[16] Evolutionary Analyses of Repetitive DNA Sequences

By MARK A. BATZER, CARL W. SCHMID, and PRESCOTT L. DEININGER

Introduction

In the mid 1970s, Britten and colleagues developed simple, reliable procedures for renaturing DNA duplexes from dissociated single strands.[1] These investigators immediately recognized that both the rate of cross-hybridization and the thermal stability of resulting interspecies DNA heteroduplexes measure the DNA sequence relatedness of divergent species. Whereas the single-copy sequences are directly suitable for phylogenetic comparisons, repetitive sequences present complications. A single-copy sequence from one species is constrained to cross-hybridize with its ortholog from a divergent species so that the mispairing of the resulting heteroduplex reflects the sequence relatedness of the two species.[2,3] In contrast, a repetitive sequence can hybridize to any one of many potential complements (i.e., paralogous sequences) and will rarely hybridize to its corresponding ortholog. The mispairing of paralogous heteroduplexes more nearly reflects the divergence of these repeated sequences from their most recent common ancestral sequence than the divergence of the species being examined.

For purposes of identification, we refer to the preceding studies as being "genomic hybridizations" in that they involve the cross-hybridization of either total DNA or some large subfraction of total DNA, that is, a mixture of many different sequences. In contrast, specific nucleotide substitutions can be targeted with short oligonucleotide probes to facilitate "sequence-specific" hybridization. Despite the limitations mentioned above, genomic hybridization of repetitive sequences has led to many worthwhile conclusions that have both stood the test of time and have indeed been verified by more direct cloning and sequencing results. These achievements are worth noting as they document the applicability of this approach. We limit this synopsis to studies of human DNA that are relevant to topics we are examining by improved techniques described below.

The rate of hybridization of repetitive human DNA suggested the existence of a single major repetitive DNA sequence family, a prediction

[1] E. H. Davidson, G. A. Galau, R. C. Angerer, and R. J. Britten, *Chromosoma* **51,** 253 (1975).

[2] B. H. Hoyer, N. W. van de Velde, M. Goodman, and R. B. Roberts, *J. Hum. Evol.* **1,** 645 (1972).

[3] D. E. Kohne, J. A. Chiscon, and B. H. Hoyer, *J. Hum. Evol.* **1,** 627 (1972).

that was subsequently verified by the identification of human *Alu* repeats.[4,5] The melting temperature of renatured human repeats indicated these sequences to be approximately 20% divergent from each other, a value confirmed and refined by detailed sequence comparisons.[6] Comparisons of the melting temperatures of human–chimpanzee DNA heteroduplexes showed that repetitive and single-copy sequence classes diverge at similar rates, a conclusion that has also been verified by sequence comparisons.[7,8] Although the thermal stability of DNA heteroduplexes accurately indicated the relative divergence of the major family of repeats in divergent primates (i.e., human, chimpanzee, monkey, and galago, a prosimian), these observations did not reveal specific sequence differences that distinguish the major family of repeats in human and galago genomes.[9,10] We think it is likely that today the same questions would be investigated by very different and more incisive procedures and do not review these earlier genomic hybridization methods.

The study of repetitive DNA sequences has been refined by advances in cloning, sequencing, and oligonucleotide synthesis. Whereas the original studies of repeated DNA evolution had to analyze whole families of repeated DNA sequences using genomic hybridization techniques, it is now possible to use cloning and DNA sequence analysis to define subfamilies of repeated DNA sequences. These subfamilies may then be characterized rapidly utilizing other approaches, such as specific oligonucleotide probes and the polymerase chain reaction (PCR).[11] It is these procedures that are covered in more detail here.

Current Approaches

It is difficult to improve on direct DNA sequence comparisons for evolution studies of the repeated DNA sequences. The only drawback is that these studies are relatively labor intensive, limiting the experimental sample to a much smaller one than can be studied using hybridization procedures. As a first step in analyzing any repeated DNA family, however, several independent copies should be sequenced in order to obtain some

[4] C. M. Houck, F. P. Rinehart, and C. W. Schmid, *Biochim. Biophys. Acta* **518,** 37 (1978).

[5] C. M. Houck, F. P. Rinehart, and C. W. Schmid, *J. Mol. Biol.* **132,** 289 (1979).

[6] P. L. Deininger, D. J. Jolly, C. M. Rubin, T. Friedmann, and C. W. Schmid, *J. Mol. Biol.* **151,** 17 (1981).

[7] P. L. Deininger and C. W. Schmid, *Science* **194,** 846 (1976).

