process. In contrast to the high copy number and rapid amplification of L1 sequences in general, there are only one or two molecular drivers per species, and during most periods of time there are no duplications to produce new molecular drivers. Yet between such duplications there is a reduced (R/S) ratio for molecular drivers, suggesting that a form of selection based on something other than amplification has occurred.

Concerted Evolution among the L1 Genes

Concerted evolution proceeds among the functional L1 genes with a bias for two to three of the members per species to preferentially donate their sequences to the others. We have discussed how either gene conversion or a transposition mechanism would fit with our results (see Results). The mechanism of the concerted evolution among the functional L1 genes or the reason some donors should be preferred is unknown. The mechanism of concerted evolution among the functional L1 genes cannot be the same as that between the functional L1 genes and their pseudogenes. If it were, then the genes that drive the concerted evolution of the functional family would also directly produce most of the pseudogenes. Then the other members of the L1 gene family would not have shown up as intermediates in the flow of sequence information from the molecular drivers out into the pseudogene family.

Concerted Evolution of the L1 Pseudogenes

We considered a variation on the mechanism of concerted evolution of the L1 pseudogenes in which, after their original dispersal, they are gene converted many times by the functional L1 genes. In that case, the dynamics of their concerted evolution would predominantly be controlled by gene conversion rather than by the dispersal process. Jubier-Maurin et al. (1985) have argued for a gene conversion mechanism based on an analysis that included the sequences discussed in this paper. Gene conversion would remove an old copy for each new copy added and therefore naturally explain the overall balance of pseudogene production and clearance indicated by our analysis.

However, we conducted a test that favored the dispersal-only model over the dispersal-conversion model. It is based on the expectation that the small direct repeats of the flanking DNA generated during the initial insertion would not be affected by subsequent conversions. Thus we timed the true age of the insertion by the divergence of these small direct repeats. We averaged the data for multiple paired repeats to
increase the accuracy of the measurement. Five sequenced repeats from the mouse β-globin cluster (Voliva et al. 1984) contain no changes out of 114 total base pairs of direct repeat. When we use the rate of sequence divergence in a neutral sequence derived by Li et al. (1981) \((4.9 \times 10^{-9} \text{ changes/site-year})\), we see that this translates to an average age of \(<2 \text{ Myr} (4.9 \times 10^{-9} \times 114 \text{ sites} = 1 \text{ change/1.8 Myr})\). It appears that the insertion-conversion model can only be salvaged if a conversion mechanism can be found that rejuvenates the direct repeats. A stronger test of this possibility will be made when it is known how often the L1 pseudogenes are in the same locations in closely related species. Meanwhile, we think that it is a reasonable conclusion that the predominant flow of sequence from the L1 genes into the genome at large is by way of new insertions.

Possible Mechanistic Explanations for the L1 Pseudogenes

The high rate of both accrual and elimination of L1 pseudogenes requires some explanation. A purely mechanistic explanation for the accrual might be that the L1 gene happens to be expressed heavily in germ cells. Why they should be lost at such a high rate is more problematic. Half of all L1 pseudogenes are deleted every few million years, yet nonessential DNA in general is not thought to disappear that fast. A mechanistic explanation might be that the short direct repeats created on insertion are sufficient to specifically guide subsequent deletion. Short direct repeats in this size range are known to guide large deletions in \(E. \text{ coli}\) (Albertini et al. 1982). However, we would like to propose an alternative to a mechanistic explanation for the behavior of the L1 family, an alternative based on selection.

Potential Evolutionary Effects of the L1 Pseudogene Family

The possibility that repetitive sequences are involved in gene regulation has long been recognized (Davidson and Britten 1979, and references therein). We emphasize that the rate of insertion and deletion implicit in our results represents a large flux of genetic change. For example, the mouse β-globin gene cluster currently has eight L1 elements (Voliva et al. 1985). It is 65 kb long and has been separated from primates for \(\sim 100 \text{ Myr}\). Of \(10^5\) L1 copies, half will insert and half be deleted per 2 Myr, and \(10^5 \text{ events/2 Myr} \times 100 \text{ Myr} \times 65 \times 10^3 \text{ bp target size/3} \times 10^9 \text{ bp genome} = 108\) events in the rodent globin cluster since the separation from primates. Our times and rates are calibrated against a single speciation event (see Methods) and may subsequently be scaled up or down should better information become available. But the number of L1 insertions and deletions is still clearly going to be significant. If L1 insertions have even subtle effects on the expression of neighboring genes, then this process will be a major source of genetic variation. Such effects could be entirely passive, by being restricted to changing the spacing of active elements within the flanking sequences. Alternatively, the L1 genes may have acquired a transcriptional enhancer in the transcription unit, thus causing an active effect of their pseudogenes on neighboring genes. In either way, the high rate of insertion and deletion could reflect selective pressures rather than some unusual mechanism.

Such a process would bear on an important question in evolutionary biology. As suggested by Wilson et al. (1977), regulatory changes are a better candidate than structural changes to explain the vast morphological diversity arising during mammalian evolution. The proposed process fits the requirements perfectly. It would provide a continuous and plentiful source of subtle variation. The induced changes could also
be qualitative in the same sense that rearranged flanking sequence can cause hereditary persistence of fetal hemoglobin (for review, see Weatherall and Clegg 1982). We emphasize that the proposed effects need not be large, since evolution proceeds by a series of small changes.

Finally, no other protein-encoding family known has the strikingly peculiar properties associated with L1. This generates for us the expectation that these unusual features are a harbinger of unusual function.

Acknowledgments

S. L. Martin is a fellow of the Jane Collin Childs Memorial Fund for Medical Research. This research was supported by Public Health Service grants AI08998 and GM21313 from the National Institutes of Health.

LITERATURE CITED


A large interspersed repeat found in mouse DNA contains a long open reading frame that evolves as if it encodes a protein. Proc. Natl. Acad. Sci. USA 81:2308-2312.


WALTER M. FITCH, reviewing editor

Received May 17, 1985; revision received November 27, 1985.
Evolutionary Implication of Heterogeneity of the Nontranscribed Spacer Region of Ribosomal DNA Repeating Units in Various Subspecies of Mus musculus

Hitoshi Suzuki,* Nobumoto Miyashita,* Kazuo Moriwaki,* Ryo Kominami,† Masami Muramatsu,† Takeharu Kancheisa,‡ François Bonhomme,§ Michael L. Petras,‖ Ze-chang Yu,# and De-yuan Lu**

*National Institute of Genetics, Japan; †Tokyo University; ‡Kobe University; §Université des Sciences et Techniques du Languedoc, Montpellier, France; ‖University of Windsor, Ontario, Canada; #National Vaccine and Serum Institute, Beijing, People's Republic of China; and **Shanghai Second Medical College, People's Republic of China

Genetic variability of the nontranscribed spacer (NTS) region within ribosomal DNA repeating units in the various subspecies of Mus musculus was determined. Mice belonging to several laboratory mouse strains were examined by means of Southern blot hybridization with a mouse ribosomal DNA probe. This probe encompasses the 3' end of the 28S ribosomal RNA (rRNA) gene and the following spacer. Restriction enzyme digestions of the liver DNAs from various wild mice revealed that each of the subspecies has a unique pattern in the spacer encompassing a distance ~10 kb downstream from the ribosomal gene. These restriction patterns permit the classification of mouse subspecies and also provide insights into the origin of the laboratory mouse strains.

Introduction

Recent studies on the restriction fragment-length polymorphism (RFLP) of mitochondrial DNA (mtDNA) suggested that laboratory mouse strains were derived primarily from the European subspecies Mus musculus domesticus (Yonekawa et al. 1980, 1982; Ferris et al. 1982). However, since mtDNA is maternally inherited (Huchinson et al. 1974) and may cross subspecies boundaries (Ferris et al. 1983a; Boursot et al. 1984), its restriction patterns may not reflect the true evolutionary relationship of subspecies involved. It was therefore considered essential to examine nuclear DNA, specifically ribosomal DNA (rDNA), for such relationships. Variations in rDNA follow Mendelian patterns of inheritance.

There are 100–200 copies of mouse rDNA genes in the genome (Atwood et al. 1976). They are tandemly repeated in clusters at several different sites on the chromosomes. Each rDNA repeating unit is composed of a coding region that is transcribed into the precursor molecule for 28S, 5.8S, and 18S rRNA and a nontranscribed spacer (NTS). The NTS in rDNA is known to evolve rapidly in most of the higher eukaryotes (Arnheim 1983), and restriction analysis of such DNA has already been shown to

1. Key words: restriction fragment-length polymorphism (RFLP), nontranscribed spacer (NTS), ribosomal DNA, wild mouse subspecies.

Address for correspondence and reprints: Kazuo Moriwaki, National Institute of Genetics, Mishima, Shizuoka-ken, 411, Japan.

© 1986 by The University of Chicago. All rights reserved.
0737-4038/86/0302-3203$02.00
provide useful phylogenetic data (Wilson et al. 1984). Restriction-site maps of the rRNA gene region including the NTS have been constructed by Corry et al. (1977) and Kominami et al. (1981) for laboratory mice.

In this study, the restriction cleavage patterns of the NTS region of rDNA from various subspecies of Mus musculus (wild and laboratory strains) were compared. Geographical distribution of the rDNA haplotypes based on RFLP provides further insight into the origin of laboratory strains.

Material and Methods

Mice

The animals used in the present study included wild mice (see table 1 for the collecting localities); laboratory mice A/WySnJ, AKR/J, AU/SsJ, C3H/HeJ, C57BL/6J, C57BL/10J, DBA/1J, DBA/2J, PL/J, SM/J, and 129/SvSicp (all originally from the Jackson Laboratory); BALB/cAnN and CBA/CaHN (both originally from the National Institutes of Health); RFM/Ms (from the National Institute of Radiological Sciences, Japan); and NZB/San (from the Institute of Medical Science, University of Tokyo).

Blot Analysis

Nuclear DNAs were prepared from the livers of the mice as described by Maniatis et al. (1982). They were digested with EcoRI and/or BamHI (Takara Biochemicals, Kyoto, Japan) and electrophoresed on 0.6%-0.7% agarose gel at 3V/cm for 8-15 h in 40 mM Tris-acetate buffer (pH 7.8) containing 2 mM ethylene diaminetetraacetate and 0.5 μg/ml ethydium bromide. Then, the double-stranded DNA fragments were transferred to a nitrocellulose filter, heated at 80 C for 3 h, and hybridized with a probe of 0.7-kb double-stranded DNA in 6 X SSC as described by Southern (1975). (1 X SSC = 0.15 M NaCl-0.015 M Na citrate solution with pH 7.0.) The probe was prepared from an EcoRI-6.6-kb rDNA fragment cloned by Kominami et al. (1982). It included the HinfI restriction site at the 3'-end region of 28s rDNA and the BamHI site in the NTS downstream of the 3' end. Labeling of the probe was carried out using (α-32P)dCTP (Amersham Searle) and DNA polymerase I (Boehringer-Mannheim) (Maniatis et al. 1982). The specific radioactivity of the probe was 1-2 × 10^8 cpm/μg.

Molecular-Weight Determination of DNA

As shown in figure 1, molecular weights of the restriction fragments of the rDNA repeating units were determined using 0.6%-0.7% agarose gel and rDNA fragments such as 2.3-kb BamHI and 6.6-kb EcoRI fragments as molecular markers (Kominami et al. 1981).
FIG. 1.—A restriction map of mouse ribosomal RNA repeating units constructed according to Kominami et al. (1981). The 45S rRNA coding region is depicted by the elongated box. The darkened segments of the box labeled 18S and 28S indicate the genes coding for 18S and 28S rRNA, respectively. The 0.7-kb probe used hybridizes with mouse rDNA from the HinfI site on the 3’-end region of the 28S rRNA gene to the BamHI site (B4) on the following spacer region.

Results
Heterogeneity Involving EcoRI or BamHI Fragments of rDNA Repeating Units from Wild Mice

The EcoRI and BamHI digests of the liver DNAs obtained from 55 individuals collected at 31 sites throughout the world (table 1) were hybridized with the 0.7-kb rDNA probe. The patterns obtained are shown in figure 2 and summarized schematically in figure 3. The restriction fragments in the region from the 3’ end of 28S rRNA gene to the downstream spacer region varied in length in the different subspecies. As seen in figure 3, at least nine different EcoRI bands and seven BamHI bands were observed. These were 5.1–10.5 and 2.3–30 kb, respectively. Each subspecies seemed to have a characteristic major band, such as 2.3-kb BamHI in Mus musculus domesticus. These restriction patterns suggest that many types of DNA repeating units occur in the wild mouse populations investigated. In this paper each type is designated as an rDNA haplotype.

Restriction-Site Maps of rDNA Haplotypes

Eight rDNA haplotypes were recognized based on the sizes of the EcoRI, BamHI, and EcoRI + BamHI digests (table 2). To construct the restriction maps for these haplotypes, the genomic DNAs from seven individuals (DOM-LBL, SK/Cam, BRV-MPL, MUS-BLG3, MOA, CAS-QZN, BAC-LAH) were digested with both EcoRI and one of the following enzymes: BamHI, EcoRV, PvuII, PstI, and Scl.

On double digestion with EcoRI and EcoRV, only a single major fragment of 1.1 kb was observed in all the cases (data not shown). This suggested that the EcoRI...
Table 1
List and Source of Wild Mouse Subspecies Employed

<table>
<thead>
<tr>
<th>Subspecies and Site of Original Collection</th>
<th>Stock Designation</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Mus Musculus domesticus:</em></td>
<td></td>
</tr>
<tr>
<td>1. Pigeon (Canada)</td>
<td>DOM-PGN (2)</td>
</tr>
<tr>
<td>2. L. Belanger (Canada)</td>
<td>DOM-LBL (2)</td>
</tr>
<tr>
<td>3. Skokholm Island (United Kingdom)</td>
<td>SK/Cam(^a) (2)</td>
</tr>
<tr>
<td>4. Pomorie (Bulgaria)</td>
<td>DOM-BLG(^b) (1)</td>
</tr>
<tr>
<td>5. Langadas (Greece)</td>
<td>DOM-GRC(^b) (1)</td>
</tr>
<tr>
<td><em>M. m. brevirostris:</em></td>
<td></td>
</tr>
<tr>
<td>6. Montpellier (France)</td>
<td>BRV-MPL (2)</td>
</tr>
<tr>
<td><em>M. m. bactrianus:</em></td>
<td></td>
</tr>
<tr>
<td>7. Lahore (Pakistan)</td>
<td>BAC-LAH (2)</td>
</tr>
<tr>
<td>8. Kabul (Afghanistan)</td>
<td>BAC-KAB (2)</td>
</tr>
<tr>
<td><em>M. m. urbanus:</em></td>
<td></td>
</tr>
<tr>
<td>9. Bandarawela (Sri Lanka)</td>
<td>URB-BDW (2)</td>
</tr>
<tr>
<td><em>M. m. castaneus:</em></td>
<td></td>
</tr>
<tr>
<td>10. Bogor (Indonesia)</td>
<td>CAS-BGR (1)</td>
</tr>
<tr>
<td>11. Quezon City (Philippines)</td>
<td>CAS-QZN (2)</td>
</tr>
<tr>
<td>12. Taichung (Taiwan)</td>
<td>CAS-TCH (2)</td>
</tr>
<tr>
<td><em>M. m. musculus:</em></td>
<td></td>
</tr>
<tr>
<td>13. Vrania (Bulgaria)</td>
<td>MUS-BLG(^3) (b)</td>
</tr>
<tr>
<td>14. Northern Jutland (Denmark)</td>
<td>MUS-NJL (2)</td>
</tr>
<tr>
<td><em>M. m. molossinus:</em></td>
<td></td>
</tr>
<tr>
<td>15. Nakashibetsu, Hokkaido (Japan)</td>
<td>MOL-NSB (1)</td>
</tr>
<tr>
<td>16. Morioka, Iwate Prefecture (Japan)</td>
<td>MOL-MRO (1)</td>
</tr>
<tr>
<td>17. Omiya, Saitama Prefecture (Japan)</td>
<td>MOL-HOM (2)</td>
</tr>
<tr>
<td>18. Mishima, Shizuoka Prefecture (Japan)</td>
<td>MOL-MSM (2)</td>
</tr>
<tr>
<td>19. Anjo, Aich Prefecture (Japan)</td>
<td>MOA(^e) (3)</td>
</tr>
<tr>
<td>20. Teine, Hokkaido (Japan)</td>
<td>MOL-TEN2 (2)</td>
</tr>
<tr>
<td>21. Ohma, Aomori Prefecture (Japan)</td>
<td>MOL-OHM (2)</td>
</tr>
<tr>
<td>22. Hakozaki, Fukuoka Prefecture (Japan)</td>
<td>MOL-HKZ (1)</td>
</tr>
<tr>
<td>23. Kagoshima, Kagoshima Prefecture (Japan)</td>
<td>MOL-KAG (2)</td>
</tr>
<tr>
<td><em>Chinese subspecies (unidentified):</em></td>
<td></td>
</tr>
<tr>
<td>24. Changchun (China)</td>
<td>sub-CHC (1)</td>
</tr>
<tr>
<td>25. Urumchi (China)</td>
<td>sub-URM (2)</td>
</tr>
<tr>
<td>26. Jiayuguan (China)</td>
<td>sub-JYG (1)</td>
</tr>
<tr>
<td>27. Lanzhou (China)</td>
<td>sub-LZH (2)</td>
</tr>
<tr>
<td>28. Chengtu (China)</td>
<td>sub-CHT (1)</td>
</tr>
<tr>
<td>29. Beijing (China)</td>
<td>sub-BJN1 (2)</td>
</tr>
<tr>
<td>30. Nanjing (China)</td>
<td>sub-NAN (3)</td>
</tr>
<tr>
<td>31. Shanghai (China)</td>
<td>sub-SHH (2)</td>
</tr>
</tbody>
</table>

\(^a\) Inbred strain obtained through Dr. K. Kondo.
\(^b\) Original nomenclatures were DBP for BLG, DGD for GRC, BRV/2 for MPL, and MRV for BLG3.
\(^e\) Inbred strain developed by Dr. K. Kondo.

and EcoRV sites are well conserved in the mouse subspecies examined. The 0.7-kb probe hybridized with the region just downstream of the E2 site (see fig. 1). Therefore, the conservative EcoRV site is probably located 1.1 kb downstream of E2. Both single and double digestions placed BamHI, PvuII, and SacI sites upstream of E2. Those sites were also well conserved in all mice investigated. The restriction maps of the eight rDNA haplotypes (r1–r8) are presented schematically in figure 4.
Fig. 2.—Southern transfer and hybridization analysis on BamHI (a) and EcoRI (b) digests of purified mouse DNA using the 0.7-kb probe described in the text. The DNA came from (1) BALB/c; (2) SK/Cam; (3) DOM-LBL; (4) BRV-MPL; (5) BAC-LAH; (6) BAC-KAB; (7) URB-BDW; (8) CAS-QZN; (9) CASTCH; (10) MUS-BLG3; (11) MUS-NJI; and (12) MOA (see table 1). The lengths (kb) of the size standards (λDNA digested with HindIII) are indicated on the right.

Geographical Distribution of rDNA Haplotypes

The appearance of the EcoRI and BamHI bands suggested that each individual investigated had several rDNA haplotypes (see fig. 3). Microdensitometry of the autoradiographic bands permitted quantification of the relative amount of each haplotype in the total mouse genome.