[8] I. Sawada, C. Willard, C.-K. J. Shen, B. Chapman, A. C. Wilson, and C. W. Schmid, *J. Mol. Evol.* **22,** 316 (1985).

[9] P. L. Deininger and C. W. Schmid, *J. Mol. Biol.* **127,** 437 (1979).

[10] G. R. Daniels, G. M. Fox, D. Loewensteiner, C. W. Schmid, and P. L. Deininger, *Nucleic Acids Res.* **11,** 7579 (1983).

[11] K. B. Mullis and F. A. Faloona, this series, Vol. 155, p. 335.

knowledge of the structure and general variability of the sequences in the family. Sequence analysis of fairly large numbers of sequences may eventually be required to determine the detailed subfamily structure of a repeated DNA family. Traditional subcloning procedures for sequencing are now being supplemented and superseded by a variety of extremely promising PCR approaches[12] and novel cloning vectors such as λ ZAP II.[13] For these reasons, we think it is both likely and desirable that future investigations of phylogenetic relatedness will rely increasingly on DNA sequence comparisons and PCR approaches. However, hybridization techniques continue to be a valuable indirect method of rapidly comparing sequences at a large number of loci or between a large number of individuals or species. Additionally, they provide simple, effective methods of isolating clones for subsequent sequence analysis.

Oligonucleotide synthesis provides precisely defined sequences for use as hybridization probes to specific sequences. Once a family, or subfamily, of repeated sequences is defined in sequence, further DNA sequences are easily determined and sequence data accumulate in readily accessible databanks. As suggested above, genomic hybridization studies involving the cross-hybridization of many different sequences are relatively insensitive to precise differences that may distinguish otherwise closely related sequences. In contrast, the utility of oligonucleotide hybridization probes for this purpose is illustrated by recent findings concerning human *Alu* repeats. The number of human *Alu* repeats that have been sequenced is especially large owing both to the ubiquity of *Alu* repeats in the human genome and to the special emphasis human DNA has received in sequence studies. The nucleotide sequences of individual subfamily members can be aligned and family or subfamily consensus sequences determined (Fig. 1A). The various subfamilies are defined by members which share common nucleotide variants. Independent analysis of *Alu* sequences by six laboratories suggested the existence of distinct *Alu* sequence subfamilies that inserted into the human genome at different times in evolution.[14-19] Examples of the consensus sequences advanced for these subfamilies are shown in Fig. 1B. Whereas six laboratories arrived at similar conclusions

[12] W. Bloch, *Biochemistry* **30,** 2735 (1991).

[13] J. M. Short, J. M. Fernandez, J. A. Sorge, and W. D. Huse, *Nucleic Acids Res.* **16,** 7583 (1988).

[14] V. Slagel, E. Flemington, V. Traina-Dorge, H. Bradshaw, Jr., and P. L. Deininger, *Mol. Biol. Evol.* **4,** 19 (1987).