The geographic distribution of each rDNA haplotype in the subspecies studied is plotted on the world map (fig. 5). Domesticus and brevirostris mice (nos. 1–6 in table 1 and fig. 5) collected from North America and Europe exhibited r1 and/or r2 haplotype(s). Wild mice obtained from Bulgaria (no. 13), the northern part of China (nos. 24–31), and the central part of Japan (nos. 15–21) generally had r4 haplotype, which appears to be characteristic of the musculus-molossimus group. In addition,
molossinus mice collected from seven localities in Japan (nos. 15–21) exhibited a characteristic minor haplotype, r5. Wild mice from Hakozaki (no. 22) in northern Kyushu had r4 and r8 haplotypes. The latter appears specific for bactrianus (nos. 7, 16).

Table 2

<table>
<thead>
<tr>
<th>rDNA HAPLOTYPES</th>
<th>EcoRI</th>
<th>BamHI</th>
<th>EcoRI + BamHI</th>
<th>TYPICAL EXAMPLES OBSERVED</th>
</tr>
</thead>
<tbody>
<tr>
<td>r1 . . . . . . .</td>
<td>6.6</td>
<td>6.5</td>
<td>5.1</td>
<td>SK/Cam, BRV-MPL,</td>
</tr>
<tr>
<td>r2 . . . . . . .</td>
<td>6.6</td>
<td>2.3</td>
<td>0.9</td>
<td>DOM-LBL, SK/Cam, BRV-MPL</td>
</tr>
<tr>
<td>r3 . . . . . . .</td>
<td>5.5</td>
<td>30</td>
<td>5.5</td>
<td>BRV-MPL</td>
</tr>
<tr>
<td>r4 . . . . . . .</td>
<td>9.0</td>
<td>6.5</td>
<td>5.1</td>
<td>MUS-BLG3, MOA</td>
</tr>
<tr>
<td>r5 . . . . . . .</td>
<td>8.8</td>
<td>6.3</td>
<td>4.9</td>
<td>MOA</td>
</tr>
<tr>
<td>r6 . . . . . . .</td>
<td>9.0</td>
<td>6.4</td>
<td>5.0</td>
<td>CAS-QZN</td>
</tr>
<tr>
<td>r7 . . . . . . .</td>
<td>6.5</td>
<td>6.4</td>
<td>5.0</td>
<td>CAS-QZN, BAC-LAH</td>
</tr>
<tr>
<td>r8 . . . . . . .</td>
<td>7.0</td>
<td>&gt;30</td>
<td>7.0</td>
<td>CAS-QZN&lt;sup&gt;b&lt;/sup&gt;, BAC-LAH</td>
</tr>
</tbody>
</table>

<sup>a</sup> Double digestion with EcoRI and BamHI.

<sup>b</sup> BamHI band in CAS-QZN was faint; length heterogeneity of the NTS BamHI fragment might be assumed: one being 30 kb and the other the longer.
8) and castaneus (nos. 10-12). A wild mouse collected from Kagoshima (no. 23) exhibited r6 and r8, both of which are castaneus type. Wild specimens collected near Nanjing (no. 30) carried r1, r4, and r8; those from Shanghai (no. 31) showed r4 and r7; and, finally, the major haplotypes of urbanus (no. 9) were r6, r7, and r8.

Southern Blot Analysis of Laboratory Mice

A blot analysis of repeating units on 15 laboratory mouse strains listed in Material and Methods was carried out. Most of the strains showed the domesticus-type NTS—that is, 6.5-kb and 2.3-kb BamHI fragments corresponding to the haplotypes r1 and r2 (see fig. 4). The SM/J strain, however, showed only the 2.3-kb BamHI band. Moreover, some strains, such as AU/SsJ, RFM/Ms, PL/J, CBA/CaHN, and NZB/San, also possessed minor fragments not found in wild domesticus. BamHI cleavage of CBA/
FIG. 5.—Geographical distribution of mouse rDNA haplotypes. Numbers refer to the collection sites of wild mice listed in table 1. Percentages of each haplotype except r1 and r2 were based on the relative concentration of "EcoRI band" corresponding to each haplotype. Percentage of r1 and r2 was based on the concentration of the 6.5-kb and 2.3-kb BamHI bands, respectively. Haplotypes containing <5% in each genome were neglected in this figure. The sites and bands represented by the unshaded portions of the circle are as follows: (4) and (5), r1; (6), r3; (8), "5.1-kb EcoRI band"; (12), "7.5-kb EcoRI band"; (13), "10.5-kb EcoRI band"; (15)–(19) and (21), r5; (20), r5 and "7.5-kb EcoRI band."
CaHN and NZB/San DNA yielded fragments more than 6.5 kb long. On EcoRI digestion, PL/J showed a 5.5-kb minor component(s). In addition, a 7.0-kb minor band was found in Au/SsJ and RFM/Ms strains.

Discussion
Genetic Classification of Wild Mice

The NTS region involving the area 10 kb downstream from the 3' end of the 28S rRNA gene was found to be highly variable with respect to the distribution of the restriction sites in wild subspecies of *Mus musculus*. At least 10 different kinds of rDNA haplotypes were recognized in the genomes of subspecies investigated. Restriction maps of eight of these are presented in figure 4. Each subspecies appears to have a characteristic combination of the haplotypes. Previously, Kuehn and Arnheim (1983) reported a polymorphism of the NTS region that lies 210 bp upstream from the origin of transcription of the rRNA precursor. Since nothing is known about the distribution of this polymorphism, its usefulness in distinguishing between subspecies is uncertain.

The patterns of rDNA organization in the seven subspecies studied—*domesticus*, *brevirostris*, *bactrianus*, *urbanus*, *castaneus*, *musculus*, and *molossinus*—suggest that this could provide us with a new set of criteria for their classification. The subspecies studied may be placed into three groups—*domesticus*, *castaneus*, and *musculus*—using the NTS haplotypes (r1-r8) summarized in table 3. This appears consistent with the conclusions previously reached following mtDNA restriction enzyme studies (Yonekawa et al. 1981; Ferris et al. 1983b) and the hemoglobin beta chain allelic frequencies (Miyashita et al. 1985). Subspecies *domesticus* and *brevirostris* are combined into the *domesticus* group because they have r1, r2, and r3; *musculus* and *molossinus* are placed in the *musculus* group because both have r4 and r5; and *bactrianus*, *castaneus*, and *urbanus* are combined into the *castaneus* group because they have r6, r7, and r8.

The above classification does not match perfectly with the three major biochemical groups (Mus1, Mus2, and Mus5) suggested by Bonhomme et al. (1984). In Bonhomme et al.'s classification, *castaneus* is placed in the *musculus* group and *bactrianus* in a distinct group. Which of these classifications is preferable remains to be settled.

Differences were also found between *brevirostris* (e.g., BRV-MPL) and *domesticus* (e.g., DOM-LBL, SK/Cam, DOM-BLG, and DOM-GRC). For instance, in BRV-

Table 3
Putative Genetic Classification of Mouse Subspecies Based on the Combinations of the rDNA Haplotypes

<table>
<thead>
<tr>
<th>Groups and Subspecies</th>
<th>Combinations of rDNA Haplotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domesticus:</td>
<td></td>
</tr>
<tr>
<td><em>Mus musculus domesticus</em></td>
<td>... r1, r2, r3</td>
</tr>
<tr>
<td><em>M. m. brevirostris</em></td>
<td></td>
</tr>
<tr>
<td>Castaneus:</td>
<td></td>
</tr>
<tr>
<td><em>M. m. bactrianus</em></td>
<td></td>
</tr>
<tr>
<td><em>M. m. urbanus</em></td>
<td></td>
</tr>
<tr>
<td><em>M. m. castaneus</em></td>
<td></td>
</tr>
<tr>
<td>Musculus:</td>
<td></td>
</tr>
<tr>
<td><em>M. m. musculus</em></td>
<td></td>
</tr>
<tr>
<td><em>M. m. molossinus</em></td>
<td></td>
</tr>
</tbody>
</table>
MPL, in addition to r1 and r2, ~50% of the haplotypes are r3, which is characteristic of *brevirostris*. Previous studies using either protein polymorphism (Bonhomme et al. 1978, 1984; Sage 1981; Thaler et al. 1981) or mtDNA (Ferris et al. 1983b) did not detect marked differences between *domesticus* and *brevirostris*. They are, therefore, often combined in a single taxonomic unit, *M. m. domesticus* (biochemical group Mus1). However, in the old classification of Schwartz and Schwartz (1943), *domesticus*, being dark colored and having a large body and long tail, is considered the northern form, whereas *brevirostris*, being light colored and white bellied and having a somewhat shorter tail and smaller body weight, is found in more southern climates. These two morphotypes may have a different ecological preference.

The rRNA haplotypes may also be used to give some insight into the evolutionary relationship among the subspecies involved. In this regard the r1, r4, r6, and r7 types, which are generally observed in wild mouse populations, may be especially informative. E2/B fragments of both r1 and r4 types are 0.1 kb longer than those of r6 and r7. Two explanations for this may be suggested. One is that r1 and r4 were derived from r6 and r7. Ancestral *castaneus*, which carried r6 and r7, moved west and in differentiating to *bactrianus* lost r6. Subsequently, in *domesticus*, the r1 was derived from r7 through the insertion of a 0.1-kb DNA in the E2/B fragment. In ancestral *castaneus*, which differentiated to *musculus* in the east or north, r6 became r4 by insertion of a 0.1-kb DNA. A second hypothesis involves the mixing of *domesticus*-like and *musculus*-like mice in ancient time. Thereafter, E2/B fragments of r1 and r4 were shortened by ~0.1 kb in the genomes of the *castaneus* ancestors. This gave rise to r6 and r7. Such changes have been well documented in repetitive DNA families (see Dover 1982 for review).

**Origin of Laboratory Mice**

The 15 inbred strains examined so far possess the *domesticus* type NTS: r1 and r2. This is consistent with the comparative analysis of chromosomal C-banding patterns (Moriwaki et al. 1982, 1985) and endonuclease-cleavage patterns of mtDNA (Yonekawa et al. 1980, 1982; Ferris et al. 1982, 1983b), which suggested that the whole mitochondrial genomes and most of the nuclear genes of inbred mice were derived from western Europe (*M. m. domesticus*). Some nuclear genes appear, however, to have originated in Asiatic mice (*M. m. molossinus* and others). Further, it is interesting to speculate on the origin of restriction fragments found in several of the inbred strains as minor components. For example, PL/J had 5.5-kb DNA, and AU/SsJ and RFM/Ms contained 7.0-kb DNA on EcoRI digestion. This suggests that *brevirostris* contributed to the establishment of the PL/J strain. However, it is not clear at present whether the 7.0-kb fragment found in AU/SsJ and RFM/Ms should be included in the category of r8 or in one of the other NTS haplotypes not used in the present study. Double digestion of 7.0-kb DNA with EcoRI and BamHI suggested the latter possibility (data not shown).

**Acknowledgments**

We are grateful to Dr. T. Gojobori for helpful discussions. The mice, MUS-NJL, were kindly supplied by Dr. J. P. Hjorth (University of Aarhus, Denmark). This study was supported in part by Grants-in-Aid for Scientific Research from the Ministry of Education, Science and Culture, Japan. This paper is contribution no. 1640 from the National Institute of Genetics, Japan.


MASATOSHI NEI, reviewing editor

Received August 6, 1985; revision received September 30, 1985.
This paper concludes that the statistical properties of protein evolution are compatible with a particular model of evolution by natural selection. The argument begins with a statistical description of the molecular clock based on a Poisson process with a randomly varying tick rate. If the time scale of the change of the tick rate of the molecular clock is assumed to be much less than the average time between substitutions, then it is shown that the substitution process must be episodic, with bursts of substitutions being separated by long periods of time with no substitutions. This analysis generalizes the recent work of Gillespie (1984a). The second part of the argument shows that a simple model of evolution by natural selection—one that incorporates a changing environment, the molecular landscape, and a simple form of epistasis—exhibits dynamics that are identical to those inferred from the statistical analysis. This leads to the conclusion that natural selection is a viable explanation for protein evolution. In addition, a correction formula for multiple substitutions is given that does not require that the substitution process be a Poisson process, and some comments on the inability of the neutral allele theory to account for the dynamics of the substitution process are presented.

Introduction

The forces responsible for the evolution of proteins remain a mystery despite many years of intensive study and debate. The two most conspicuous theories to account for molecular evolution are the neutral allele theory (recently reviewed by Kimura [1983]) and a less precisely defined theory of evolution by natural selection. From a mathematical point of view, the neutral allele theory is the more thoroughly studied and, partly for this reason, is the more generally cited as the probable cause of many of the differences that are observed in the amino acid sequences both between and within species. One of the most convincing arguments for neutrality and against natural selection has been the near constancy of the evolutionary rates of proteins. Kimura, in particular, has repeatedly argued that natural selection would lead to different rates of evolution in different lineages whereas neutrality would give similar rates. Thus, the issue of the mechanism of evolution is intimately connected with the variability of the rates of molecular evolution.

Despite the central importance of the variability of rates of evolution on our understanding of the mechanism of evolution, there appears to be very little published work that attempts to estimate the variability of rates using sequence data. When this has been attempted, as in Ohta and Kimura (1971), a naive model of rate variation is usually invoked. In this model the variation in the rate of evolution is due to dif-
ferences in rates between the lineages under study rather than to variation in rates within a lineage. This model allows us to infer the variability of the rates of evolution between lineages in a relatively straightforward fashion and has led to the generality that the rates of evolution do not exhibit much variation.

Recently, Gillespie (1984a) examined a somewhat different model of rate variation, one in which the rates of substitutions are allowed to vary through time down each of the lineages under study. Such a model allows variation both between and within lineages. This seemingly innocent generalization led to a complex statistical analysis that suggests that the variability of rates of evolution may be much larger than previously suspected. Paradoxically, this model has the property that large variabilities in rates of evolution are not necessarily manifest in large variabilities in the numbers of substitutions in different lineages. With a few additional assumptions, this analysis went even further to suggest that molecular evolution may be an episodic process, with bursts of substitutions being separated by long periods without any substitutions. It is this property of molecular evolution that suggested that the molecular clock be called an “episodic clock.”

These results are purely statistical, i.e., they do not depend in any way on a particular mechanism of molecular evolution. They are, however, strongly suggestive of a model of molecular evolution by natural selection that will be described in this paper. The model is particularly attractive because it exhibits the same episodic dynamics that were described in the statistical analysis. It is a model that appears to fit the statistical aspects of the sequence data as well as does the neutral allele model and that shares with that model a very appealing biological description.

Unlike the neutral allele model, the most important properties of this model of molecular evolution by natural selection cannot be described in a couple of simple equations. To make the results accessible to those unfamiliar with the requisite mathematics, the next section will provide an overview of the remainder of the paper. One could even view the next section as the paper proper and all of the subsequent sections as extended appendices.

Overview

The present paper consists of two logically independent parts: a statistical part that provides a description of the properties of the molecular clock and a mechanism part that describes a model of molecular evolution by natural selection. The main observation of the paper is that both of these parts arrive at the same description of the stochastic dynamics of molecular evolution. From this we conclude that natural selection is compatible with the statistical aspects of the sequence data.

The statistical part of the paper is further subdivided into two parts: the first part derives a very general formula for correcting sequence data for multiple substitutions at a site, and the second part generalizes the statistical procedure of Gillespie (1984a) for characterizing the episodic nature of the clock. The data for this statistical investigation come from “star” or “bush” phylogenies. A star phylogeny is one in which the radiation into a series of \( n \) species has occurred in a small period of time relative to the length of the lineages. Kimura (1983) was the first to recognize the value of star phylogenies for estimating the statistical properties of the substitution process. The best-known example of a star phylogeny is the mammalian radiation.

The object of study is \( N_i(t) \), the number of substitutions that occurred on the \( i \)th lineage of a radiation that occurred \( t \) years ago. Our goal is to estimate the mean and variance of the \( N_i(t) \): \( \mu = E[N_i(t)] \), \( \sigma^2 = \text{Var}[N_i(t)] \), where \( E \) stands for “expected value
of.” The data at hand for the estimation are the observed number of amino acids that differ between species \(i\) and \(j\), \(D_{ij}\).

To estimate the moments of \(N_i(t)\) using the \(D_{ij}\), corrections must be made for the fact that the \(D_{ij}\) are correlated with one another and for the occurrence of multiple substitutions at a site. The details of this correction are given in the next section. The results are as follows. Let \(M\) be Kimura’s statistic that is one-half the average value of the \(D_{ij}\):

\[
M = [n(n - 1)]^{-1} \sum_{i<j} D_{ij}. \tag{1}
\]

Let \(S^2\) be Kimura’s statistic that is proportional to the variance in the \(D_{ij}\):

\[
S^2 = [(n - 1)(n - 2)]^{-1} \sum_{i<j} (D_{ij} - 2M)^2. \tag{2}
\]

The results of the next section suggest that an estimator for the mean number of substitutions per lineage, \(\mu\), that corrects for multiple substitutions at a site, is given by

\[
\hat{\mu} = \log(1 - 2M/m)/2 \log(1 - 1/m). \tag{3}
\]

The estimator for the variance, \(\sigma^2\), is given by

\[
\hat{\sigma}^2 = (S^2 - VC)(1 - 1/m)^{-4\hat{\mu}}. \tag{4}
\]

In these expressions, \(m\) is the number of amino acids in the protein and \(VC\) is a small correction for compulsive types. When \(VC\) is set equal to zero, an error of no more than 5% will result in typical protein data. This error is much smaller than the sampling variance. Finally, much of the subsequent discussion will focus on the ratio of these two estimators:

\[
\hat{R} = \hat{\sigma}^2/\hat{\mu}, \tag{5}
\]

which is an estimator of the ratio of the variance to the mean of the number of substitutions per lineage,

\[
R(t) = \text{Var}\{N_i(t)\}/E\{N_i(t)\}. \tag{6}
\]

These estimators are used on actual data in table 2. Of particular interest in this table are the estimates of \(R(t)\). It has generally been assumed (incorrectly) that, under the neutral allele model or for constant rates of evolution, \(R(t)\) will equal one. This is because in both of these situations the \(N_i(t)\) are supposed (again incorrectly) to be Poisson distributed. When various authors (e.g., Ohta and Kimura 1971; Langley and Fitch 1974; Kimura 1983) have claimed to show that the rates of evolution are not constant, they have really shown that \(R(t)\) is significantly larger than one or that the \(N_i(t)\) are not Poisson distributed. Thus most of the discussion over the variability of rates is really a discussion about the values of \(R(t)\). As in other work, our results show that \(R(t)\) is always larger than one. The striking aspect of the estimates, however, is the narrow range of \(R(t)\) values: \(1.0 < R(t) < 3.4\). This is an awkward result: if one favors neutrality, then one must account for the fact that \(R(t)\) is larger than one (perhaps
by variable rates); if one favors natural selection, then one must account for the fact that \( R(t) \) is not larger than 3.4. It is the latter problem that will be solved in this paper.

A statistical description of the substitution process with a fluctuating rate of evolution is presented in the second of the two statistical sections. The aim of this section is to generalize the results of Gillespie (1984a). The starting point is a process called a doubly stochastic Poisson process, or Cox process, which is used as the statistical model for the substitution process, \( N_i(t) \). This process is just a Poisson process with a randomly changing rate of evolution. One major assumption is introduced in this section: the rate of change of the rate of molecular evolution is assumed to occur on a time scale that is much shorter than the length of the lineages under study. A typical lineage as used in table 2 is tens to hundreds of millions of years long. If the changes in the rates of evolution occur on a time scale that is on the order of thousands to tens of thousands of years—the time scale of the recent ice ages—then this assumption will be met. Given this assumption, the data support the view that the substitution process is an episodic process with bursts of substitutions followed by long periods with no substitutions. Table 3 gives estimates for the mean number of episodes and for the mean number of substitutions per episode for six proteins using two particular examples of the Cox process.

It is also shown in this section that the substitution process may be represented as a Poisson sum of positive random variables:

\[
N(t) = Y_1 + Y_2 + \ldots + Y_{M(t)}.
\]

In this representation the \( Y_i \) represent the number of substitutions that occurred in the \( i \)th episode of evolution and \( M(t) \) represents the number of episodes that occurred in the lineage during a time interval of \( t \) generations. In this formula, \( M(t) \) is a Poisson process. This representation formula is the main result of the statistical portion of the present paper. It gives a representation of the substitution process that will be used as a guide for the development of a model of molecular evolution by natural selection in the subsequent section.

The following section of the paper presents a model of molecular evolution by natural selection whose dynamics may also be represented by a formula of the form given above in equation (7). The model has three components: (1) a changing environment, (2) an epistatic scheme in which each environmental change presents a challenge to the species that may be met by substitutions at one of several nearly equivalent loci, and (3) a mutational landscape that makes it unlikely that any particular locus will experience the substitution of an allele that is two mutational steps away from the allele that is currently fixed in the population. To achieve the representation of equation (7), we must assume that the rate of environmental change is large and that the number of loci that are capable of responding to each environmental change is also large. A remarkable property of this model is that the values of \( R(t) \) that it predicts are in a very restricted range, say from 1.0 to \( \sim 3.5 \), just as observed in the data. This agreement, and the fact that the dynamics are compatible with the episodic structure of the data, show that a very plausible model of molecular evolution by natural selection fits the sequence data just as well as does the neutral allele model. Various caveats about the interpretation of these results are given in the general discussion at the end of this paper.
Correcting for Multiple Substitutions

In this section a correction formula for multiple substitutions in a single amino acid will be derived that does not require the number of substitutions per lineage to be Poisson distributed. The basic setting is that of a star phylogeny in which \( n \) lineages radiate from a single common ancestor at some point in the remote past. We will represent each of the \( m \) amino acids in a protein in each of the \( n \) extant species under study by an urn and each of the \( N_i \) amino acid substitutions that occur in the \( i \)th lineage by a ball (in this section the argument \( t \) in \( N_i(t) \) will be suppressed). If \( N_i \) substitutions have occurred in the \( i \)th lineage, then we throw \( N_i \) balls randomly into the \( m \) urns for that lineage with the probability of a ball going into the \( k \)th urn being \( p_k \). The number of balls in a lineage will be viewed as a random quantity. Models of molecular evolution frequently assume that \( N_i \) is Poisson distributed. Here we will assume only that the \( N_i, i = 1, \ldots, m, \) are independent, identically distributed, non-negative, integer-valued, random variables. The number of substitutions that separate two species is then the total number of balls found in their two sets of urns. Note that the assumption of independence of the \( N_i \) is a strong assumption that may be violated if the autocorrelation of the rate of molecular evolution is high relative to the length of the lineages.

The statistical problem arises because it is possible for two or more balls to land in the same urn. As experimentalists we can only say that a particular amino acid site differs between two species. We cannot know if that difference arises from one, two, or even more substitutions at a site. Thus we need to infer the number of substitutions that occurred when all that we know is the number of sites in which at least one substitution occurred. In terms of the urn model, we would like to infer the number of balls by observing the number of occupied urns. In fact, this goal is too ambitious. All that we can do is to estimate the mean and variance of \( N_i \) using the mean and variance of the number of occupied urns.

Let \( D_{ij} \) represent the number of urns that are occupied in either species \( i \) or species \( j \); that is, to calculate the contribution of the \( k \)th urn to \( D_{ij} \) we examine urn \( k \) in species \( i \) and species \( j \) and increment \( D_{ij} \) if the urn is occupied in either or both of the two species. \( D_{ij} \) is therefore a reasonable representation of the number of amino acids that are observed to be different between two species. It is not a perfect representation because we are not allowing the possibility that if two or more balls land in the same urn then a back substitution may occur. While this possibility could be included in the model, its occurrence in the data to be examined is so rare that ignoring it will not introduce any significant bias into the analysis.

Kimura (1983) suggested the two summary statistics based on \( D_{ij} \), \( M \) and \( S^2 \), given in equations (1) and (2) of the previous section. His development differed somewhat from ours in that he first transformed the observed number of differences between two species, \( D_{ij} \), by a function that is supposed to correct for multiple substitutions at a site. From that point on he assumed that \( D_{ij} = N_i + N_j \). In our development the correction will be carried out after the summary statistics are calculated. Our primary goal is to use these statistics to estimate the mean and variance of \( N_i \). This will be accomplished by first evaluating the expectations of \( M \) and \( S^2 \) and then using these results to suggest suitable estimators.

In urn models it is often more convenient to describe the number of unoccupied urns rather than the number of occupied urns. Thus we will let \( U_{ij} \) represent the number of urns that are unoccupied in both species \( i \) and \( j \):
The moments of the statistics $M$ and $S^2$, expressed in terms of the moments of $U_{ij}$, are

\[ E(M) = \frac{m - E(U_{ij})}{2}, \]
\[ E(S^2) = \frac{(n + 1)[2(n - 1)]^{-1}\text{Var}(U_{ij}) - 2(n - 1)^{-1}\text{Cov}(U_{ij}, U_{ik})}{2}. \]

Thus the task before us is to find approximations to the moments of the $U_{ij}$, plug these into the expressions for the expectations of $M$ and $S^2$, and from these guess some good estimators of the moments of $N_i$. (The details of the approximations to the moments of $U_{ij}$ are complex, tedious, and inappropriate for this journal; they will not be given here. The author will supply the details of the calculations to anyone who requests them.)

The approximations are

\[ E(U_{ij}) \approx m(1 - 1/m)^2a, \]
\[ \text{Var}(U_{ij}) \approx 2\sigma^2(1 - 1/m)^4a, \]
\[ \text{Cov}(U_{ij}, U_{ik}) \approx \sigma^2(1 - 1/m)^4a. \]

In arriving at these approximations, we took as our main assumption that the variation in the $p_k$—i.e., the probability of a substitution in the $k$th urn—is not excessive. To be precise, if we write $p_k = (1 + a_k)/m$, then we require that the sum

\[ m^{-1}\sum_i a_i^2 \]

be around one or less. In using this formula it may be necessary to exclude some sites that are known not to evolve.

Using these approximations, we get

\[ E(M) \approx m(1 - 1/m)^3a/2, \]
\[ E(S^2) \approx \sigma^2(1 - 1/m)^4a, \]

which immediately suggest the estimators

\[ \hat{\mu} = \log(1 - 2M/m)/2 \log(1 - 1/m), \]
\[ \hat{\sigma}^2 = S^2(1 - 1/m)^{-4a}, \]
\[ \hat{r} = \hat{\sigma}^2/\hat{\mu}. \]

At the level of approximation used in arriving at these estimators we have

\[ E(\hat{\mu}) \approx E[N_i(t)], \]
\[ E(\hat{\sigma}^2) \approx \text{Var}[N_i(t)]. \]

In deriving these approximations we assumed that the variance introduced by
the random placement of balls in urns is small relative to the variance introduced by the variation in $N_i$. In some cases, when $\mu$ and $\sigma^2$ are small, it may be necessary to take account of this extra element of variation. This is accomplished by using the estimator

$$\hat{\sigma}^2 = (S^2 - VC)(1 - 1/m)^{-4\bar{\mu}},$$

(15)

where $VC = (n + 1)(2(n - 1))^{-1}C_1 - 2(n - 1)^{-1}C_2$, and $C_1 = m(1 - 1/m)^{\bar{\mu}} + m^2(1 - 1/m)^{4\bar{\mu}} + m(m - 1)(1 - 2/m)^{\bar{\mu}}$, $C_2 = m(1 - 1/m)^{3\bar{\mu}} - m^2(1 - 1/m)^{4\bar{\mu}} + m(m - 1)(1 - 1/m)^{2\bar{\mu}}(1 - 2/m)^{\bar{\mu}}$. Fortunately, this correction is unnecessary for most data.

Because of the various approximations and biases that are inherent in these estimators, they must be checked for accuracy by simulations. Table 1 gives the results for a protein that has 150 amino acids with the mean number of substitutions per lineage being 14. This is very typical of the actual proteins that will be analyzed. As is obvious, the estimators are very good up to an $R(t)$ of approximately 5–6. Even here the bias in estimating $R$ is only $\sim 5\%$. The real data fall in the range $1.5 < R(t) < 3.5$, so the bias in the estimate of $R(t)$ can safely be ignored. The origin of the bias is from the covariance in $M$ and $S^2$ that is generated by the variability in $N_i$. While a correction for the bias could be derived, the computations are very complex for the small gain in precision. This point of view is reinforced by examining the SD of the estimator of $\hat{\sigma}$ in table 1. The SD exceeds by a large amount any correction that might be undertaken for $\hat{\sigma}$.

In table 2 the estimators are used to examine actual data. A comparison of the estimators obtained by our method with those obtained by Kimura (1983) shows that there is very little difference between them. This is surprising, since the justification

<table>
<thead>
<tr>
<th>$\mu$</th>
<th>$R(t)$</th>
<th>Species</th>
<th>$M$</th>
<th>$S^2$</th>
<th>$\hat{\mu}$</th>
<th>$\hat{\sigma}^2$</th>
<th>$\hat{\tau}$</th>
<th>$\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td></td>
<td>0</td>
<td>12.81</td>
<td>1.27</td>
<td>14.00</td>
<td>0.00</td>
<td>0.07</td>
<td>0.09</td>
</tr>
<tr>
<td>14</td>
<td></td>
<td>0</td>
<td>12.81</td>
<td>1.14</td>
<td>14.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.05</td>
</tr>
<tr>
<td>14</td>
<td>4</td>
<td>4</td>
<td>12.76</td>
<td>10.93</td>
<td>13.97</td>
<td>14.12</td>
<td>1.00</td>
<td>0.90</td>
</tr>
<tr>
<td>14</td>
<td>4</td>
<td>6</td>
<td>12.78</td>
<td>10.76</td>
<td>13.98</td>
<td>14.04</td>
<td>1.00</td>
<td>0.67</td>
</tr>
<tr>
<td>14</td>
<td>4</td>
<td>4</td>
<td>12.73</td>
<td>20.49</td>
<td>13.95</td>
<td>28.35</td>
<td>2.00</td>
<td>1.71</td>
</tr>
<tr>
<td>14</td>
<td>4</td>
<td>6</td>
<td>12.72</td>
<td>20.32</td>
<td>13.93</td>
<td>28.14</td>
<td>2.00</td>
<td>1.32</td>
</tr>
<tr>
<td>14</td>
<td>4</td>
<td>4</td>
<td>12.67</td>
<td>29.85</td>
<td>13.93</td>
<td>42.52</td>
<td>2.98</td>
<td>2.50</td>
</tr>
<tr>
<td>14</td>
<td>4</td>
<td>6</td>
<td>12.67</td>
<td>29.6</td>
<td>13.88</td>
<td>41.93</td>
<td>2.98</td>
<td>1.93</td>
</tr>
<tr>
<td>14</td>
<td>4</td>
<td>4</td>
<td>12.64</td>
<td>38.89</td>
<td>13.89</td>
<td>56.36</td>
<td>3.93</td>
<td>3.31</td>
</tr>
<tr>
<td>14</td>
<td>4</td>
<td>6</td>
<td>12.67</td>
<td>38.86</td>
<td>13.90</td>
<td>55.95</td>
<td>3.94</td>
<td>2.53</td>
</tr>
<tr>
<td>14</td>
<td>6</td>
<td>4</td>
<td>12.57</td>
<td>57.13</td>
<td>13.85</td>
<td>85.06</td>
<td>5.83</td>
<td>4.88</td>
</tr>
<tr>
<td>14</td>
<td>6</td>
<td>6</td>
<td>12.58</td>
<td>56.94</td>
<td>13.81</td>
<td>83.47</td>
<td>5.85</td>
<td>3.78</td>
</tr>
</tbody>
</table>

**Note.**—Each number represents the average of 50,000 replicates. For $R = 1$ a Poisson distribution for $N_i$ was used; for $R > 1$ a Polya-Aeppli distribution was used. $\mu$ = true mean of the number of substitutions per lineage; $R(t)$ = true value of the variance to mean ratio of the number of substitutions per lineage; $M$ and $S^2$ estimators given by eq. (1) and (2), respectively; $\hat{\mu}$ = estimated mean number of substitutions per lineage; $\hat{\sigma}^2$ = estimated variance in the number of substitutions per lineage; $\hat{\tau}$ = estimated value of $R(t)$; and $\sigma$ = average SD of estimator of $R(t)$. 

Table 1
Simulation Results Illustrating Accuracy of Estimators for Moments of $N_i$
Table 2
Estimates of Moments of $N_i(t)$ for Six Proteins

<table>
<thead>
<tr>
<th>Protein*</th>
<th>$N$</th>
<th>$m$</th>
<th>$M$</th>
<th>$S^2$</th>
<th>$\hat{\mu}$</th>
<th>$\hat{\sigma}^2$</th>
<th>$\hat{r}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta$-Hemoglobin</td>
<td>6</td>
<td>146</td>
<td>13.43</td>
<td>31.19</td>
<td>14.79</td>
<td>44.93</td>
<td>3.04</td>
</tr>
<tr>
<td>$\alpha$-Hemoglobin</td>
<td>6</td>
<td>141</td>
<td>11.57</td>
<td>11.29</td>
<td>12.59</td>
<td>14.74</td>
<td>1.17</td>
</tr>
<tr>
<td>Myoglobin</td>
<td>6</td>
<td>153</td>
<td>11.37</td>
<td>15.15</td>
<td>12.27</td>
<td>19.67</td>
<td>1.60</td>
</tr>
<tr>
<td>Cytochrome c</td>
<td>4</td>
<td>104</td>
<td>7.58</td>
<td>19.81</td>
<td>8.16</td>
<td>26.26</td>
<td>3.22</td>
</tr>
<tr>
<td>Ribonuclease</td>
<td>4</td>
<td>123</td>
<td>17.42</td>
<td>25.14</td>
<td>20.39</td>
<td>43.81</td>
<td>2.15</td>
</tr>
<tr>
<td>$\alpha$-Crystallin</td>
<td>6</td>
<td>175</td>
<td>4.10</td>
<td>10.42</td>
<td>4.19</td>
<td>11.36</td>
<td>2.71</td>
</tr>
</tbody>
</table>

* Data for all of the proteins except $\alpha$-crystallin were taken from tables 4.3-4.4 in Kimura (1983). The $\alpha$-crystalline data is from De Jong (1982) for human, rat, rabbit, dog, horse, and cow.

NOTE.—$N =$ no. of species; $m =$ no. of amino acids; $M$ and $S^2 =$ estimators given by eqq. (1) and (2) respectively; $\hat{\mu} =$ estimated mean number of substitutions per lineage; $\hat{\sigma}^2 =$ estimated variance in the number of substitutions per lineage; and $\hat{r} =$ estimated value of $R(t)$.

The General Clock

The estimates of $R(t)$ and the mean rate of substitution should provide some insight into the process of molecular evolution. Such insight will only come after the adoption of a specific stochastic model of molecular evolution. In this section, the model to be employed is a statistical model, being totally independent of any genetically based models. The intent is to use such a model to gain insight into the dynamics of the process without reference to the underlying mechanism. Discussion of the mechanism will be deferred until the next section.

As was done in Gillespie (1984a), assume that the substitution process along any lineage is a doubly stochastic Poisson process, $N_i(t)$, with the rate of evolution being

$$\lambda \theta(t),$$

where $\theta(t) > 0$, $E[\theta(t)] = 1$, $\text{Cov}[\theta(t), \theta(t + x)] = r(x)$; that is, the probability of a substitution occurring in a small period of time, $dt$, is

$$Pr[dN_i(t) = 1] = E[dN_i(t)]dt = \lambda \theta_i(t)dt.$$
A good treatment of Cox processes is given by Cox and Isham (1980). The process has been indexed by $\tau$, which will play a role in the limiting operation that will be described shortly. A well-known result that follows immediately from these assumptions is that the total number of substitutions in a lineage, $N_\tau(t)$, given a single trajectory of $\theta_\tau(t)$, is Poisson distributed with mean

$$\lambda \xi_\tau(t) = \lambda \int_0^t \theta_\tau(x)dx. \quad (18)$$

In Gillespie (1984a) it was argued that the time scale of change of the "molecular clock," $\theta_\tau(t)$, is probably much shorter than the length of a typical lineage under study. This assumption is motivated by the fact that major environmental changes, such as the recent ice ages, occur on a time scale of thousands to tens of thousands of years while the time between substitutions is typically on the order of millions of years. If the rate of evolution is changing on a time scale of thousands of years, the autocorrelation of the rate will approach zero very quickly on the time scale of the substitution process. Thus it is natural to consider the behavior of the integral $\xi_\tau(t)$ as the autocorrelation as the process approaches zero. However, as the autocorrelation approaches zero, so does the variance in $\xi_\tau(t)$ as given by the equation

$$\text{Var}[\xi_\tau(t)] = 2 \int_0^t (t - x)r_\tau(x)dx. \quad (19)$$

In order to keep the variance from going to zero (required by the fact that $R(t)$ is in the range $1.5-3.5$), we must let the variance of $\theta_\tau(t)$ approach infinity as the autocovariance for nonzero lags approaches zero in such a way that the variance of $\xi_\tau(t)$ approaches a nonzero limit. As these limits are taken, we want to discover the limiting behavior of the process $\xi_\tau(t)$.

Suppose we have a sequence of processes, $\theta_\tau(t)$, such that, as $\tau \to 0$,

$$r_\tau(x) \to 0 \quad \text{for} \quad x \neq 0,$$

$$r_\tau(0) \to \infty, \quad (20)$$

such that $\text{Var}[\xi_\tau(t)] \to vt$ and such that the limiting process exists and is a process with stationary, independent increments. We will call the limiting process $\xi(t)$. $\xi(t)$ will obviously be a process with orthogonal increments. This is true because the integrals of $\theta_\tau(t)$ over disjoint intervals will be asymptotically uncorrelated since the autocovariance function approaches zero (as $\tau \to 0$) for nonzero lags. However, orthogonality does not imply independence so, to be cautious, we will restrict the domain of our discussion to those sequences of processes with limiting processes having independent increments.

A nondecreasing process with stationary, independent increments, such as $\xi(t)$, is called a subordinator. Its distribution is determined by the Levy equation

$$E[e^{-\xi(t)}] = e^{-\psi(z)}, \quad (21)$$

where $\psi(z) = \int_0^\infty [1 - e^{-zx}]\mu(dx)$ (see Kingman [1975] for a discussion of subordinators.
and Breiman [1968, section 14.4] for the background on processes with independent increments). The measure, $\mu(x)$, called a Levy measure, characterizes the process. This equation is particularly useful for our purposes. The number of substitutions, $N(t)$, is a Poisson process randomized by $\xi(t)$; thus its probability-generating function is

$$E[e^{-\xi(t)}] = e^{-\eta(t)}.$$  

(See Feller [1968, chap. 11] for a discussion of probability-generating functions [pgf] and his formula 2.9 for the pgf of the Poisson distribution.)