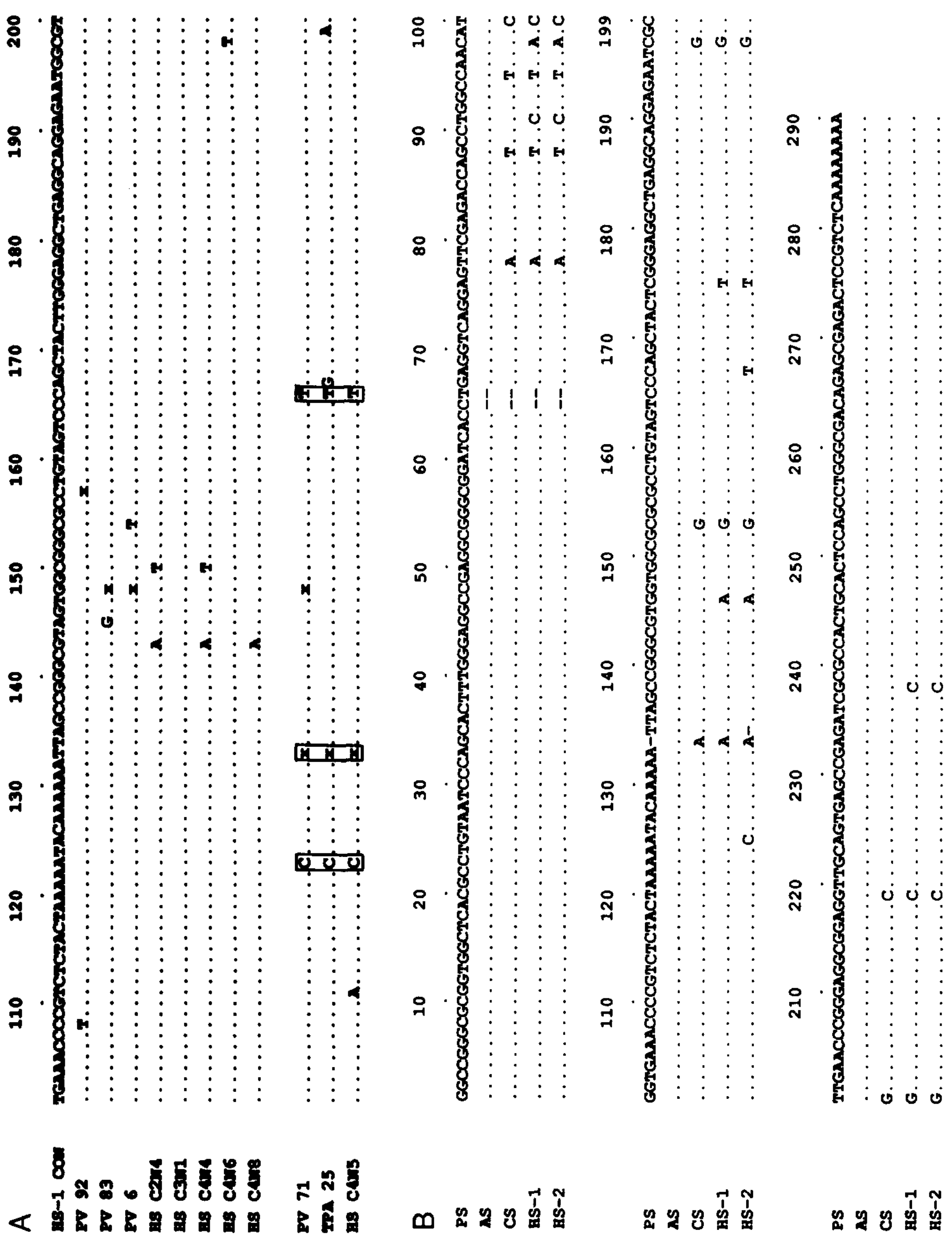
[15] C. Willard, H. T. Nguyen, and C. W. Schmid, *J. Mol. Evol.* **26,** 180 (1987).

[16] R. J. Britten, W. F. Baron, D. B. Stout, and E. H. Davidson, *Proc. Natl. Acad. Sci. U.S.A.* **85,** 4770 (1988).

[17] J. Jurka and T. Smith, *Proc. Natl. Acad. Sci. U.S.A.* **85,** 4775 (1988).

[18] Y. Quentin, *J. Mol. Evol.* **27,** 194 (1988).

[19] D. Labuda and G. Striker, *Nucleic Acids Res.* **17,** 2477 (1989).


A
HS-1 CON
PV 92
PV 83
PV 6
HS C2H4
HS C3H1
HS C4H4
HS C4H6
HS C4H8
PV 71
TPA 25
HS C4H5
110
120
130
140
150
160
170
180
190
200
B
PS
AS
CS
HS-1
HS-2
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
199
210
220
230
240
250
260
270
280
290

concerning the existence of *Alu* subfamilies,[14-19] there are naturally some differences in details concerning the number of subfamilies and refinement in the corresponding consensus sequences. There is also no common nomenclature for the *Alu* subfamilies. Because this chapter is not intended to resolve these issues, but rather to describe procedures that can be employed to distinguish between closely related sequences, we arbitrarily adopt the subfamily names identified by the Deininger group[20,21] to provide specific examples for our discussion. Matera *et al.*[22,23] have studied the identical subfamilies, albeit under different names, so that the hybridization procedure and results of our two laboratories can be directly compared. The human-specific (HS-1) subfamily differs by five concerted mutations from the cattarhine-specific (CS) subfamily (Fig. 1B). Two of

[20] M. A. Batzer and P. L. Deininger, *Genomics* **9,** 481 (1991).

[21] M. R. Shen, M. A. Batzer, and P. L. Deininger, *J. Mol. Evol.* **33,** 311 (1991).

[22] A. G. Matera, U. Hellmann, and C. W. Schmid, *Mol. Cell. Biol.* **10,** 5424 (1990).

[23] A. G. Matera, U. Hellmann, M. F. Hintz, and C. W. Schmid, *Nucleic Acids Res.* **18,** 6019 (1990).