In order to achieve the desired generalization of the results in Gillespie (1984a), we must see whether or not a random variable with this pgf may be represented as a Poisson sum of positive random variables, say

$$N(t) = Y_1 + Y_2 + \ldots + Y_{M(t)},$$  

where $M(t)$ is Poisson distributed and the $Y_i$ are independent, identically distributed positive random variables. The answer is clearly in the affirmative. If the pgf of $Y_i$ is

$$g(s) = \psi(\lambda)^{-1} \int_0^{\infty} e^{-\lambda x} e^{\lambda x} dx,$$  

and if

$$E[M(t)] = \alpha \psi(\lambda),$$

then the pgf of the sum will be given by $e^{-\eta(t)}$, as required.

Somewhat more about the distribution of the $Y_i$ can be learned from the pgf, $g(s)$. By expanding $g(s)$ in powers of $s$, we have

$$Pr(Y_i = j) = \lambda[\psi(\lambda)^{-1} \int_0^{\infty} e^{-\lambda x} x^j dx].$$

The moments of $Y_i$ may be obtained by differentiating $g(s)$:

$$E(Y_i) = \lambda \psi(\lambda)^{-1},$$

$$\text{Var}(Y_i) = \lambda \psi(\lambda)^{-1}[\lambda \mu_2 + 1 - \lambda \psi(\lambda)^{-1}],$$

$$\mu_2 = \int_0^{\infty} x^2 \mu(dx).$$

From these moments can be derived the moments of $N(t)$,

$$E[N(t)] = \lambda t,$$

$$\text{Var}[N(t)] = \lambda t(\lambda \mu_2 + 1).$$

We will now give two examples of processes that fulfil the assumptions of this development. The first is taken from Gillespie (1984a). For $\theta_d(t)$ we use a two-state Markov process. The process remains in state zero for an exponentially distributed
time with mean $1/\mu_0$ and then jumps to the value $1/(\mu_0 \tau)$, where it remains for an exponentially distributed time with mean $\tau$ before returning to zero, and so forth. This process will be called the two-state clock. As shown in Gillespie (1984a), as $\tau \to 0$, the limiting process may be represented as a Poisson sum of exponentially distributed random variables:

$$\xi(t) = Z_1 + Z_2 + \ldots + Z_{k(t)}.$$  

(29)

This process has the Levy measure

$$\mu(dx) = \mu_0^2 e^{-\mu_0 dx},$$  

(30)

with $\mu_2 = 2/\mu_0$. Using these results we see that $Y_i$ is geometrically distributed with mean $(\mu_0 + \lambda)/\mu_0$ and that $M(t)$ is Poisson distributed with mean $\mu_0 \lambda t/(\mu_0 + \lambda)$. The distribution of a Poisson sum of geometric random variables is called a Polya-Aeppli distribution (Johnson and Kotz 1969). These results agree with those obtained in Gillespie (1984a) by a very different approach.

Another example that is very different from the previous one is a process that starts at zero, where it remains for an exponentially distributed length of time with mean $\tau$ and then jumps to a gamma-distributed height with mean $l/2$ and variance $\tau/\lambda^2$, where it remains for time $\tau^2$ before returning to zero, and so forth. This process will be called the gamma clock. Again we let $\tau \to 0$ and discover that the limiting process, $\xi(t)$, is gamma distributed with Levy measure

$$\mu(dx) = (v x)^{-1} e^{-v x} dx,$$  

$$\mu_2 = v.$$  

(31)

This process is examined in detail in Kingman (1975). Unlike the previous case, $\xi(t)$ cannot be represented as a Poisson sum of random variables. Rather, $\xi(t)$ may be thought of as increasing in a countably infinite number of steps, “most” of which are “zero” in magnitude. There is some similarity with the previous case in that the number of jumps that are larger than any number larger than zero is Poisson distributed. For this case $Y_i$ has a logarithmic series distribution with mean $\lambda \nu/\log(1 + \lambda \nu)$ and $M(t)$ is Poisson distributed with mean $t \log(1 + \lambda \nu)/\nu$. Since $N(t)$ is a Poisson sum of a logarithmic series of distributed random variables, it has a negative binomial distribution.

The contrast between these two clocks stems from the fact that one has a Levy measure with finite total mass while the other has infinite total mass. In the former case the clock “ticks” a Poisson number of times in a finite interval, so it is natural to think of each of these ticks as an episode with a rapid rate of evolution even if no substitutions actually occur. For the latter clock, however, one cannot represent the clock as a Poisson sum of random variables (even though $N(t)$ can be represented in this way), so one cannot as naturally conceive of some quantity that represents the number of episodes of high evolutionary rate. Clearly, the common ground between these two clocks is in the representation of the number of substitutions rather than in the number of episodes. We will henceforth refer to an episode of evolution as occurring whenever at least one substitution occurs. This differs from the usage in Gillespie (1984a), where an episode referred only to the two-state clock and then to a case in
which the clock was not zero even if a substitution did not occur. The new usage is
dictated by the more general treatment of the clock.

Using the expressions from the previous sections for the expectation of the esti-
mators $\hat{\mu}$ and $\hat{\sigma}^2$, we see that

$$E(\hat{\mu}) \approx E[N(t)] = \lambda t,$$

$$E(\hat{\sigma}^2) \approx \text{Var}[N(t)] - \lambda t(\mu^2 + 1). \tag{32}$$

A minor calculation using these results shows that, for the two-state clock, $(\hat{t} - 1)/2$
is a consistent estimator of $\lambda/\mu_0$, the mean number of substitutions per episode, and
$2\hat{\mu}/(\hat{\sigma}^2 - \hat{\mu})$ is a consistent estimator of the mean number of episodes per lineage.
For the gamma clock, $(\hat{\sigma}^2/\hat{\mu} - 1)/\log(\hat{\sigma}^2/\hat{\mu})$ is a consistent estimator for the mean
number of substitutions per episode and $\hat{\mu}^2\log(\hat{\sigma}^2/\hat{\mu})/(\hat{\sigma}^2 - \hat{\mu})$ is a consistent estimator
of the mean number of episodes per lineage. These estimators are applied to actual
data in table 3. One remarkable aspect of this analysis is the similarity of the inferred
episodic events for the two clocks. Given the very different characters of the two
clocks, this suggests that the inferred episodic structure is robust to the assumptions
made about the clock, at least for the restricted range of values of $R(t)$ for the currently
available data.

Models of Molecular Evolution

The analysis of the previous section suggests that the substitution process for a
single locus may be represented by the random sum

$$N(t) = V_1 + V_2 + \ldots + V_{M(t)}, \tag{33}$$

where the $V_i$ are independent, identically distributed, positive, integer-valued random
variables and $M(t)$ is a Poisson process. In this section an attempt will be made to
attach some biological significance to this representation by exploring models of mo-
lecular evolution based on the action of natural selection.

For various technical reasons theoretical work on models of molecular evolution
based on natural selection has lagged far behind the work on neutral models. However,

Table 3
Inferred Episodic Structure of Evolution for Two Clocks

<table>
<thead>
<tr>
<th>Protein</th>
<th>Mean No. of Substitutions per Episode</th>
<th>Mean No. of Episodes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Two-State</td>
<td>Gamma</td>
</tr>
<tr>
<td>$\beta$-Hemoglobin</td>
<td>2.02</td>
<td>1.83</td>
</tr>
<tr>
<td>$\alpha$-Hemoglobin</td>
<td>1.09</td>
<td>1.08</td>
</tr>
<tr>
<td>Myoglobin</td>
<td>1.30</td>
<td>1.28</td>
</tr>
<tr>
<td>Cytochrome c</td>
<td>2.11</td>
<td>1.90</td>
</tr>
<tr>
<td>Ribonuclease</td>
<td>1.57</td>
<td>1.50</td>
</tr>
<tr>
<td>$\alpha$-Crystalline</td>
<td>1.86</td>
<td>1.71</td>
</tr>
</tbody>
</table>
in a recent series of papers, Gillespie (1983a, 1983b) has developed a boundary layer-approximation technique for multiple allele-diffusion models that allows one to describe molecular evolution in a particularly simple way. The approach is similar to that used by Maynard Smith (1976) in another context. In summary, it was shown that if the mutation rate to an advantageous allele is very small, say $10^{-9} - 10^{-8}$, and if the selection coefficient of the allele is large relative to the reciprocal of the population size, say 100-1,000 times larger, then the advantageous allele will spend an exponentially distributed length of time in the boundary layer, i.e., at a very low frequency, before sweeping through the population. The time required to sweep through the population will be short relative to the time spent in the boundary layer. The only alleles that are likely to exhibit this behavior are alleles that are one mutational step away from alleles that are currently fixed or segregating at high frequency in the population. If there is more than one advantageous allele, then a particular allele's probability of being the first one to become fixed is its selective advantage divided by the sum of the selective advantages of all of the advantageous alleles.

Using this boundary-layer description of the substitution process we can easily describe a population's response to a changing environment. Let $P(t)$ be a stationary-point process that represents the times of change of those aspects of the environment that cause a change in the relative fitnesses of one or more of the alleles at the locus under study. With each environmental change the locus could respond with zero, one, or more substitutions with a boundary-layer dynamic. This suggests representing the substitution process as the random sum

$$ S(t) = X_1 + X_2 + \ldots + X_{P(t)}. \quad (34) $$

In this representation, the $X_i$ represent the number of substitutions at the locus that occur following the $i$th environmental change. In general, we would expect $X_i$ to be greater than zero only if there are alleles that are both more fit than and one mutation step away from those alleles that are currently in the population. Given that $X_i > 0$, its properties are determined to a great extent by the "mutational landscape" as described in Gillespie (1984b). There it is argued, using some results from extreme value theory, that the mean value of the $X_i$, given that $X_i > 0$, are likely to be in the range 1.7-2.7.

The similarity of this representation of molecular evolution to the representation based on the statistical analysis suggests equating $M(t)$ with $P(t)$ and $Y_i$ with $X_i$. There are, however, two problems with this equivalence. The first is that $X_i$ can equal zero while $Y_i$ cannot. The second is that $P(t)$ will not, in general, be a Poisson process. The representations can be made more similar by modifying $P(t)$ to mark only those events for which $X_i > 0$ and to substitute for $X_i$ the random variable, call it $Z_i$, whose distribution is equal to that of $X_i$ conditioned on $X_i > 0$. The substitution process could now be written as

$$ S(t) = Z_1 + Z_2 + \ldots + Z_{Q(t)}, \quad (35) $$

where $Q(t)$ is a point process for which each event is an event of $P(t)$ that led to a substitution. This simple expedient obviously handles the fact that the $X_i$ might equal zero and, less obviously, makes $P(t)$ more Poisson-like. The transformation of the process $P(t)$ into the process $Q(t)$ is called thinning in the point-process literature. It
is a well-known result that thinning of a point process leads, asymptotically, to a Poisson process (see Cox and Isham 1980, pp. 98–100). However, even though the process \( P(t) \) is made more Poisson-like by considering only those environmental changes that lead to substitutions, there is an extension to this one-locus model that will make the convergence of \( P(t) \) to a Poisson process even more compelling. It is an argument based on a simple form of epistasis.

Suppose that each environmental change presents the population with a challenge that can be met in equivalent ways by substitutions at many different loci. Suppose, for example, that there are \( L \) loci that are capable of responding to an environmental challenge by the fixation of one or more alleles. We suppose, for simplicity, that these \( L \) loci are equivalent in a long-term evolutionary sense. By this we mean that with each environmental challenge, each of the \( L \) loci is equally likely to experience an allelic substitution that will mollify the challenge. However, the reason that a particular locus is the one to experience the substitution depends on a number of factors. First among these is that, as discussed in the one-locus model, the locus must have at least one allele that is more fit than the alleles currently in the population and is a neighbor on the “mutational landscape,” that is, an allele that is more fit and is one mutational step away. As was argued in Gillespie (1984b), alleles that are two or more mutational steps away from those currently in the population are unlikely to enter the population because of the extraordinarily low rate of production of these double mutants. Thus, with each environmental change, some of the \( L \) loci will be sitting on peaks in the mutational landscape, with further evolution in the current environment unlikely, while other loci will be on the sides of peaks, with the potential for allelic substitutions.

When the environment changes, only a subset of the \( L \) loci will be on the sides of peaks on the mutational landscape. Among these, there will be a tremendous range in the strength of selection acting on the advantageous alleles. Those loci whose advantageous alleles have the largest selective advantage are the ones that are most likely to experience the substitutions that satisfy the environmental challenge. This is the second factor that will determine the probability that a particular locus among the \( L \)-equivalent loci will be the one that is chosen.

There are undoubtedly other factors that will contribute to the probability of a particular locus being chosen, but these two appear to be the most important. Thus, with each environmental change there will be a small subset of the \( L \) loci that are likely to respond: those off peaks with the strongest selection. In some cases, there may be no loci that can respond. In those cases when more than one locus are capable of responding with similar selection coefficients, the substitutions should occur with equal probability among them. With each environmental change, we picture another randomly chosen set of loci being the ones that are likely to respond with substitutions. These ideas will now be investigated mathematically through an approximation that assumes that the number of loci and the rate of environmental change are large.

If \( L \) is large, the probability of at least one of the substitutions occurring at any particular locus must be small, say equal to \( \alpha/L \), where \( \alpha \) will depend on the particulars of the model. Given that at least one substitution occurs, let the pgf of the total number that occurs at a given locus be \( g_L(s) \). The pgf for the number of substitutions at a locus following an environmental change is then

\[
1 + (\alpha/L)(g_L(s) - 1).
\]
If the total number of environmental changes is represented by the point process \( P(t) \), then the number of substitutions at a particular locus has the pgf

\[
E\{ [1 + (\alpha/L)[g_L(s) - 1]]^{P(t)} \}. \tag{37}
\]

To arrive at an approximation of this formula equation, let the number of environmental changes and the number of loci increase such that \( P(t)/L \to \lambda t \) and let \( g_L(s) \to g(s) \). The pgf of the total number of substitutions will then become, asymptotically,

\[
E\{ e^{P(t)\log[1+\alpha/L][gL(s)-1]} \} \sim E\{ e^{\lambda g(s)-1}P(t)/L} \}
\sim E\{ e^{\lambda g(s)-1} \}. \tag{38}
\]

This will be immediately recognized as the pgf for a Poisson sum of random variables of precisely the same form as that obtained by the statistical model

\[
S(t) = Z_1 + Z_2 + \ldots + Z_{Q(t)}, \tag{39}
\]

where \( Z_i \) has the pgf, \( g(s) \), and where \( Q(t) \) is a Poisson process with mean \( \lambda t \).

As in the single-locus case, the limiting argument used to achieve this representation involves thinning. The effect of thinning on the variance-to-mean ratio is given by

\[
R(t) = \frac{\text{Var}[S(t)]}{E[S(t)]} \sim E(Z_i) + \frac{\text{Var}(Z_i)}{E(Z_i)} + (\alpha/L)E(Z_i)t(t), \tag{40}
\]

where \( t(t) = \text{Var}[P(t)]/E[P(t)] \) is the index of dispersion of the process of environmental change. This expression shows how the variance-to-mean ratio is lowered as the number of loci is increased.

The expression

\[
\kappa = E(Z_i) + \frac{\text{Var}(Z_i)}{E(Z_i)} \tag{41}
\]

requires some additional discussion. In was shown in Gillespie (1984b) that excursions through the mutational landscape lead to values of \( \kappa \) in the range 2.0–3.5. In the context of the epistatic model, this range would be expected to extend down to one. The reason is that with each environmental change there could be more than one locus that is a candidate for responding with a substitution. If the total number of candidate loci is larger than the total number of substitutions that occurs, then each locus is unlikely to experience more than one substitution. Since \( Z_i \) is conditioned to be greater than zero, \( Z_i \) would equal one with a probability of nearly one, yielding a \( \kappa \) of nearly one. In the original description of the model this would correspond to \( \alpha = 1 \) and \( g(s) = s \). Thus the prediction of our model is that the \( R(t) \) should be in the range \( 1.0 < R(t) < \approx 3.5 \), exactly as observed in the data.

This argument is only valid if \( L \), the number of loci that can respond to a particular environmental change, is large. It may well be the case that for certain environmental
changes $L$ is very small, perhaps even equal to one. For these cases the value of $R(t)$ could be rather large if the process of environmental change, $P(t)$, has a large index of dispersion.

**General Discussion**

This paper presents a statistical analysis of protein sequence data and a model of natural selection that share a common episodic structure. Moreover, the model specifically predicts that the values of $R(t)$ observed in the data should be in the range $1.0 < R(t) < \approx 3.5$ if enough loci are capable of responding to particular environmental changes. Thus the statistical aspects of the data appear to fit a model of evolution by natural selection even better than they fit a constant-rate neutral model. This conclusion is made more forceful by the recent demonstration (Gillespie 1985) that, under neutrality, $R(t)$ should be

$$R(t) = 1 + V(4Nr)4Nu/(1 + t/2N), \quad (42)$$

where $r$ is the intragenic recombination rate, $N$ is the population size, $u$ is the neutral mutation rate, and $V(\cdot)$ is a function described in Hudson (1983b) that varies from one to zero as its argument varies from zero to infinity. It had already been pointed out that for neutrality to be compatible with the observed values of $R(t)$ in a model without intragenic recombination, $4Nu$ would have to be too large, say $1 < 4Nu < 10$, to be compatible with the polymorphism data (Gillespie and Langley 1979; Hudson 1983a). When recombination is introduced, the effect is to lower the value of $R(t)$ and thus make it even more difficult to account for the data under a neutral model. Thus, from the perspective of purely statistical considerations, our model of evolution by natural selection actually seems to fit the data better than does the neutral model.

The neutral allele model could be modified to incorporate variable mutation rates and thus be made more compatible with the data. This has, in fact, been attempted both explicitly (e.g., Wu and Li 1985) and implicitly (e.g., Kimura 1983). In each case the model seems to be the same: the mutation rates for each of the lineages of the star phylogeny are chosen independently and remain fixed from the point of origin of the lineage until the present. For this model, the variance-to-mean ratio of the number of substitutions per lineage, $R(t)$, is a good measure of the variability of the rates of substitution. However, this model seems quite artificial. Why should the rates of mutation be altered only at the time of origin of the orders of mammals and then remain unaltered for the next 60 Myr even though other lineages are branching off to form families, genera, and species? A more realistic model would have the neutral mutation rates varying through time in a single lineage to the same extent that they vary between lineages. Such a model has not yet been analyzed. We can speculate that it may well result in a prediction of episodic mutation rates whose range would be outside the ranges that are commonly observed in laboratories. One conclusion is clear: the neutral allele theory cannot, at present, account for the statistical patterns in the sequence data without the use of a very artificial model of rate variation. While the theory could be modified to be in better agreement with the data, such modifications will make the theory less compelling as a uniquely parsimonious explanation for molecular evolution.

Another generalization of the neutral model that could elevate the value of $R(t)$ is Ohta's (1976) theory of mildly deleterious alleles. Under this model, the rate of evolution depends on the population size as well as on the mutation rate. Fluctuations
in the population size could potentially increase the value of $R(t)$, although there are no published results that suggest that this theory can elevate the values of $R(t)$ to the level observed in the data.