FIG. 1. Alignment of several *Alu* subfamily members and comparison of five *Alu* consensus sequences. (A) Partial alignment of the TPA 25 *Alu* family member [S. J. Friezner Degen, B. Rajput, and E. Reich, *J. Biol. Chem.* **261,** 6972 (1986)] and several additional *Alu* HS subfamily members (see the references below). The consensus sequence (CON) is depicted at the top and represents the most common nucleotide found within the subfamily members at each position. Positions in the individual sequences that are the same as the consensus sequence are represented as dots. Substitutions are marked with the appropriate nucleotide, and deletions are indicated with an x or –. The boxed nucleotides represent HS-2 subfamily diagnostic mutations. (B) Representation of the five *Alu* consensus sequences as reported by Shen *et al.* [M. R. Shen, M. A. Batzer, and P. L. Deininger, *J. Mol. Evol.* **33,** 311 (1991)]. Each of the consensus sequences is defined by a number of diagnostic mutations and has been given a biologically relevant name. The PS (primate-specific) *Alu* consensus sequence represents the oldest and largest subfamily of *Alu* sequences found within primate genomes. The AS (anthropoid-specific) *Alu* consensus sequence differs from the PS consensus by a single 2-base pair deletion at position 65. The CS (catarrhine-specific) subfamily consensus sequence shares nine diagnostic mutations that are not found in the AS consensus sequence. The HS-1 (human-specific-1) and HS-2 (human-specific-2) consensus sequences are defined by five and three unique diagnostic mutations from the CS and HS-1 consensus sequences, respectively. The observation that each of the subfamilies has all of the diagnostic changes of the previous subfamily, as well as unique changes, supports the sequential appearance of different subfamilies within the genome. The HS subfamilies represent the most recently amplified *Alu* family members found within the human genome [M. A. Batzer, G. E. Kilroy, P. E. Richard, T. H. Shaikh, T. D. Desselle, C. L. Hoppens, and P. L. Deininger, *Nucleic Acids Res.* **18,** 6793 (1990); A. G. Matera, U. Hellmann, and C. W. Schmid, *Mol. Cell. Biol.* **10,** 5424 (1990); A. G. Matera, U. Hellmann, M. F. Hintz, and C. W. Schmid, *Nucleic Acids Res.* **18,** 6019 (1990); M. A. Batzer and P. L. Deininger, *Genomics* **9,** 481 (1991)].

these five differences, a C at position 91 and an A at position 98, are close enough so that both laboratories experimentally confirmed and extended these sequence predictions using oligonucleotide hybridization probes directed to this region (see Methods).

Special Considerations for Repetitive DNA Evolution

Many of the other chapters in this volume deal with evolutionary analyses of specific genes and unique DNA sequences.[24-28] There are, however, some evolutionary aspects unique to repeated DNA sequences. The most important of these factors is the amplification dynamics. Sequences become repetitive because there are amplification processes that make extra copies of them. These include retroposition and transposition mechanisms that would explain the majority of interspersed repeated DNA sequences, as well as recombination or replication slippage mechanisms that would probably explain most tandem replications. For any given repeated sequence, various factors may combine to increase or decrease the amplification rate of that sequence at various times in the evolutionary process. Thus, the dynamics of the amplification process could greatly affect the observed evolution of the family. This is particularly important in cross-species comparisons, because the amplification dynamics of a specific repeated DNA family may be altered in one species, relative to another.

Once a sequence amplification event occurs, the nature of any selection on the copies is important. In many (or even most) cases, it appears that the majority of repeated DNA sequences represent pseudogenes, which mutate at a neutral rate of evolution.[8] Along with amplification dynamics, the possible removal of repeated sequences must also be considered. Removal does not seem to play a major role with the interspersed repeated DNA elements,[8,29,30] but it is likely to be important in tandemly repeated satellite elements. Other mechanisms might also alter evolution of parts of a repeated DNA sequence. For instance, human *Alu* family copies are initially rich in CpG dinucleotides. These sites appear to be approximately 10-fold more subject to mutation than other sites in the genome,[19,31]

[24] D. Stahl, this volume [27].

[25] E. P. Lessa, this volume [31].

[26] J. M. Chesnick and R. A. Cattolico, this volume [13].

[27] D. B. Stein, this volume [12].

[28] R. DeSalle, A. K. Williams, and M. George, this volume [14].

[29] I. Sawada and C. W. Schmid, *J. Mol. Biol.* **192,** 693 (1986).

[30] B. F. Koop, M. M. Miyamoto, J. E. Embury, M. Goodman, J. Czelusniak, and J. L. Slightom, *J. Mol. Evol.* **24,** 94 (1986).

[31] M. A. Batzer, G. E. Kilroy, P. E. Richard, T. H. Shaikh, T. D. Desselle, C. L. Hoppens, and P. L. Deininger, *Nucleic Acids Res.* **18,** 6793 (1990).

probably because of methylation of these sites in the copies.[32,33] Other sequences, such as regions containing short repeated segments or homopolymeric runs, also seem subject to higher rates of mutation.

Methods

The study of repetitive sequences utilizes many relatively routine techniques in molecular genetics that are described in a number of excellent manuals[34,35] and are not discussed here. Instead, we consider the aspects of experimental design and data analysis that are specific for the study of repeated DNA families.

Cloning Repetitive DNA Sequences

Several special considerations arise when cloning repeated DNA sequences. One consideration is the stability of repeated sequences cloned into *Escherichia coli.* Instability is generally attributed to recombinations between tandemly repeated or inverted repeated sequences. These problems may be minimized by keeping the insert size as small as possible. Several genetic factors also influence instabilities in *Escherichia coli.* These include the general host restriction and modification systems, as well as *RecA* and *RecB* (homologous recombination) and *uvrC* and *umuC* (recombination involving inverted repeats).[36] Methylation has also been found to have a significant effect on the cloning of methylated DNA fragments, with hosts deficient in *mcrA* and *mcrB* host methylation being the best choice.[37]

Second, it is important to consider whether a clone library will be representative of a particular repeated sequence. Besides the genetic factors, above, unusual patterns of restriction sites in some repeated sequences may influence their relative abundance in a library. This would be more likely for a tandemly repeated sequence or a very long repeated sequence than for short, interspersed repeated DNA sequences. Traditional λ or plasmid libraries would be sufficient for most studies, but in certain situations it might be necessary to resort to DNA libraries of randomly frag-

[32] C. Coulondre, J. H. Miller, P. J. Farabaugh, and W. Gilbert, *Nature (London)* **244,** 775 (1978).

[33] A. P. Bird, *Nucleic Acids Res.* **8,** 1499 (1980).

[34] F. M. Ausabel, R. Brent, R. E. Kingston, D. D. Moore, J. G. Seidman, J. A. Smith, and K. Struhl, "Current Protocols in Molecular Biology." Wiley, New York, 1987.

[35] J. Sambrook, E. F. Fritsch, and T. Maniatis, "Molecular Cloning: A Laboratory Manual." Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 1989.

[36] A. Greener, *Strategies* **3,** 5 (1990).

[37] J. P. Doherty, M. W. Graham, M. E. Linsenmeyer, P. J. Crowther, M. Williamson, and D. M. Woodcock, *Gene* **98,** 77 (1991).

mented DNA derived from sonicated DNA for short fragments[38] or from DNA sheared through a syringe for large fragments.[20,31]

Screening of a library with standard hybridization conditions [42°, 1 *M* NaCl in 50% formamide, with a final wash at 65° in 0.1 × standard saline citrate (SSC)] will detect sequences having a maximum 20–30% mismatch. For more divergent repetitive sequences, a screening may also be attempted under somewhat lower stringency (e.g., 37° hybridization with a final wash at 50° in 1 × SSC) to determine whether a large number of sequences can be detected. The source of the probe may represent either sequences from a previously isolated member of a repeated DNA family or simply radiolabeled genomic DNA. In the latter screening, only those sequences that are present at a fairly high copy number (i.e., represent greater than 0.1% of that genome) will produce a hybridization signal in this experiment. The sensitivity of this approach could easily be increased by utilizing a C_0t fractionation to isolate various repetitive fractions that could then be utilized to probe the library.[38a,39]

Sequence Determination

Routine sequencing may be carried out using either shotgun or sequential deletion procedures.[40] The latter strategy is particularly useful to help align segments of a long tandemly repeated sequence. However, for experiments involving sequence analysis of multiple members of an interspersed repeated DNA family, sequence determination using sequencing primers from within the repeated DNA sequence can greatly streamline the analysis. The primers are generally 17–20 bases in length. By utilizing primers in both orientations and sequencing both strands of the sequence, many copies of a repeated DNA family, including their immediate flanking regions, can be rapidly and accurately sequenced (Fig. 2). In a repeated DNA family with a great deal of sequence mismatch this may not work well, as it is important that the primer match the sequence reasonably well, particularly at the last several 3′ bases.[41] Difficulties may also arise if more than one copy of the repeated sequence are present in a single recombinant DNA molecule. This could result in determination of a mixed sequence. However, it is also possible to minimize this problem in some cases. For instance, if efforts are being made to sequence members of a specific repeated DNA subfamily (see below), a primer can be made which will only sequence members of that subfamily by placing one of the subfamily

[38] P. L. Deininger, *Anal. Biochem.* **129,** 216 (1983).