A basic conclusion of the present paper is that molecular evolution is compatible with an episodic model. To arrive at this conclusion we began with a model with varying rates of evolution. However, other starting points could lead to models without an episodic structure that are nonetheless compatible with the observed values of $R(t)$. One such model would involve interactions of substitutions at different sites. It is well known (e.g., Blaisdell 1985) that there is some correlation in the identities of neighboring bases in DNA sequences. This suggests that the occurrence of a substitution at one site may increase the likelihood of a substitution at a nearby site. If the increase were adequate, this could lead to an elevation in the value of $R(t)$ without the necessity of an episodic model. A common name given to this phenomenon is “compensatory substitutions.”

A second class of models without an episodic structure would be models in which the rate of change of the molecular clock occurs on a time scale similar to that of the length of the lineages. While this assumption does not alter the conclusion that the variance in the rates of evolution is very much larger than previously thought (Gillespie 1984a), it does not require an episodic process. A complication of this model involves a fundamental issue as to whether the rates of evolution that remain nearly constant for long periods of time are better represented as random quantities chosen from a common probability distribution or as fixed quantities that differ from one part of a lineage to another. This is similar to the distinction between fixed-effects and random-effects models in the analysis of variance.

Clearly, the next step in the development of stochastic models of molecular evolution must be to incorporate all three elements in the model: episodes of evolution, compensatory substitutions, and slow changes in rates. Unfortunately, the current data will probably be inadequate to determine the relative importance of these three elements.

As a final caveat it should be noted that the analysis presented here depends on the fact that the data come from a phylogeny that is closely approximated by a star phylogeny. If this assumption is inappropriate, then the high values of $R(t)$ could be due to the variation in the lengths of the lineages. The analysis of Langley and Fitch (1974) is not subject to this problem and yields an overall estimate of $R(t)$ of $\sim 2.5$. This is similar to the average value of $R(t)$ estimated from the proteins used in this study, suggesting that the assumption of a star phylogeny is fairly accurate. As more sequences become available, attention should turn to the distribution of $M$ and $S^2$ for phylogenies other than star phylogenies.

Acknowledgments

This paper has benefitted from many useful comments from Brian Charlesworth, Walter Fitch, and Michael Turelli.

LITERATURE CITED


WALTER M. FITCH, reviewing editor

Received February 21, 1985; revision received December 13, 1985.
The relative contributions of germline gene variation and somatic mutation to immunoglobulin diversity were studied by comparing germline gene sequences with their rearranged counterparts for the mouse $V_\mu$, $V_\kappa$, and $V_\lambda$ genes. The mutation rate at the amino acid level was estimated to be 7.0% in the first and second complementarity-determining regions (CDRs) and 2.0% in the framework regions (FRs). The difference in the mutation rate at the nucleotide level between the CDRs and FRs was of the same order of magnitude as that for the amino acid level. Analysis of amino acid diversity or nucleotide diversity indicated that the contribution of somatic mutation to immunoglobulin diversity is $\sim 5\%$. However, the contribution of somatic mutation to the number of different amino acid sequences of immunoglobulins is much larger than that estimated by the analysis of amino acid diversity, and more than 90% of the different immunoglobulins seem to be generated by somatic mutation. Examination of the pattern of nucleotide substitution has suggested that clonal selection after somatic mutation may not be as strong as generally believed.

Introduction

Immunoglobulins are composed of heavy and light chains, each of which consists of a variable and a constant region. The variable region is responsible for antigen binding, whereas the constant region is responsible for effector function (see, e.g., Leder 1982; Honjo 1983). The variable region is composed of four framework regions (FRs) and three complementarity-determining regions (CDRs) (see, e.g., Tonegawa 1983). CDRs are the major antigen specificity-determining sites and known to be highly variable. This high variability is generated by both germline gene variation and somatic diversity-generating mechanisms.

Germline gene variation represents an accumulation of germline mutations that occurred in the past and have been stored in an apparently large number of V-region genes. The V-region genes are known to undergo a slow process of concerted evolution, so that a large amount of genetic variability can be maintained (Smith et al. 1971; Hood et al. 1975; Gojobori and Nei 1984). There are two different somatic diversity-generating mechanisms. One is DNA rearrangement of three sets of H-chain ($V_\mu$, $D$, $V_\kappa$).
and \( J_H \) and two sets of L-chain (\( V_\kappa \) and \( J_\kappa \) or \( V_\lambda \) and \( J_\lambda \)) germline gene segments coding for the \( V \) regions. This mechanism generates much immunoglobulin diversity, though it is accomplished largely by utilizing the genetic variation existing in germline genes. Moreover, the joining of various gene segments is not very precise and thus creates new sequence variations at the points of recombination. The other somatic diversity-generating mechanism is somatic (point) mutation, which occurs exclusively in the \( V \) and \( J \) gene segments and their neighboring regions (Kim et al. 1981; Gearhart and Bogenhagen 1983).

It is often argued that germline gene variation and somatic DNA rearrangement alone are sufficient for generating the observed level of antibody diversity (see, e.g., Leder 1982). However, recent data suggest that the rate of somatic mutation is extraordinarily high (see, e.g., Baltimore 1981; Tonegawa 1983), and thus it is possible that somatic mutation plays a major role in generating antibody diversity.

Indeed, Bentley (1984) and Klobeck et al. (1984) have argued that the major contributor to the human \( V_\kappa \) diversity is somatic mutation. However, their arguments are not very convincing because they have not examined the relative contributions of germline gene variation and somatic mutation quantitatively. Previously, Ohta (1978, 1980) analyzed the amino acid diversities of immunoglobulin \( V \) regions within and between species and concluded that "somatic mutation cannot be the major cause of the high variability of CDRs and the observations available can readily be explained by the germline theory." Her conclusion, however, should be reexamined in light of recent data on immunoglobulin DNA sequences, since the region she studied contained the V-D-J joining regions and at that time somatic mutation was not clearly identified.

The purpose of this paper is to examine the relative contributions of germline gene variation and somatic mutation to the \( V \)-region diversity of immunoglobulins by using two different methods. One method used is Nei's (1982) analysis of nucleotide or amino acid diversity, and the other is the examination of the proportion of new amino acid sequences generated by somatic mutation.

**Nucleotide Sequences Used**

Surveying the literature, we have collected data on germline DNA sequences and their rearranged counterparts for the mouse \( V_H, V_\kappa, \) and \( V_\lambda \) genes. The germline sequences used were those obtained from sperm or placentas, and their rearranged sequences were those obtained from myeloma or hybridoma cells. Since somatic mutations occur during B-cell differentiation, i.e., at the time of DNA rearrangement (class switching) or later, they are detected by comparing a rearranged sequence with its original germline sequence.

We made 16 pairwise comparisons of rearranged genes to their germline counterparts. Eight of these comparisons involved six \( V_H \) germline genes, six involved four \( V_\kappa \) germline genes, and two involved two \( V_\lambda \) germline genes (see table 1 for details and references to the sequences used).

In the present study, only the coding regions were examined, including the leader (signal peptide) region, the first three FRs, and the first two CDRs (see fig. 1 of Gojobori and Nei 1984). The third CDR (CDR3) of the \( V_H \) gene consists of the joining region of \( V_H, D, \) and \( J_H \) segments (Kim et al. 1981), and thus it was not included in the present study. The CDR3 of the \( V_\kappa \) and \( V_\lambda \) genes was included, however, because it is coded by the \( V \) segment itself.
<table>
<thead>
<tr>
<th>GENE FAMILY AND GERMLINE GENE (DNA)</th>
<th>REARRANGED GENE (DNA or mRNA)</th>
<th>NO. OF CODONS EXAMINED</th>
<th>Codon Position</th>
<th>All Nucleotides</th>
<th>Amino Acids</th>
<th>REFERENCE(S)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>1st 2d 3d</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>V&lt;sub&gt;H&lt;/sub&gt;:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VH186-2&lt;sup&gt;b&lt;/sup&gt;</td>
<td>S43</td>
<td>116</td>
<td>2.59 2.59 3.45</td>
<td>2.87</td>
<td>6.03</td>
<td>1</td>
</tr>
<tr>
<td>VHSPT15&lt;sup&gt;b&lt;/sup&gt;</td>
<td>M167</td>
<td>119</td>
<td>5.88 2.52 3.36</td>
<td>3.92</td>
<td>7.56</td>
<td>2</td>
</tr>
<tr>
<td>VHSPT15</td>
<td>M603</td>
<td>105</td>
<td>1.90 .95 3.81</td>
<td>2.22</td>
<td>2.86</td>
<td>2</td>
</tr>
<tr>
<td>VH101&lt;sup&gt;b&lt;/sup&gt;</td>
<td>MC101</td>
<td>115</td>
<td>.87 2.61 1.74</td>
<td>1.74</td>
<td>4.35</td>
<td>3</td>
</tr>
<tr>
<td>VH101</td>
<td>S9</td>
<td>101</td>
<td>.99 0 1.98</td>
<td>.99</td>
<td>.99</td>
<td>4</td>
</tr>
<tr>
<td>VHP114&lt;sup&gt;b&lt;/sup&gt;</td>
<td>UPC10</td>
<td>115</td>
<td>1.74 3.84 1.74</td>
<td>2.32</td>
<td>5.22</td>
<td>5</td>
</tr>
<tr>
<td>VHH41&lt;sup&gt;b&lt;/sup&gt;</td>
<td>ID3660&lt;sup&gt;c&lt;/sup&gt;</td>
<td>97&lt;sup&gt;c&lt;/sup&gt;</td>
<td>2.06 0 1.04</td>
<td>1.03</td>
<td>2.06</td>
<td>7</td>
</tr>
<tr>
<td><strong>V&lt;sub&gt;k&lt;/sub&gt;:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V&lt;sub&gt;k&lt;/sub&gt;24</td>
<td>MOPC167</td>
<td>119</td>
<td>1.68 .84 .84</td>
<td>1.12</td>
<td>1.68</td>
<td>8, 9</td>
</tr>
<tr>
<td>V&lt;sub&gt;k&lt;/sub&gt;24</td>
<td>HPCG13</td>
<td>119</td>
<td>.84 0 .84</td>
<td>.56</td>
<td>.84</td>
<td>9</td>
</tr>
<tr>
<td>V&lt;sub&gt;k&lt;/sub&gt;24</td>
<td>M511</td>
<td>119</td>
<td>0 0 .84</td>
<td>.28</td>
<td>.84</td>
<td>9</td>
</tr>
<tr>
<td>V&lt;sub&gt;k&lt;/sub&gt;16</td>
<td>T1</td>
<td>114</td>
<td>1.75 1.75 1.75</td>
<td>1.75</td>
<td>3.51</td>
<td>10</td>
</tr>
<tr>
<td>V&lt;sub&gt;k&lt;/sub&gt;27</td>
<td>T2</td>
<td>114</td>
<td>1.75 .88 2.63</td>
<td>1.75</td>
<td>4.39</td>
<td>10</td>
</tr>
<tr>
<td>V&lt;sub&gt;k&lt;/sub&gt;12</td>
<td>MOPC149</td>
<td>114</td>
<td>0 0 0</td>
<td>0</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>V&lt;sub&gt;k&lt;/sub&gt;18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V&lt;sub&gt;k&lt;/sub&gt;12</td>
<td>MOPC315.26</td>
<td>116</td>
<td>.86 1.72 2.59</td>
<td>1.72</td>
<td>3.45</td>
<td>12</td>
</tr>
<tr>
<td>V&lt;sub&gt;k&lt;/sub&gt;11</td>
<td>H2020</td>
<td>115</td>
<td>.87 .87 0</td>
<td>.58</td>
<td>1.74</td>
<td>13</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>1813</td>
<td>1.54 1.32 1.65</td>
<td>1.51</td>
<td>3.09</td>
<td></td>
</tr>
</tbody>
</table>

* The references for the sequences used are as follows: 1, Bothwell et al. (1981); 2, Kim et al. (1981); 3, Kataoka et al. (1982); 4, Yaoita et al. (1983); 5, Sakano et al. (1980); 6, Olo et al. (1981); 7, Near et al. (1984); 8, Genesiful et al. (1981); 9, Gearhart and Bogenhagen (1983); 10, Pech et al. (1981); 11, Nishioka and Leder (1980); 12, Wu et al. (1982); 13, Bernard et al. (1978).

* Denotes six independent pairs of sequences used for data analysis in table 3.

* The number of nucleotides compared at the third position is 96 because of deletion of a nucleotide in the rearranged sequence, ID3660.
Results

Rates of Somatic Mutation

Let us first estimate the rate of somatic mutation for \( V_H \) and \( V_L \) (\( V_\kappa \) and \( V_\lambda \)) genes, comparing the germline sequences with their rearranged counterparts. The results obtained are presented in table 1. The estimate of the rate of somatic mutation is on the average 1.51% per nucleotide site or 3.09% per amino acid site and is roughly the same for all genes examined except for the \( V_\kappa 2 \) gene, where no mutation was observed. This confirms earlier estimates based on a smaller number of genes (see, e.g., Kim et al. 1981; Nei 1983). Since other proteins, such as thymidine kinase and hypoxanthine phosphoribosyl transferase, are known to undergo somatic mutation at a rate of \( 10^{-6} - 10^{-7} \) per locus (Teillaud et al. 1983), the rate for the immunoglobulin V-region genes is extraordinarily high.

Table 1 also gives the mutation rate for each of the three nucleotide positions of codons. It is clear that there is little difference among the three positions. This pattern is quite different from that of nucleotide substitution in long-term evolution. In most genes so far examined, including the immunoglobulin V-region genes, the rate of nucleotide substitution in the evolutionary process is usually lower at the first and second nucleotide positions than at the third position, apparently because the first- and second-position changes are subject to stronger purifying selection (Li et al. 1981; Kimura 1983; Gojobori and Nei 1984). This suggests that somatic mutation occurs at the same rate for the three positions and that if there is somatic selection, there is no difference among the three nucleotide positions.

Since the CDRs are directly involved in antigen recognition, we studied the rates of somatic mutation in the CDRs and the other regions separately. Interestingly, the rate of somatic mutation in the CDRs is much higher than that in the remaining regions. As shown in table 2, the rate of somatic mutation in the CDRs is nearly three times higher than that of the other regions (also see fig. 1), the difference between the two regions being statistically significant at both the nucleotide and amino acid levels. It is also noted that the rate of somatic mutation is higher in the CDRs than in the other regions at all the nucleotide positions of codons.

These results clearly show that the rate of somatic mutation is unusually high in the variable region, particularly in the CDRs. This suggests that somatic mutation plays an important role in generating immunoglobulin diversity.

Contributions of Germline Gene Variation and Somatic Mutation to Antibody Diversity

1. Amino Acid Diversity and Nucleotide Diversity

Nei and Li (1979) and Nei (1983) suggested that the extent of DNA sequence variation be measured by nucleotide diversity defined by

\[
d = \sum_{i<j} d_{ij}/n_c,
\]

where \( d_{ij} \) is the proportion of different nucleotides between the \( i \)th and \( j \)th sequences and \( n_c \) is the total number of comparisons available. If there are \( n \) sequences, \( n_c = n(n-1)/2 \). Note that \( d \) is equal to the heterozygosity at the nucleotide level. The variance of \( d \) is given by
Table 2
Rates (%) of Somatic Mutation per Nucleotide or Amino Acid Site in Complementarity-determining Regions (CDRs) and Other (framework and signal peptides) Regions

<table>
<thead>
<tr>
<th>GENE AND GENE REGION</th>
<th>NO. OF CODONS EXAMINED</th>
<th>Codon Position</th>
<th>All Nucleotides</th>
<th>Amino Acids</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1st</td>
<td>2d</td>
<td>3d</td>
</tr>
<tr>
<td><strong>V_h:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDRs</td>
<td>176</td>
<td>3.98</td>
<td>5.11</td>
<td>5.11</td>
</tr>
<tr>
<td>Others</td>
<td>707</td>
<td>1.70</td>
<td>1.13</td>
<td>1.42</td>
</tr>
<tr>
<td><strong>V_k:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDRs</td>
<td>165</td>
<td>1.21</td>
<td>.61</td>
<td>2.42</td>
</tr>
<tr>
<td>Others</td>
<td>534</td>
<td>.94</td>
<td>.56</td>
<td>.75</td>
</tr>
<tr>
<td><strong>V_x:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDRs</td>
<td>57</td>
<td>1.75</td>
<td>5.26</td>
<td>1.75</td>
</tr>
<tr>
<td>Others</td>
<td>174</td>
<td>.57</td>
<td>0</td>
<td>1.15</td>
</tr>
<tr>
<td>Total:</td>
<td>398</td>
<td>2.51</td>
<td>3.27</td>
<td>3.52</td>
</tr>
<tr>
<td>CDRs</td>
<td>1,415</td>
<td>1.27*</td>
<td>0.78*</td>
<td>1.13**</td>
</tr>
<tr>
<td>Others</td>
<td>1,415</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Significantly different from the mutation rate for the CDRs at the 6% level.
** Significantly different at the 0.1% level.

\[
V(d) = 4\left\{ (6 - 4n)\left( \sum_{i<j} d_{ij}^2 \right) + n(n - 2) \sum_{i<j} d_{ij}d_{ik} + n^2\left( \sum_{i<j} d_{ij}^2 \right) \right\} / \{n^5(n - 1)\}
\] (2)

(Nei and Tajima 1981). If there are several independent factors contributing to nucleotide diversity, the total diversity can be partitioned into its components (Nei 1982). This principle can be extended to "amino acid diversity" by applying equation (1) to

Fig. 1.—Distribution of somatic mutations in various regions of the \( V_h \) genes. Each column indicates the number of mutations for a region of 15 nucleotides (five codons). The numbering system for codons is identical with that of Gojobori and Nei (1984).
amino acid sequence data. In the present case, there are two factors, i.e., germline gene variation and somatic mutation. The contribution of the former factor \((d_G)\) can be computed by applying equation (1) to the amino acid sequences inferred from the germline DNA sequences. The contribution of somatic mutation \((d_M)\) is obtained by subtracting \(d_G\) from the total diversity \((d_T)\), which is obtained by applying equation (1) to the rearranged DNA sequences. The relative contribution of somatic mutation is then given by

\[
R_M = d_M/d_T,
\]

whereas the relative contribution of germline gene variation is \(R_G = 1 - R_M\).

We used the above method only for the \(V_H\) genes, because the numbers of DNA sequences for the \(V_K\) and \(V_\lambda\) genes were small. To assure independent comparisons of DNA sequences, we used only one rearranged sequence for each germline gene. These are marked by a superscript b in table 1. Furthermore, the signal peptide region was excluded from this analysis because this peptide is excised after immunoglobulins are secreted from plasma cells. The estimates of \(d_T, d_G, d_M,\) and \(R_M\) are presented in table 3. It is seen that amino acid diversity is very high even in germline genes, the \(d_G\) for CDRs being 0.676. That is, \(-68%\) of amino acid sites in the CDRs seem to be different between two randomly chosen sequences. Somatic mutation increases amino acid diversity to 0.707 for the CDRs, the increment being \(d_M = 0.031\). Thus, the contribution of somatic mutation \((R_M)\) is 4.4%. Therefore, if we use amino acid diversity as a criterion, the contribution of somatic mutation to immunoglobulin diversity is very small despite the high mutation rate. This conclusion is similar to Ohta's (1978, 1980), though she had not evaluated the relative contributions quantitatively. It is noted that the amino acid diversity in the FRs is considerably lower than that for the CDRs, but since the rate of somatic mutation is also lower in the FRs, \(R_M\) is nearly the same for both CDRs and FRs.