[38a] M. S. Springer and R. J. Britten, this volume [17].

[39] P. E. Nisson, P. C. Watkins, J. C. Menninger, and D. C. Ward, *Focus* **13,** 42 (1991).

[40] P. L. Deininger, *Anal. Biochem.* **135,** 247 (1983).

[41] M. A. Batzer and P. L. Deininger, unpublished data (1989).

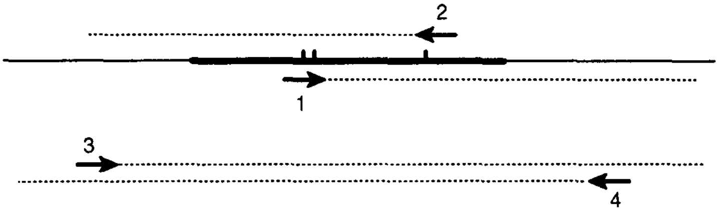


FIG. 2. Schematic representation of a DNA sequencing strategy for the analysis of repetitive DNA sequences. The individual *Alu* (repetitive) family member is depicted by the thick line. DNA sequencing primers are indicated by the arrows (1–4). Diagnostic mutations unique to the particular subfamily are indicated by the tick marks near the 3′ end of sequencing primers 1 and 2. Initially primers that anneal within the repetitive element are used to generate DNA sequence information, which begins within the element and proceeds out to unique DNA sequences 3′ and 5′ of the element. The design of these primers allows exact base pairing only with subfamily members (owing to the 3′ unique subfamily diagnostic mutations), permitting the analysis of relatively large clones that may contain more than one repetitive element. Subsequent primers complementary to the unique 5′ and 3′ flanking sequences (3 and 4, respectively) can be made for generating overlapping nucleotide sequence information or for PCR analysis of the locus.

diagnostic mutations at the 3′ end of the sequencing primer. This primer will be very ineffective at sequencing members of the same repeated DNA family that do not have the diagnostic change.

Analysis of Sequences

There are two basic strategies for comparing repeated sequences. The first, and most common, is to align the sequences (as in Fig. 1A) to develop a consensus sequence. Each individual sequence can then be compared to the consensus sequence. Many repetitive DNA sequences will vary from their consensus by 0.5% to more than 30%.[6,14] The consensus sequence provides an improved estimate of what the ancestral or parental repeated DNA sequence looked like prior to accumulation of mutations in the individual copies. In sequence families that have distinct subfamilies, such studies may be somewhat misleading unless the consensus used is that for the appropriate subfamily. However, such an alignment may help detect changes within subgroups of the repeated DNA family members that may represent subfamilies (Fig. 1A). There are multiple alignment programs (e.g., CLUSTAL in the PC/GENE suite from Intelligenetics) which can also help align sequences. However, the alignments almost always will require manual improvement, as these programs tend to include more insertions and deletions than necessary.

The alternate form of analysis would be to carry out pairwise comparisons of individual repeated DNA sequence members.[14,42] Such pairwise comparisons can provide an excellent method for suggesting subfamily structure, as well as dating a subfamily age. However, it is important to check the alignment criteria and manually refine the alignments in these analyses. Other chapters of this[43] and other texts[44] cover phylogenetic tree formation from such alignments in detail.

In making the alignments and determining subfamily structure, one must be aware that some sequence changes may occur in parallel in totally different members of a family. Changes that are held in common do not always indicate a subfamily (see Fig. 1A). Some analyses deal with this by including statistics on the probability of multiple common changes occurring in two family members.[45] One must also consider sequences (such as potential CpG methylation sites and simple sequences) that may be especially prone to specific types of sequence changes and may mimic subfamily relations, when they really represent parallel changes in random family members (Fig. 1A). For example, position 143 in Fig. 1A shows parallel mutations of a CpG dinucleotide to CpA. If subfamily changes are suspected, they may then be confirmed by further sequencing and oligonucleotide hybridization studies as described below.