Nucleotide diversity was computed by using equation (1) for the three nucleotide positions of codons separately. The results are presented in table 4. In both CDRs and FRs, \(d_G\) is largest in the third position and lowest in the second position. This pattern is similar to that of the hemagglutinin gene of the influenza A virus (Nei 1983) and is apparently caused by the difference in the intensity of purifying selection among the three different positions. In the CDRs, however, somatic mutation increases the

<table>
<thead>
<tr>
<th>Table 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Amino Acid Diversities in the (V_H) Regions before and after Somatic Mutation</strong> *</td>
</tr>
<tr>
<td>REGION</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>CDRs</td>
</tr>
<tr>
<td>FRs</td>
</tr>
<tr>
<td>CDRs and FRs</td>
</tr>
</tbody>
</table>

* Based on the 15 possible pairwise comparisons of six sequences for both rearranged and germline genes (those denoted by a superscript b in table 1).

b SE of \(d_T\) and \(d_G\) are computed by using eq. (2). For the computation of the variance of \(d_M\), the covariance of \(d_T\) and \(d_G-Cov(d_T, d_G)\) is needed. Since mutations occur independently of germline variation, we assumed \(Cov(d_T, d_G) = V(d_G)\).
Table 4
Nucleotide Diversities in the V<sub>H</sub> Region Genes before and after Somatic Mutation*

<table>
<thead>
<tr>
<th>REGION</th>
<th>After Somatic Mutation ((d_T))</th>
<th>Of Germline DNAs ((d_G))</th>
<th>Due to Somatic Mutation ((d_M = d_T - d_G))</th>
<th>(R_M = \frac{d_M}{d_T} \times 100)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDRs:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st</td>
<td>0.500 ± 0.054</td>
<td>0.500 ± 0.055</td>
<td>0.0 ± 0.009</td>
<td>0.014 ± 0.015</td>
</tr>
<tr>
<td>2d</td>
<td>0.498 ± 0.067</td>
<td>0.398 ± 0.066</td>
<td>0.050 ± 0.013</td>
<td>0.011 ± 0.013</td>
</tr>
<tr>
<td>3d</td>
<td>0.579 ± 0.076</td>
<td>0.582 ± 0.079</td>
<td>-0.003 ± 0.020</td>
<td>-0.003 ± 0.020</td>
</tr>
<tr>
<td>All</td>
<td>0.509 ± 0.059</td>
<td>0.493 ± 0.058</td>
<td>0.016 ± 0.009</td>
<td>0.016 ± 0.009</td>
</tr>
<tr>
<td>FRs:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st</td>
<td>0.357 ± 0.049</td>
<td>0.343 ± 0.046</td>
<td>0.014 ± 0.015</td>
<td>0.014 ± 0.015</td>
</tr>
<tr>
<td>2d</td>
<td>0.233 ± 0.034</td>
<td>0.234 ± 0.034</td>
<td>-0.001 ± 0.005</td>
<td>-0.001 ± 0.005</td>
</tr>
<tr>
<td>3d</td>
<td>0.463 ± 0.059</td>
<td>0.461 ± 0.058</td>
<td>0.002 ± 0.014</td>
<td>0.002 ± 0.014</td>
</tr>
<tr>
<td>All</td>
<td>0.351 ± 0.045</td>
<td>0.346 ± 0.044</td>
<td>0.005 ± 0.009</td>
<td>0.005 ± 0.009</td>
</tr>
<tr>
<td>CDRs and FRs:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st</td>
<td>0.390 ± 0.049</td>
<td>0.377 ± 0.047</td>
<td>0.013 ± 0.015</td>
<td>0.013 ± 0.015</td>
</tr>
<tr>
<td>2d</td>
<td>0.280 ± 0.039</td>
<td>0.270 ± 0.036</td>
<td>0.010 ± 0.016</td>
<td>0.010 ± 0.016</td>
</tr>
<tr>
<td>3d</td>
<td>0.489 ± 0.060</td>
<td>0.487 ± 0.059</td>
<td>0.002 ± 0.005</td>
<td>0.002 ± 0.005</td>
</tr>
<tr>
<td>All</td>
<td>0.386 ± 0.047</td>
<td>0.378 ± 0.046</td>
<td>0.008 ± 0.010</td>
<td>0.008 ± 0.010</td>
</tr>
</tbody>
</table>

*For explanation see the Note to table 3.

second-position nucleotide diversity substantially, so that \(R_M\) is 11%. This increment is clearly responsible for the increase of amino-acid diversity by somatic mutation. We note that all nucleotide substitutions at the second position result in amino acid changes, whereas at the third position the proportion of nucleotide substitutions resulting in amino acid changes is ~28%. At the first position, the proportion is ~95%. In the first and third positions, \(R_M\) is zero or negative. This peculiar result has occurred because new mutations sometimes cancel the existing variation because there are only four different nucleotides. The cancellation of existing variation is observed when \(d_G\) is already high and the number of nucleotides examined is relatively small. Note that the maximum possible value of expected nucleotide diversity is not 1 but 0.75. In the FRs, \(R_M\) is higher in the first position than in the second or third position.

2. Proportion of V-Region Sequences Generated by Somatic Mutation

It is now clear that the amino acid or nucleotide diversity for immunoglobulins is very high and that a large proportion of this diversity is caused by germline gene variation. However, there is a problem in using this measure as a criterion of immunoglobulin diversity. The problem arises because the number of copies of V-region genes is limited. For example, the number of copies of V<sub>H</sub> genes in the mouse seems to be about 200 (Kemp et al. 1981). Therefore, if we assume that all 200 loci are heterozygous in a diploid organism, there will be 400 different V<sub>H</sub> germline gene sequences available for producing immunoglobulins. Although this is not a large number, the number of different immunoglobulins certainly increases substantially if we consider the D and J gene segments as well as the V<sub>k</sub> and V<sub>1</sub> genes (see, e.g., Leder 1982; Honjo 1983). Moreover, if somatic mutation occurs at the CDRs, the number of different V<sub>H</sub> chains produced becomes very large. We have estimated that the somatic
mutation rate at the amino acid level is 7.0% in the CDRs (table 2). Since there are, on the average, 25 amino acids in the CDRs, each sequence has an average of 1.8 mutations in these regions. Therefore, if we assume that every amino acid replacement in the CDRs changes antigen specificity (Teillaud et al. 1983), somatic mutation is expected to generate many different immunoglobulins.

To determine the contribution of somatic mutation to the total number of immunoglobulins produced, let us compute the expected proportion of mutant sequences generated. It is given by

$$P_M = 1 - \prod_{i=1}^{n} (1 - \mu_i) = 1 - e^{-\mu n},$$

where $\mu_i$ is the mutation rate at the $i$-th site, $\mu$ is the average mutation rate per site, and $n$ is the number of amino acids involved in the region concerned. In the present case, $n = 25$ and $\mu = 0.070$, so that $P_M = 0.83$. That is, 83% of V regions are expected to have mutant sequences in the CDRs. This is slightly lower than the observed value (15/16 = 0.94) for the genes in table 1. If we consider both the heavy and light chains of immunoglobulins, the proportion of immunoglobulins having somatic mutations in the CDRs will be $1 - (1 - 0.83)^2 = 0.97$. In other words, the majority of immunoglobulins produced is expected to have one or more mutant amino acids in the CDRs. Thus, if a single amino acid in the CDRs changes antigen specificity, somatic mutation will play a very important role in generating antibody diversity. If the FRs are also involved in determining antigen specificity, the contribution of somatic mutation would be even greater.

Pattern of Somatic Mutation

It is known that in most eukaryotic genes the nucleotide substitution in the evolutionary process does not occur at random among the four nucleotides adenine (A), thymine (T), cytosine (C), and guanine (G) (Fitch 1967; Vogel and Kopun 1977; Gojobori et al. 1982). This nonrandom substitution occurs partly because purifying selection eliminates certain types of mutation and partly because mutation itself occurs nonrandomly (Gojobori et al. 1982). It is, therefore, interesting to see whether or not somatic mutations follow the same rule as that of usual spontaneous mutation. We have, therefore, studied the pattern of somatic mutation, examining the relative frequencies ($f_{ij}$) of nucleotide substitution (Gojobori et al. 1982).

Let $n_{ij}$ be the observed number of substitutions from nucleotide $i$ to nucleotide $j$ ($i, j = A, T, C, \text{ or } G$). The relative substitution frequency from nucleotide $i$ to nucleotide $j$ is then defined as

$$f_{ij} = n_{ij} / \sum_{i \neq j} n_{ij}. \quad (5)$$

In the present case, $n_{ij}$ can be determined unambiguously because the nucleotide sequences before and after mutation are known. In the previous studies, the $n_{ij}$ was always inferred under certain assumptions. At any rate, using a total of 82 mutations observed from comparison of germline and rearranged sequences, we computed $f_{ij}$'s, and the results obtained are presented in table 5.

Table 5 shows that the substitution frequency varies considerably with nucleotide pair. In the case of "All codon positions," the highest frequency (G → A) is about
Table 5
Relative Frequencies (%) of Nucleotide Substitution Due to Somatic Mutation*

<table>
<thead>
<tr>
<th>ORIGINAL NUCLEOTIDE</th>
<th>Mutated Nucleotide</th>
<th>A</th>
<th>T</th>
<th>C</th>
<th>G</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>8.5</td>
<td>9.8</td>
<td>13.4</td>
<td></td>
<td>31.7</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>6.9</td>
<td>10.4</td>
<td>2.3</td>
<td></td>
<td>19.6</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>5.2</td>
<td>5.2</td>
<td></td>
<td>6.5</td>
<td></td>
<td>16.9</td>
</tr>
<tr>
<td>G</td>
<td>19.6</td>
<td>2.4</td>
<td>9.8</td>
<td></td>
<td>31.8</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>31.7</td>
<td>16.1</td>
<td>30.0</td>
<td>22.2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ORIGINAL NUCLEOTIDE</th>
<th>Mutated Nucleotide</th>
<th>A</th>
<th>T</th>
<th>C</th>
<th>G</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>12.4</td>
<td>9.0</td>
<td>9.0</td>
<td></td>
<td>30.4</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>3.8</td>
<td>10.2</td>
<td>2.5</td>
<td></td>
<td>16.5</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>5.8</td>
<td>8.7</td>
<td></td>
<td>4.4</td>
<td></td>
<td>18.9</td>
</tr>
<tr>
<td>G</td>
<td>15.3</td>
<td>11.8</td>
<td>7.1</td>
<td></td>
<td>34.2</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>24.9</td>
<td>32.9</td>
<td>26.3</td>
<td>15.9</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Total frequencies of transitional mutation are 48.0% and 43.2% for the upper and lower matrices, respectively.

**Nucleotide substitution inferred from the amino acid sequences (data from Crews et al. 1981) are included.

nine times as high as the lowest one (T → G). Note that transitional mutations (mutations between A and G and between T and C) are about 1.5 times the frequency expected (0.33) under random substitution. Furthermore, the pattern of nucleotide substitution is virtually the same for all nucleotide positions of codons. (Data for the third position are not shown). Gojobori et al.'s (1982) method of correlation analysis also showed that, unlike the usual germline gene mutations, there is no tendency for the rate of nucleotide substitution from somatic mutation to be affected by the chemical property of the amino acids involved in the substitution. These results suggest that purifying selection, which is generally observed in functional germline genes, is not operating for somatic mutations.

Kaartinen et al. (1983) proposed the hypothesis that, in V-region genes, natural selection at both individual and cellular levels favors amino acid–altering nucleotide substitution over silent substitution. At the individual level, this hypothesis is not supported because the rate of nucleotide substitution in long-term evolution is higher at the third position than at the first and second positions (Gojobori and Nei 1984). Data on somatic mutation also fail to support this hypothesis. Our data show that the proportion of silent somatic mutations in 29%, which is very close to the theoretical value of 30% expected from the pseudogene pattern of spontaneous mutation (Gojobori et al. 1982). That is, it seems that there is not very strong somatic selection after somatic mutation occurs.
Discussion

We have examined the relative contributions of germline gene variation and somatic mutation to immunoglobulin diversity by using two different methods. The conclusions obtained by the two methods are quite different. The reason for this is that in the first method (analysis of amino acid or nucleotide diversity) the extent of amino acid or nucleotide differences between different V-region sequences is considered, whereas in the second method two sequences that differ by a single amino acid are considered to recognize different antigens. In practice, there is evidence showing that a single amino acid replacement can change antigen specificity (Teillaud et al. 1983; Knossow et al. 1984). Therefore, the conclusion obtained by the second method seems to be more realistic than that obtained by the first method. However, a small difference in the amino acid sequence in the CDRs may not always be effective in distinguishing between different antigens, so that the contribution of somatic mutation to antibody diversity may not be as great as our second method shows. It should also be noted that somatic mutation apparently occurs during or after DNA rearrangement (Tonegawa 1983) and that a certain proportion of somatic mutations may not contribute effectively to antigen specification. Furthermore, since mutations occur at random, some of them may not be really useful for determining antigen specificity. Nevertheless, the contribution of somatic mutation seems to be much more important than the analysis of amino acid diversity shows.

Somatic mutation seems to be correlated with the class of the antibody. Most somatic mutations are found in the V regions derived from antibodies of the IgA and IgG classes (Crews et al. 1981). The IgM molecules express the rearranged V gene identical to the germline V sequence exclusively, with a few exceptions (Rudikoff et al. 1984). Thus, if the IgM molecules are produced in quantity throughout the ontogeny of a B-cell line committed to antibody production, the value of $P_M$ obtained by the second method may be overestimated.

It has been shown that somatic mutation occurs at the J gene segments as well (Gearhart and Bogenhagen 1983). If a high rate of somatic mutation also occurs at the entire CDR3 region, the contribution of somatic mutation to immunoglobulin diversity will be even higher than our previous estimate. It should also be noted that the joining of V and J or of V, D, and J gene segments is known to be imprecise, and this imprecise joining also causes amino acid sequence variation.

The possible number of immunoglobulins produced is often computed by multiplying the numbers of V, D, and J gene segments in the genome. Thus, if there are $2 V_\lambda$, $3 J_\lambda$, $300 V_\kappa$, and $4 J_\kappa$ segments in the genome, the maximum number of different light chain V regions that can be produced by DNA rearrangement is $1,206 (-2 \times 3) + [300 \times 4])$. Similarly, if there are $200 V_H$, $12 D$, and $4 J_H$ segments, the maximum number of different $V_H$ regions would be $9.6 \times 10^3 (=200 \times 12 \times 4)$. If the light and heavy chain V regions are combined at random, the total possible number of immunoglobulins is $1.2 \times 10^7$. This number further increases if a majority of gene segment loci carry different alleles in heterozygous condition. The variability caused by imprecise joining of V, D, and J gene segments would also increase the number by two or three times. Therefore, the total number of different immunoglobulins would be about $3 \times 10^7$ without any contribution from somatic mutation. If 90% of sequence variation in the V regions of immunoglobulins is caused by somatic mutation, the total number of different immunoglobulins that can be produced will be enormous and more than sufficient for producing different antibodies that are required in higher vertebrates.
Acknowledgments

We thank Tommy Douglas, Takashi Miyata, and Shozo Yokoyama for their comments on the manuscript. This study was supported by research grants from the National Institutes of Health and the National Science Foundation to M.N. and by a grant in-aid from the Ministry of Education, Science, and Culture of Japan to T.G.

LITERATURE CITED


WALTER M. FITCH, reviewing editor

Received August 28, 1985; revision received October 17, 1985.
Variation in the Number of α-Globin Loci in Sheep

A. Rando, L. Ramunno, and P. Masina
Facoltà di Agraria, Università di Napoli

Southern blot analysis was used to compare sheep and goat restriction-endonuclease maps of the DNA region containing the α-globin genes. The identical digestion patterns observed in both species with three endonucleases (BamHI, BstEII, and PstI) show that in sheep a single chromosome normally bears two nonallelic α-globin genes positioned at the same distance as in goat. Variant digestion patterns with enzymes that cleave outside (BamHI and HindIII) and within (EcoRI) the α-globin loci allowed us to infer that chromosomes with different numbers of α-globin loci are also present in sheep. In particular, in the 60 sheep considered, four individuals were heterozygous (aa/aaa) and one was homozygous (aaa/aaa) for chromosomes with three loci and one individual was heterozygous for a chromosome with four loci (aa/aaaa). This variation in the number of copies of α-globin loci can be explained by means of unequal crossovers.

Introduction

Closely related pairs of nonallelic α-globin genes are found not only in several mammalian species but also in chickens. This implies that the loci coding for the adult α-chains may have been in the duplicate state prior to the mammalian radiation (~300 Myr ago) (Zimmer et al. 1980). Concerted evolution has been proposed to explain the extremely high similarity that exists between the duplicated α-globin genes in man (Liebhaber et al. 1981; Michelson and Orkin 1983), apes (Zimmer et al. 1980), and goat (Schon et al. 1982).

Until 1980 only one type of α-chain was detected in sheep; therefore, there was no evidence supporting the hypothesis of the duplication of α-globin loci in this species. Careful analyses at the protein level (Vestri et al. 1980, 1983a, 1983b; Vestri and Salmaso 1981) have shown that the α-globin locus is also duplicated in sheep. Three alleles occur in each of the 1α and 2α loci. According to Vestri et al.'s nomenclature, the alleles are referred to as α113Leu, 1α8Ala,113Leu, 1α15Asp,113Leu, 2α113Leu, 2α113His, and 2α113Hist. The 1α product is approximately 1.8 times more abundant than the 2α one. The 1α113Leu and the 2α113Leu alleles produce identical chains. The 2α113Hist and the 2α113Hist alleles produce identical chains, but the former is less active (~30%) than the latter. Finally the 1α15Asp,113Leu allele (referred to as αD by Huisman et al. 1968) is found only in cis phase with the 2α113Hist allele. Goats, which are closely related to sheep, also have two nonallelic α-globin loci, 1α and 2α, which produce two α-chains, differing at four amino acid positions, in a quantitative ratio of 3:1 (Adams et al. 1969). The sheep 1α113Leu chain differs from the goat 1α-chain at only one position (Thr 104 in the sheep vs. Ser 104 in the goat) (Wilson et al. 1968).

Key words: restriction endonuclease map, DNA polymorphism, unequal crossover, concerted evolution.
The approach pioneered by Southern (1975) to study specific sequences in DNA has been applied to provide direct physical evidence for \( \alpha \)-globin gene duplication and for the close linkage of the two loci within a small region of DNA in man (Orkin 1978), apes (Zimmer et al. 1980), and goat (Schon et al. 1982). Furthermore, chromosomes carrying either one or three \( \alpha \)-globin loci have been detected in man (Embry et al. 1979, 1980; Orkin et al. 1979; Goossens et al. 1980; Higgs et al. 1980; Lie-Injo et al. 1981) and in chimpanzee (Zimmer et al. 1980).

In this paper we compare sheep and goat restriction maps of the DNA region containing the \( \alpha \)-globin genes. As in goat, a single sheep chromosome normally bears two \( \alpha \)-globin loci, but in the 60 animals considered in the present study we also found six chromosomes with three and one with four \( \alpha \)-globin loci.

Material and Methods

By using EDTA as an anticoagulant, blood samples were collected from 20 goats (Capra hircus) and 15 sheep (Ovis aries) of a local undefined genetic type, from 20 purebreed Ile de France, and from 25 purebreed Apulian Merino sheep. Restriction endonucleases, listed in the legend of figure 2, were purchased from Amersham, (England) and Boehringer Mannheim. The Nick-translation kit and \( \alpha \)-\(^{32}\)PdCTP (400 Ci/mmol) were purchased from Amersham, England and used according to the supplier's directions to obtain a specific activity of \( \sim 2 \times 10^8 \) cpm/\( \mu \)g of DNA.