Specific Sequence Hybridization

Specific sequence hybridizations utilize specific oligonucleotides or longer probes to detect repeated sequence subfamilies (as discussed below). The use of repeated DNA probes to screen recombinant DNA libraries for new sequence members has been discussed in general above. We consider this approach to be one of the best and most direct methods for determination of repeated DNA sequence copy number as well. If the repetitive family is randomly represented in the library, the most direct count of repetitive sequence members can be estimated by screening the library and determining how many hybridizing positive members are obtained relative to the number of plaques screened and the average insert size in the library. Dot blots, Southern blot hybridizations, or traditional C_0t plots are alternatives, but such measurements rely on relative renaturation rates. These rates depend not only on copy number, but also on sequence length and mismatching, potentially necessitating significant corrections to the data. In addition, as a result of the subfamily structure of repeated DNA se-

[42] P. L. Deininger and V. K. Slagel, *Mol. Cell. Biol.* **8,** 4566 (1988).

[43] D. M. Hillis, M. W. Allard, and M. M. Miyamoto, this volume [34].

[44] D. L. Swofford and G. J. Olsen, *in* "Molecular Systematics" (D. M. Hillis and C. Moritz, eds.), p. 411. Sinauer, Sunderland, Massachusetts, 1990.

[45] J. Jurka and A. Milosavljevic, *J. Mol. Evol.* **32,** 105 (1991).

quences, these hybridization techniques probably have a much higher signal-to-noise ratio than library screening owing to the background caused by related but nonidentical sequences.

Hybridization techniques are also the methods of choice to look at the RNA expression of repeated DNA family members. Such studies[46-48] are not discussed here, but they can be an important part of understanding the function and evolutionary mechanisms associated with a repeated DNA family.

Either cloned sequences or chemically synthesized oligonucleotides might be used as specific sequence hybridization probes. However, the most thermally stable region in a long duplex determines the temperature at which denatured single strands separate and the probe elutes from DNA immobilized on the filter, usually the critical parameter for the observations described below. A long duplex consisting of both poorly base-paired and exactly base-paired regions might denature at the same temperature as exact sequence complements. For this reason we generally expect oligonucleotide hybridization probes to be more discriminating than probes using longer cloned sequences and recommend their use whenever possible. This expectation is documented by Southern blot hybridization of a cloned HS subfamily member to a restriction digest of total human DNA (Fig. 3), where the cloned HS subfamily member hybridizes to a prominent *Bam*HI restriction fragment of 1 kilobase (kb). Higher stringency washing eliminates hybridization to both the 1-kb *Bam*HI band and the higher molecular weight smear, so that the stability of the hybrid formed by this band is indistinguishable from that of HS subfamily members. However, sequence analysis of the 1-kb *Bam*HI fragment demonstrates that is not a member of the HS subfamily reported in Fig. 1. Rather, the sequence of this restriction fragment reveals the presence of two complete *Alu* repeats and one partial *Alu* repeat interrupted by the *Bam*HI cloning site. Included within the sequence of one *Alu* repeat is a short (31 nucleotides) GC-rich (66%) sequence that only differs by two mispairs (one of which is a GT mispair) from the cloned *Alu* hybridization probe. Plausibly the high genomic copy number of this fragment, its multiple *Alu* composition, and the excellent sequence match between short regions of the hybridization probe and the fragment might all contribute to their cross-hybridization under stringent conditions. Regardless of the correct explanation, long cloned sequences do not provide the specificity required to identify sequence subfamilies.

Selection of Oligonucleotide Hybridization Probes. The shortest possi-

[46] K. E. Paulson and C. W. Schmid, *Nucleic Acids Res.* **14,** 6145 (1986).

[47] J. B. Watson and J. G. Sutcliffe, *Mol. Cell. Biol.* **7,** 3324 (1987).

[48] J. Skowronski, T. G. Fanning, and M. F. Singer, *Mol. Cell. Biol.* **8,** 1385 (1988).