Plasmid pGa2 (Schon et al. 1982) containing the goat chromosomal \( \beta \)-globin gene was a gift from Dr. J. B. Lingrel. Plasmid JW101 (Wilson et al. 1978), containing the human \( \alpha \)-globin cDNA, was a gift from Prof. L. Felicetti. The whole plasmids were used as probes.

Cellular DNA was obtained from leukocytes according to the method of Goossens and Kan (1981). Digestion with restriction endonucleases, electrophoresis of DNA in 0.8% agarose gels, transfer of DNA onto nitrocellulose filters, prehybridization, hybridization, washing of filters, and autoradiography were carried out as previously described (Masina et al. 1984). Plasmid JW101 was hybridized in less-stringent conditions (see Rando and Masina 1985). HindIII and PstI single digests of \( \lambda \)DNA were used as molecular size markers.

Isoelectric focusing (pH range, 6.0–8.5) of hemolysates was accomplished in an LKB 2117 Multiphor apparatus following the supplier's directions. Densitometric ratios were determined with an LKB Ultrascan densitometer and a Hewlett-Packard 3380A integrator. Hematologic analyses were accomplished according to the classical methods.

Results

During our survey on restriction fragment-length polymorphisms (RFLPs) in \( \alpha \)- and \( \beta \)-globin DNA regions in farm animals (Masina et al. 1984; Rando and Masina 1985), we digested 20 goat DNA samples with six endonucleases—BamHI, BstEII, EcoRI, HindIII, KpnI, and PstI—and probed them with plasmid pGa2. The restriction patterns obtained with these endonucleases are shown in figure 1. No individual differences appeared in the samples considered. Double digestions and the restriction map presented by Schon et al. (1982) allowed us to locate the cleavage sites of these enzymes (fig. 2).

The same probe and the same endonucleases were also used to test 60 sheep DNA samples. Among these, 20 DNA samples obtained from the Ile de France breed, 24 from the Apulian Merino breed, and 10 from the undefined genetic type showed
FIG. 1.— Autoradiograms of restriction enzyme–digested goat (G) and sheep (S) DNAs hybridized with plasmid pGnu2. The numbers indicate fragment lengths in kilobase pairs. The 0.6-kb KpnI band and the 0.4-kb PstI band were detected on autoradiography with longer exposure.
Fig. 2.—Diagrams showing goat and sheep restriction maps of the α-globin genes' DNA region. B = BamHI; Bs = BsiEII; E = EcoRI; H = HindIII; K = KpnI; P = PstI. The length of each duplicated region is 3.6 kb. Asterisks indicate boundaries of the DNA region in which crossovers occurred.
the restriction patterns presented in figure 1. It can be seen that \textit{BamHI}, \textit{BstEII}, and \textit{PstI} show the same digestion pattern in goat as in sheep; \textit{EcoRI} shows one identical fragment (1.5 kb) and a shorter one (8.7 kb rather than 9.3 kb); \textit{HindIII} shows a 13.5-kb fragment whose length corresponds to the sum of the two fragments determined in goat (11.6 kb and 1.9 kb); \textit{KpnI} shows three fragments (9.8 kb, 3.6 kb, and 0.6 kb) whose lengths total the single fragment determined in goat (\(\sim\)14.0 kb). Fragment length determination, however, does not exclude the possibility of small differences that could be the result of insertion/deletion of short DNA fragments.

To obtain the restriction map of the sheep \(\alpha\)-globin region (see fig. 2), we performed double-digestion analyses. The most important differences between the two species are the lack of the intergenic \textit{HindIII} site and the presence of the two corresponding intragenic \textit{KpnI} sites in sheep. Double digestions allowed us to locate the intragenic \textit{KpnI} sites \(\sim 200\) bp upstream of the identically positioned intragenic \textit{BstEII} sites. Therefore, the 3.6-kb \textit{KpnI} fragment, like the one of identical length produced by \textit{BstEII}, represents the bridging fragment containing the intergenic region plus the 3' end of the \(\alpha\) locus and the 5' end of the \(\beta\alpha\) locus. These results provide direct evidence that in these 54 sheep a single chromosome bears two nonallelic \(\alpha\)-globin genes, positioned at the same distance as in goat.

Five DNA samples, all obtained from the sheep belonging to the undefined genetic type, showed normal digestion patterns with \textit{BstEII}, \textit{KpnI}, and \textit{PstI} but different patterns with \textit{BamHI}, \textit{EcoRI}, and \textit{HindIII} (fig. 3). Four of them, digested with \textit{BamHI} and \textit{HindIII}, showed, besides the normal fragments (6.0 kb and 13.5 kb, respectively), a fragment 3.6 kb longer (9.6 kb and 17.1 kb, respectively). This longer fragment was always denser than the normal one. One sample showed only the 9.6-kb and the 17.1-kb variant fragments. All of these five DNA samples, when digested with \textit{EcoRI}, showed an extra fragment 3.6 kb long that was denser in the individual showing only the variant fragments. This extra fragment has the same length as the bridging fragment and is visible even after hybridization of these samples with plasmid JW101 containing the human \(\alpha\)-globin cDNA (not shown). Therefore, this extra fragment contains coding sequences of the \(\alpha\)-globin gene.

Since digestion with enzymes that cleave outside the \(\alpha\)-globin loci (\textit{BamHI} and \textit{HindIII}) produces a fragment \(\sim 3.6\) kb longer than normal and since digestion with the enzyme that cleaves within the two \(\alpha\)-globin loci (\textit{EcoRI}) produces an extra 3.6-kb fragment, we infer, according to Goossens et al. (1980), Higgs et al. (1980), and Lie-Injo et al. (1981), that these five individuals carry chromosomes with three \(\alpha\)-globin loci (fig. 2). In particular, the four individuals showing both the normal and the variant patterns would be heterozygotes, having one chromosome with two loci and the other one with three loci (\(\alpha\alpha/\alpha\alpha\alpha\)), whereas the individual showing only the variant pattern would be homozygous, having three loci (\(\alpha\alpha\alpha/\alpha\alpha\alpha\)) on both chromosomes.

This situation is the same as that described in man by Goossens et al. (1980, Higgs et al. (1980), and Lie-Injo et al. (1981) and in chimpanzee by Zimmer et al. (1980). Therefore, the mechanism originating the triple-loci chromosome can be explained in the same way, that is, by hypothesizing an unequal-crossover phenomenon between two normal chromosomes.

Blood samples of 15 sheep of undefined genetic type were collected at the local slaughterhouse. Unfortunately, it was impossible to determine hematologic findings and hemoglobin isoelectric-focusing patterns. We do not have family data, but, con-
Fig. 3.—Autoradiograms of restriction enzyme-digested sheep DNA hybridized with plasmid pGα2 and showing the four phenotypes determined by different numbers of α-globin loci as follows: 1, αα/αα; 2, αα/ααα; 3, ααα/ααα; 4, αα/αααα.
considering that the triple-α-loci chromosome was found only in this group—and that it occurred there at high (20%) frequency—we suppose that these individuals were genetically related.

One DNA sample, belonging to a 3-year-old vasectomized male of the Apulian Merino breed, showed the normal fragments and a denser one 7.2 kb (2 × 3.6 kb) longer (i.e., fragments ~13.2 kb and ~20.7 kb, respectively) (fig. 3) when digested with BamHI and HindIII. This individual showed normal patterns with BstEII, KpnI, and PstI and the same EcoRI pattern as the homozygous carrier of triple-α-loci chromosomes. In this case, we infer that this individual carries one normal chromosome and one with four α-globin loci (aa/aaaaa). This chromosome (fig. 2) could be derived by unequal crossover either between a triple-loci chromosome and a normal one or between two triple-loci chromosomes. The former event would also create a single-locus chromosome, whereas the latter one would also create a normal two-loci chromosome.

The fact that the observed different patterns reflect differences in the number of α-globin loci is also confirmed by the differences in the intensity of some bands obtained with enzymes (i.e., BstEII, KpnI, and PstI) showing invariant fragments. In particular, as would be expected given the restriction maps presented in figure 2, the individuals with five or six α-globin loci per diploid cell show the BstEII and KpnI bridging fragments and two of the four PstI fragments (i.e., those of 1.0 and 0.4 kb) to be always denser than the normal ones.

Compared to seven other vasectomized males, the animal with the chromosome carrying four loci had slightly higher hemoglobin and hematocrit values, but the data are too variable to infer that the extra α-globin genes are functional. Isoelectric focusing of the hemolysate of this sheep shows two major fractions (amounting to 63%) with the mobility of HbB and HbA and two minor ones (amounting to 37%) with the mobility of HbD. According to Tucker (1981) and Vestri et al. (1983a), this result suggests that this individual is heterozygous at the β-globin locus (βA/βB) and heterozygous also at the 1α-globin locus (1α113His/αD). Even though we have not found hemoglobin fractions containing α113His chains, it is not possible to exclude their presence in small quantities (as is the case in the 12–15:1 phenotype described by Vestri et al. [1983a]). We subjected to isoelectric focusing the 44 (20 Ile de France and 24 Apulian Merino) other hemoglobin samples, all belonging to individuals carrying chromosomes with two loci. None of them contained HbD. Therefore, the unique sheep that produces αD-chains also carries the chromosome with four α-globin loci. Since the αD allele and the four-loci chromosome have a low (0.02) frequency in the Apulian Merino breed, the probability of finding an αDα/aaaaa individual is extremely low (1/1,250). It seems more likely, therefore, that the genotype of this sheep is αα/αDaaaa.

Discussion

We have provided evidence at the DNA level that, in sheep as in goats, two nonallelic α-globin loci are normally present. Furthermore, sheep carrying chromosomes with three and four α-globin loci were found. The presence in sheep—as well as in man (Goossens et al. 1980; Higgs et al. 1980; Lie-Injo et al. 1981) and in chimpanzee (Zimmer et al. 1980)—of chromosomes with different numbers of α-globin loci confirms that in this DNA region the occurrence of unequal crossovers is particularly frequent.

The identical restriction patterns obtained with BstEII digestion of sheep and goat DNA show that the two loci are located at the same distance in both species. The
most interesting difference between the restriction maps presented in figure 2 is the presence only in sheep of the intragenic KpnI sites, located in the same position in both genes. Double digestions allowed us to locate these KpnI sites ~200 bp upstream of the identically positioned intragenic BstEII sites. On the basis of the analyses of the complete nucleotide sequence of the two goat a-globin genes, (determined by Schon et al. [1982]), we hypothesize that these KpnI sites are located in the positions corresponding to the amino acids 78 and 79 (Gly and Thr, respectively). In both goat a-globin genes these two amino acids are encoded by the nucleotide sequence GGTACT, which is located 186 bp upstream of the BstEII sites. Since the a-globin chains of both sheep and goat are identical in the 78 and 79 residues (Wilson et al. 1968), the presence of the intragenic KpnI cleavage sites in sheep indicates that these residues are encoded by the nucleotide sequence GGTACC rather than by GGTACT.

Since it is likely that the loci coding for the a-chains were in the duplicated state prior to mammalian radiation (~300 Myr ago), concerted evolution has been proposed to explain the observed high similarity between the two nonallelic a-globin genes of several species (Zimmer et al. 1980). According to this hypothesis the divergence between the two duplicated loci would be prevented by means of cycles of unequal crossovers, which, in an intermediate state, would cause the appearance in the population of chromosomes with one and three a-globin loci (Zimmer et al. 1980). However, in the sheep we considered no chromosome with a single a-globin locus was found. It could be that natural and/or artificial selection act against this chromosome.

Since the intragenic KpnI site is present in both a-globin loci of the sheep species that we studied but is absent from Barbary sheep (Ammotragus lervia) (A. Rando, L. Ramunno, and P. Masina, unpublished observations) and goat, then, according to the monophyletic tree of the Caprini tribe proposed by Bunch and Nadler (1980), the KpnI site probably arose by a single point mutation in sheep after its separation from goat and Ammotragus and before a round of concerted evolution. The following two findings therefore suggest that concerted evolution is more active in sheep than in goat: (1) the greater similarity in the amino acid sequences of the nonallelic a-chains and (2) the presence of triple- and quadruple-a-loci chromosomes, which can be considered an intermediate step in concerted evolution.

Acknowledgments

We thank Prof. J. B. Lingrel and Prof. L. Felicetti for providing recombinant plasmids pGa2 and JW101, respectively, and Prof. L. L. Cavalli-Sforza for critical review of the manuscript. We also thank Dr. D. Pagnini for hematologic analyses and Drs. R. Rubino and F. Grasso for collecting some of the blood samples. This work was supported by the Ministero della Pubblica Istruzione.

LITERATURE CITED


WALTER M. FITCH, reviewing editor

Received July 22, 1985; revision received September 25, 1985.
Book Review


Since the series was founded in 1977, the editors of Isozymes: Current Topics in Biological and Medical Research have provided us with up-to-date reviews of the advances in isozyme research from a wide variety of biological disciplines. In the previous 11 volumes, they have stressed the interdisciplinary nature of isozyme research by selecting reviews that cover a diversity of topics, ranging from the study of the biochemical structure and subcellular localization of isozymes, through the mechanisms of genetic regulation and adaptive significance of isozyme variation, to the use of isozymes in swine breeding and forensic research. This diversity of experimental approaches that utilize isozymes is reflected in the six review articles in the most recent volume in the series.

The practical problem of identifying and defining new cultivars using isozymes is considered in the article by Gunnar Nielsen. Nielsen emphasizes the need for systems of genetic markers to supplement the morphological classification of plant varieties and lists 47 plant species in which studies of protein polymorphisms have revealed sufficient variation to distinguish cultivars within a species. Although the ability of isozyme patterns to define cultivars is clear (for instance, 80% of 300 inbred lines and 94% of 155 hybrid lines of maize have unique isozyme patterns), a potential problem in the award of “breeders rights,” as pointed out by the author, is the maintenance of stable allele frequencies in an outcrossing cultivar as it is propagated and grown in different environments.

An evolutionary perspective on the regulation of enzyme loci and on the adaptive nature of variation in natural populations is given in the reviews by C. C. Laurie-Ahlberg and Ward B. Watt, respectively. Laurie-Ahlberg summarizes the methods for the detection of polymorphism of genes that modify the expression of enzyme-encoding loci and reviews the extensive data on alcohol dehydrogenase, alpha-amylase, and several other enzyme systems in Drosophila. One must be impressed with the wealth of variability in genes that regulate isozyme expression in natural populations of Drosophila, and particularly with the substantial interlocus correlations in enzyme activities revealed by Laurie-Ahlberg’s own experiments with chromosome substitution lines. Watt argues that much information about the adaptive significance of allozymes may be lost in the conventional statistics that summarize genetic variation in natural populations. He reasons that a bioenergetic approach—one providing a detailed understanding of the in vitro differences in enzyme activities, the position of an enzyme in metabolic pathways, and the distribution of environments in which an organism lives—can be used in concert to generate predictions about the relative fitness of specific genotypes. Although this approach represents a substantial improvement over the simple correlation of allozyme frequencies with ecological variables, I was unconvinced of the general predictive power of the bioenergetic approach primarily because of poor development of testable genetic hypotheses. We will have to wait to see what generalizations concerning the mechanisms of biochemical adaptation emerge from future applications of the bioenergetic research program.

The three remaining articles focus on specific isozyme systems. John L. VandeBerg gives a good overview of the comparative studies of the phosphoglycerate kinase (PGK) isozyme system in a number of mammalian species. He documents the experimental evidence indicating that PGK is encoded by a pair of loci, an X-linked gene that is expressed in all somatic cells and an autosomal locus that is expressed only during
spermatogenesis. These loci apparently arose by a gene duplication event predating the divergence of the eutherian and metatherian mammals. VandeBerg concludes that the marked differences in regulation between these two loci make PGK an ideal system for the further study of the evolutionary divergence in function and regulation shown by duplicate genes. William L. Daniel gives a similarly thorough review of the phylogenetic distributions, comparative biochemistry, and genetic regulation of arylsulfatase C and steroid sulfatases. The final article, by Don Mahuran, Anton Novak, and J. Alexander Lowden, provides a detailed account of the molecular structure and practical application of the lysosomal hexosaminidase (HEX) isozymes. The use of hexosaminidase assays for carrier detection and prenatal diagnosis of Tay-Sachs disease is one of the success stories in the clinical application of isozyme research and has inspired much interest in the molecular properties that determine HEX activity.

The text is well formatted and contains few typographical errors; however, I never became comfortable with the nonitalicized species names. Each article includes a generous bibliography, and the literature cited is complete through 1983, with a few examples of more recent publications. There is also a useful index at the end of the volume.

Overall, the editors of the Isozyme series—Mario C. Rattazzi, John G. Scandalios, and Gregory S. Whitt—should be commended for assembling reviews that provide a broad coverage of the versatile use of isozymes in biological and medical research.

THOMAS S. WHITTAM
Pennsylvania State University
Information for Contributors

*Molecular Biology and Evolution* is a bimonthly journal devoted to the interdisciplinary science between molecular biology, and evolutionary biology. The journal emphasizes experimental papers, but theoretical papers are also published if they have a solid biological basis. Although this journal is primarily for original papers, review articles and book reviews normally written by solicited authors are also published. Brief discussion and comment on material published in this journal or on issues particularly relevant to readers of this journal will be published as "Letters to the Editor." Letters that refer to a paper handled by an Associate Editor should be sent to that editor or to the Editor in Chief.

To minimize publication delays, authors should follow the instructions given here and should also provide their telephone numbers.

Submission of Manuscripts

Send manuscripts (one original and two high-quality copies) to the Editor in Chief, Managing Editor, or any Associate Editor (addresses below). Any manuscript or any part of a manuscript which has been published or submitted for publication elsewhere cannot be accepted for publication. Anyone who wishes to write a review article should contact the Managing Editor. Correspondence about book reviews should also be addressed to the Managing Editor. Decision on acceptance of papers will be made as rapidly as possible. Papers that are not suitable to the journal will be returned immediately to authors without detailed review.

After a manuscript is accepted, its author will be requested to sign an agreement transferring copyright to the publisher. No published material may be reproduced or published elsewhere without the written permission of the copyright owner. The journal will not be responsible for the loss of manuscripts at any time.

Publication is taken to imply that the authors are prepared to make available to the public any unpublished sequences on which the paper is based and any clone of cells, DNA, or antibodies used in the experiments reported. This principle also applies to computer programs.

Preparation of Manuscripts

Papers must be written in English and organized in the sequence described below. Each section must be typed double-spaced on heavyweight, nonerasable bond; the page margins should be 1½ inches wide to allow for corrections and manuscript editor's notes and queries. Special typefaces (e.g., italic or sans serif) should not be used, and right-hand margins should not be justified. Word-processing output on dot matrix printers is acceptable only if it is the quality of the standard typewriter. Handwritten items (e.g., Greek letters) must be identified in the margin. Non-English words must have correct diacritics. Although each major part of the paper (e.g., Literature Cited) must begin on a new page, the pages should be numbered consecutively throughout, beginning with the title page and continuing through the abstract, text, Appendix, Literature Cited, footnotes, tables, and ending with figure legends.

**Title page.**—This page should contain the paper's title, the names of all authors, the institution(s) at which the research was done, the current affiliations of all authors, the name and address for correspondence, and a footnote on nonstandard abbreviations used, if any (see below). Finally, the title page should also provide a running head (maximum of 50 characters and spaces).

**Abstract.**—The abstract should be a one-manuscript-page factual condensation of the entire paper, including a statement of purpose, a clear description of observations and findings, and a concise presentation of conclusions. It should not assert that the findings are discussed.

**Key words.**—A list of three to six words or phrases should be provided that will accurately index the subject matter of the article.

**Text.**—The text should comprise the following sections: (1) Introduction, (2) Material and Methods, (3) Results, (4) Discussion, and (5) Acknowledgments (if any). Papers should be concise but will not be restricted in length.
All organisms mentioned must be identified by their scientific binomials and underlined. Symbols for genetic loci must also be underlined and should follow the established rules of genetic nomenclature for the various organisms (consult M. Demerec et al., *Genetics* 54 [1966]:61-76). Include the formal IUB name and number of all enzymes mentioned.

Do not use abbreviations for words or phrases used less than five times. Abbreviations used by the *Journal of Biological Chemistry* will be regarded as standard; nonstandard abbreviations should be defined collectively in a footnote.

Mathematical equations must be carefully typewritten: spacing between characters should be correct as typed. It will be assumed that all characters in equations and their counterparts in the text will be set in italics unless the author specifies otherwise the first time a character appears. Equations should be numbered sequentially, in arabic numerals in parentheses, on the right-hand side of the page.

These and other guidelines can be found in the *Council of Biology Editors Style Manual* (4th ed., 1978). In general, all material should conform to the CBE format. See also recent issues of this journal.

**Terminology.**—A satisfactory interdisciplinary communication requires using words with careful attention to their precise meaning in both disciplines. Authors may use any word they choose provided only that its meaning is clear, consistent, and serves to increase the paper's comprehensibility. The following preferred usages are *not* prescriptive but will be assumed unless authors define them otherwise.

Where the alignments disagree, they are *differences* rather than *changes* since there may have been multiple changes to create a single difference. Differences or changes are *replacements* if the sequences are amino acids, *substitutions* if they are nucleotides. *Mutations* should be restricted to changes before selection has operated. *Homology* must be defined since it has two common meanings: (1) observed similarity and (2) inferred common ancestry. The term *similarity* is preferred for meaning 1 because sequences may have similarity acquired by convergence (analogy) rather than retained after divergence (homology). When homology arises via a gene duplication (all or part), it is properly called *paralogy*; when it arises via speciation, it is properly called *orthology*. *Gaps* are introduced into sequences to increase their similarity rather than to *optimize* similarity (homology), unless an algorithm is employed that guarantees an optimized result according to the way similarity (homology) is defined (e.g., as maximum matches—a third meaning of homology). Similarity should not be asserted to be *significant* unless patently obvious or accompanied by a probability statement and its method of determination ($\chi^2$, standard measure, binomial, etc.).

As recommended to the IUB, the preferred single letter code for nucleotide bases including ambiguity is: A = adenine, C = cytosine, G = guanine, T = thymine, U = uracil, R = A/G (purine), Y = C/T (pyrimidine), M = A/C, W = A/T, S = C/G, K = G/T, B = C/G/T (not A), D = A/G/T (not C), H = A/C/T (not G), V = A/C/G (not T), N = X = A/C/G/T (any or unknown). For ambiguous nucleotides, T and U are equivalent.

**Literature Cited.**—Literature in the text should be cited by author and year and, where citation is to a book, the relevant pages thereof. Text citations of two or more works at a time should be given in chronological order. When a paper written by three or more authors is cited, write the name of the first author plus et al. The Literature Cited section at the end of the paper should be arranged alphabetically and then chronologically and should contain only works specifically cited in the text. References to papers that have not yet been published will be as for articles (see below), except that "accepted" (along with the journal name) will replace the volume and page numbers. "In press" will not be used. (When such papers include authors of the submitted manuscript, copies of those papers must accompany the submitted manuscript.)

For the style of citations, please note the following examples:

**Journal articles:**

Books:


Book chapters:


The abbreviations of periodicals should be those used by the Council of Biology Editors. Periodical titles may also be written out.

Articles should include the name of the reviewing editor (the Editor or Associate Editor with whom author has corresponded) at the end of the Literature Cited section.

Footnotes.—Footnotes should be used sparingly. When necessary, they should be indicated in the text by superscript arabic numerals; the notes themselves should be typed on a page separate from the text. Footnotes to tables are referenced by superscript letters, except for significance levels, which use asterisks; table footnotes should be typed on the same page as the table to which they pertain.

Tables.—Each table must have a brief and self-explanatory title, be numbered with arabic numerals in order of its appearance in the text, and be typed on a separate page. Large, complex tables are discouraged. Guidelines for table format may be found in the CBE Style Manual and the Chicago Manual of Style or may be obtained by writing the editors of this journal.

Legends.—Figure legends should be typed on pages at the end of the manuscript, after tables. Each legend must be descriptive so that the illustration can be understood apart from the text and must define abbreviations used in the illustration.

Illustrations.—Each illustration (figure) should be an original, not a photocopy. Illustrations should be separate and have uniform lettering. They should be numbered consecutively, following the sequence in which they are mentioned in the text. The place where each illustration is to be inserted may be indicated by a circled note in the margin of the typescript. Names of authors, figure number, and an arrow indicating proper orientation should be written lightly in pencil on the back of each figure. Line drawings must be of high quality; typewritten or hand lettering is unacceptable. Photographs should be high-contrast, glossy prints. Magnifications may be indicated by a micron bar or in the legend. Photographs to be reproduced without further reduction must be so marked and may not exceed 4 3/4 inches wide by 7 5/8 inches long (122 by 194 mm) in order to fit the journal format. Please keep in mind that these dimensions are maxima. Because of the need for a figure legend, the illustration cannot be the maximum size in both dimensions.

Proofs and Reprints

Offprint order forms will be sent to the author (or in the case of multiple authors, to the senior author) with page proofs. There will be no page charge for publication.
Information for Contributors

Editor in Chief
Walter M. Fitch
Department of Physiological Chemistry
University of Wisconsin—Madison
1300 University Avenue
Madison, Wisconsin 53706

Managing Editor
Masatoshi Nei
Center for Demographic and Population Genetics
University of Texas Health Science Center at Houston
P.O. Box 20334
Houston, Texas 77225

Associate Editors
Roy J. Britten
Kerckhoff Marine Laboratory
California Institute of Technology
Corona del Mar, California 92625

Ken W. Jones
Department of Genetics
University of Edinburgh
West Mains Road
Edinburgh EH9, 3JN Scotland

Wesley M. Brown
Division of Biological Sciences
University of Michigan
Ann Arbor, Michigan 48109

Richard K. Koehn
Department of Ecology and Evolution
State University of New York at Stony Brook
Stony Brook, New York 11794

Richard B. Flavell
Department of Cytogenetics
Plant Breeding Institute
Cambridge CB2 2LQ
England

Robert K. Selander
Department of Biology
University of Rochester
Rochester, New York 14627
A critical publication for everyone working in molecular biology. . .

Now in its fifth year of publication

DNA
A Journal of Molecular Biology
Editors:
Walter L. Miller, M.D.
and
James Roberts, Ph.D.

FROM LATEST ISSUES

Expression of Rat NADPH—Cytochrome P450 Reductase cDNA in Saccharomyces cerevisiae.
By H. Murakami, Y. Yabusaki and H. Ohkawa

A General Method for Retrieving the Components of a Genetically Engineered Fusion Protein.
By P.R. Szoka, A.B. Schreiber, and H. Chan

Synthesis of Bovine Proactin in tscherchii coli.
By D.N. Luck, J.K. Ngsee, F.M. Rottman and M. Smith

Isolation and Characterization of the α-1-Antitrypsin Gene of Mice.

Construction of a Systematic Set of tRNA Mutants by Ligation of Synthetic Oligonucleotides into Defined Single-Stranded Gaps.
By S.W. Cline, M. Yarus and P. Wier

Binding of the Human Glucocorticoid Receptor to Defined Regions in the Human Growth Hormone and Placental Lactogen Genes.
By P.H. Eliard, M.J. Marchand, G.G. Rousseau, P. Vornstecher, M. Mathy-Hartert, A. Belayew, and J. A. Martial

Molecular Cloning of Bovine Viral Diarrhea Viral Sequences.
By A. Renard, C. Guis, D. Schmetz, L. Dagenais, P. Pastoret, D. Dina and J.A. Martial

Hormone-Dependent Covalent Modification and Processing of Human Progesterone Receptors in the Nucleus.
By K.B. Horowitz, M.D. Francis, and L.L. Wei

A Highly Modular Cloning Vector for the Analysis of Eukaryotic Genes and Gene Regulatory Elements.
By D.S. Pfarr, G. Sathe, and M.E. Reff

Characterization of the Polypeptide Composition of Human Factor VIII:C and the Nucleotide Sequence and Expression of the Human Kidney cDNA.
By M.A. Truett et al.

An Ubiquitous Interspersed DNA Sequence Family in An Insect.
By J.Y. Bradfield, J. Locke, and G.R. Wyatt

Cloning and isoation of Human Cytochrome P-450 cDNA's Homologous to Dioxin-Inducible Rabbit mRNAs Encoding P-450 4 and P-450 6.
By L.C. Quattrochi, S.T. Okino, and U.R. Pendurthi

Structure of the Human Vasoactive Intestinal Polypeptide Gene.

Note to authors: There are no page charges in DNA.
Volume 5, 1986, 6 issues: USA, $135; Overseas Air, $175
Volume 4, 1985, 6 issues: USA, $135; Overseas Air, $175
Volume 3, 1984, 6 issues: USA, $135; Overseas Air, $175
Volume 2, 1983, 4 issues: USA, $135; Overseas Air, $175
Volume 1, 1982, 4 issues: USA, $135; Overseas Air, $175

Send subscription orders to: Mary Ann Liebert, Inc. publishers
157 East 86th Street, New York, N.Y. 10028 • (212) 289 2300
This Publication is available in Microform.

**University Microfilms International**

Please send additional information for
Name (name of publication)
Institution
Street
City
State Zip
300 North Zeeb Road
Dept. P.R.
Ann Arbor, Mi. 48106

---

**STATEMENT OF OWNERSHIP, MANAGEMENT AND CIRCULATION**
(Required by 39 U.S.C. 3685)

1. Title of publication: Molecular Biology and Evolution
   a. Publication number: 07374038
2. Date of filing: November 23, 1985
3. Frequency of issue: Bi-Monthly
   a. No. of issues published annually: 6
   b. Annual subscription price: $65.00
4. Location of known office of publication: 5801 S. Ellis Avenue, Chicago, Cook, Illinois 60637.
5. Location of the headquarters or general business offices of the publishers: 5801 S. Ellis Avenue, Chicago, Cook Illinois 60637.
6. Names and addresses of publisher, editor, and managing editor:
   Publisher: The University of Chicago Press, 5801 S. Ellis Avenue, Chicago, Illinois 60637
   Editor: Walter M. Fitch, University of Wisconsin, Madison, Wisconsin 53706
   Managing Editor: None
7. Owner: The University of Chicago Press, 5801 Ellis Avenue, Chicago, Illinois 60637
8. Known bondholders, mortgagees, and other security holders owning or holding 1 percent or more of total amount of bonds, mortgages or other securities: None
9. The purpose, function, and nonprofit status of this organization and the exempt status for Federal income tax purposes have not changed during preceding 12 months.
10. Extent and nature of circulation:
    | Average Number Copies | Single Issue | Each Issue | Nearest During | Preceding 12 Months |
    |------------------------|-------------|------------|----------------|--------------------|
    | Total number copies printed | 2,166 | | | 1,800 |
    | Paid circulation: |
    | 1. Sales through dealers and carriers, street vendors and counter sales | 806 | 862 |
    | 2. Mail subscriptions | 806 | 862 |
    | Total paid circulation | 1,612 | 1,724 |
    | Free distribution: |
    | 1. Samples, complimentary, and other free copies | 52 | 56 |
    | Total distribution (sum of C & D) | 858 | 918 |
    | Copies not distributed: |
    | 1. Office use, leftovers, unaccounted, spoiled after printing | 1,308 | 882 |
    | 2. Returns from news agents | | |
    | Total (sum of E and F) | 2,166 | 1,800 |
11. I certify that the statements made by me above are correct and complete.

ROBERT SHIRRELL, Journals Manager
LIFE SCIENCES
The International Medium for Rapid Publication of Communications in the Life Sciences

Executive Editor: R BRESSLER, Department of Pharmacology, College of Medicine, University of Arizona, Tucson, AZ 85724, USA

Life Sciences publishes papers concerning the pharmacological aspects of biochemistry, chemistry, endocrinology, immunology, medicinal chemistry, microbiology, molecular biology, pathology, physiology and toxicology. Reviewed on the basis of their scientific merit and the need for rapid publication, these papers cover new information; minireviews on selected facets of a scientific field undergoing rapid change; brief conceptual papers based on original and/or literature data; relevant clinical discoveries; collections of selected papers on current concepts in a particular field; and occasional symposia papers.

Software Survey Section — This new section reports developments in appropriate specialist software.

Subscription Information
Published weekly (Volumes 36 & 37)
Annual subscription (1985) US$800.00
Two-year rate (1985-86) US$1520.00

FREE SPECIMEN COPIES AVAILABLE ON REQUEST

A selection of papers
Interaction between 2-chloroadenosine and \( \alpha \)-adrenoceptors in rat vas deferens, RONG-TSAN LAI et al.
Promoting effect of concanavalin A on transport of sodium cefoxitin and phenol red from rat rectal compartment, T NISHIHATA & T HIGUCHI.
Possible mechanism of uptake for several compounds in ionized form through human erythrocyte membrane, T NISHIHATA et al.
Salicylate-promoted permeation of cefoxitin, insulin and phenylalanine across red cell membrane, possible mechanism, T NISHIHATA et al.
Presynaptic \( \alpha \)-adrenergic stimulation leads to growth hormone release in the dog, S G CELLA et al.
Ovariectomized Sprague-Dawley and Long-Evans rats release prolactin differentially in response to estrogen, D M LAWSON et al.
Pain threshold and morphine activity in vitamin D-deficient rats, C BAZZANI et al.
Effects of dopaminergic and serotonergic drugs on ethanol-induced hypothermia, S YAMAWAKI et al.
Proglumide: selective antagonism of the rumination but not gastric motor effects induced by pentagastrin in sheep, L BUENO et al.
Alterations in creatine kinase, ornithine decarboxylase, and transglutaminase during muscle regeneration, M SADEH et al.

Advertising rate card available on request.
Back issues and current subscriptions are also available in microform.
Prices are subject to change without notice. Journal prices include postage and insurance. Sterling prices are available to UK and Eire customers on request.
Forthcoming

Long Interspersed L1 Repeats in Rabbit DNA are Homologous to L1 Repeats of Rodents and Primates in an Open Reading Frame Region
G. William Demers, Kilian Brech, and Ross C. Hardison

Nucleotide Sequence of the Genes for Tryptophan Synthase in Pseudomonas aeruginosa
Ayele Hadero and Irving P. Crawford

Molecular Evolution of Mammalian Lactate Dehydrogenase-A Genes and Pseudogenes: Association of a Mouse Processed Pseudogene with a B1 Repetitive Sequence
Kayoka M. Fukasawa, Wen-hsiung Li, Kiyohito Yagi, Chi-cheng Luo, and Steven S.-l. Li

Protein Sequence Evidence for Monophyly of the Carnivore Families Procyonidae and Mustelidae
Wilfried W. de Jong

Molecular Evolution of Pancreatic-type Ribonucleases
Jaap J. Beintema, Walter M. Fitch, and Antonella Carsana

Concerted Transpositions of Mobile Genetic Elements Coupled with Fitness Changes in Drosophila melanogaster

Phylogeny and DNA-DNA Hybridization (Letter to the Editor)
Maryellen Ruvolo and Temple F. Smith

On the Delta Q Test of Templeton (Letter to the Editor)
Naruya Saitou

A Response to the Above Letters (Letter to the Editor)
Alan R. Templeton
Molecular Biology and Evolution

Editor in Chief
Walter M. Fitch, University of Wisconsin

Managing Editor
Masatoshi Nei, University of Texas

Associate Editors
Roy J. Britten, California Institute of Technology
Wesley M. Brown, University of Michigan
Richard B. Flavell, Cambridge University
Ken W. Jones, University of Edinburgh
Richard K. Koehn, State University of New York at Stony Brook
Robert K. Selander, University of Rochester

Editorial Assistant
Jon Bartels

Editorial Board
Fred W. Allendorf
University of Montana
Norman Arnheim
SUNY at Stony Brook
John C. Avise
University of Georgia
Francisco J. Ayala
University of California, Davis
Jaap Beintema
Rijksuniversiteit, Groningen
Anthony Hugh D. Brown
CSIRO, Canberra
L. L. Cavalli-Sforza
Stanford University
Elizabeth A. Craig
University of Wisconsin—Madison
Irving P. Crawford
University of Iowa
W. Ford Doolittle
Dalhousie University
Gabriel Dover
Cambridge University
Marshall H. Edgell
University of North Carolina
George Fox
University of Houston
Morris Goodman
Wayne State University
R. Grantham
Université de Lyon
Michael Grunstein
University of California, Los Angeles
Barry G. Hall
University of Connecticut
Daniel L. Hartl
Washington University
Alec Jeffreys
University of Leicester
Motoo Kimura
National Institute of Genetics, Japan

Costas Krimbas
Agricultural College of Athens, Greece
Charles H. Langley
National Institute of Environmental Health Sciences, North Carolina
Richard C. Lewontin
Harvard University
Wen-Hsiung Li
University of Texas at Houston
Alan Maxam
Harvard Medical School
George L. Gabor Miklos
Australian National University
Roger Milkman
University of Iowa
Takashi Miyata
Kyushu University, Japan
Tomoko Ohta
National Institute of Genetics, Japan
Dennis A. Powers
Johns Hopkins University
Nils Ryman
University of Stockholm, Sweden
Barbara Schaal
Washington University
Temple F. Smith
Northern Michigan University
Howard M. Temin
University of Wisconsin—Madison
Michael S. Waterman
University of Southern California
Sherman Weissman
Yale University Medical School
Gregory S. Whitt
University of Illinois at Urbana Champaign
Eleutherios Zouros
Dalhousie University
Molecular Biology and Evolution
Volume 3, Number 2, March 1986

99 Nomenclature for Incompletely Specified Bases in Nucleic Acid Sequences
   Nomenclature Committee of the International Union of Biochemistry (NC-IUB)

109 An Analysis of Replacement and Synonymous Changes in the Rodent L1 Repeat Family

126 Evolutionary Implication of Heterogeneity of the Nontranscribed Spacer Region of Ribosomal DNA Repeating Units in Various Subspecies of Mus musculus
   Hitoshi Suzuki, Nobumoto Miyashita, Kazuo Moriwaki, Ryo Kominami, Masami Muramatsu, Takeharu Kanehisa, François Bonhomme, Michael L. Petras, Ze-chang Yu, and De-yuan Lu

138 Natural Selection and the Molecular Clock
   John H. Gillespie

156 Relative Contributions of Germline Gene Variation and Somatic Mutation to Immunoglobulin Diversity in the Mouse
   Takashi Gojobori and Masatoshi Nei

168 Variation in the Number of α-Globin Loci in Sheep
   A. Rando, L. Ramunno, and P. Masina

Book Review

177 Isozymes: Current Topics in Biological and Medical Research, Volume 12, edited
   by M. C. Rattazzi, J. G. Scandalios, and G. S. Whitt
   Thomas S. Whittam