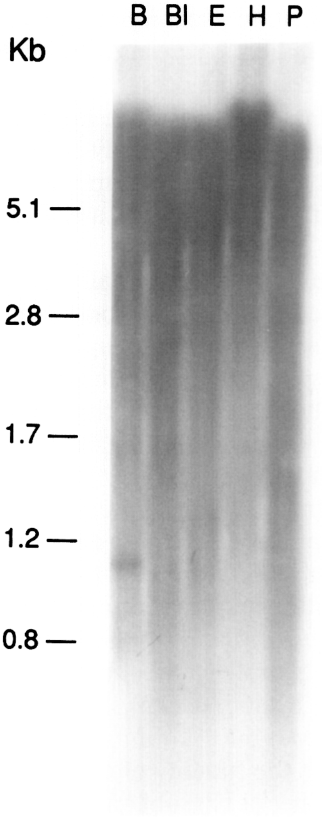


FIG. 3. Low specificity of long hybridization probes. Human DNA was digested with the following restriction enzymes: *Bam*HI (B), *Bgl*II (Bl), *Eco*RI (E), *Hin*dIII (H), and *Pst*I (P). The DNA was then transferred to filters and hybridized to a 600-bp fragment containing the polymorphic *Alu* repeat situated near the human tissue plasminogen activator gene [S. J. Friezner Degen, B. Rajput, and E. Reich, *J. Biol. Chem.* **261**, 6972 (1986)]. The blot was washed at 0.04 × SSC and 60° and exposed for 4.5 hr. These washing conditions approximate that of exactly paired sequence complements. The *Bam*HI band persists even after washing at 0.025 × SSC and 60°.

ble oligonucleotide that targets the maximum number of diagnostic base changes provides the most selective hybridization probe. As a lower limit on the size of the oligonucleotide, sequences of 16 or fewer nucleotides would occur at random in the human genome, which is about 2.5 billion base pairs in length. In our experience, oligonucleotides that consist of about 20 residues are sufficiently long to target a particular complement but are also sufficiently short to be sensitive to single nucleotide mismatches. To reduce background hybridization, all four nucleotides should

TABLE I
RELATIVE ORDER OF BASE-PAIRING STABILITY[a]

Watson-Crick	Non-Watson-Crick	Noncontributing
		T-T
	G-T	A-A
G-C A-T	G-A	C-C
	G-G	C-A
		C-T

[a] The noncontributing base pairs are the most disruptive to the hybridization, whereas the Watson-Crick base pairs act as the most positive contributors as originally shown by Ikuta *et al.* [S. Ikuta, K. Takagi, R. B. Wallace, and K. Itakura, *Nucleic Acids Res.* **15**, 797 (1987)].

be represented in the target sequence in an approximately even distribution, and targets that include runs of a particular base should be avoided if possible. However, as shown below, even a run of T residues on the end of an oligonucleotide can be used successfully.

Both the position and type of sequence mismatch determine duplex stability.[49] Pyrimidine-pyrimidine mispairs tend to be maximally destabilizing, whereas mispairs involving G tend to be least destabilizing (e.g., see Table I). By judicious choice of the complementary strand to be targeted for hybridization, the most destabilizing base mispairs can be selected for the oligonucleotide sequence. The mispair provides the maximum effect on thermal stability by being centrally located in the oligonucleotide. The terminal base pairs on the two ends of a DNA duplex are only "half-stacked" so that the duplex ends are already somewhat destabilized compared to the middle; a short duplex effectively melts from its ends. A base mispair, centrally located, essentially destabilizes the region that has the greatest effect on the strand-dissociation temperature.

The thermal stability of a short DNA duplex can be estimated by the simple 4 + 2 rule; each GC pair contributes 4° and each AT pair contributes 2° to the duplex melting temperature in 0.9 *M* NaCl solution.[50] Although more rigorous estimates of duplex stability are possible,[51] this simple method is reasonably accurate, and, in any event, we find it useful

[49] S. Ikuta, K. Takagi, R. B. Wallace, and K. Itakura, *Nucleic Acids Res.* **15,** 797 (1987).

[50] R. B. Wallace, J. Shaffer, R. F. Murphy, J. Bonner, T. Hirose, and K. Itakura, *Nucleic Acids Res.* **6,** 3543 (1979).

[51] C. R. Cantor and P. R. Schimmel, "Biophysical Chemistry, Part I: The Conformation of Biological Macromolecules." Freeman, San Francisco, 1980.

to compare the stabilities of perfectly paired and imperfectly paired duplexes empirically.

Optimization of Washing Conditions. One approach to setting exact hybridization and washing conditions is to determine the temperature at which the oligonucleotide elutes from filter-bound hybrids (Fig. 4). For example, a 22-nucleotide probe melts from its exact HS complement (TPA) at 67° compared to 66° as predicted by the 4 + 2 rule. Incorporating the two mispairs depicted in Fig. 4 lowers the duplex melting temperature by 10° (AFP) or an average of 5° per each base mispair. In a similar calibration experiment involving a different oligonucleotide sequence, we also observe a 10° depression in duplex melting temperature resulting from two base mispairs. As a possible generalization of these observations, there is approximately a 1° depression in DNA melting temperature for each 1%

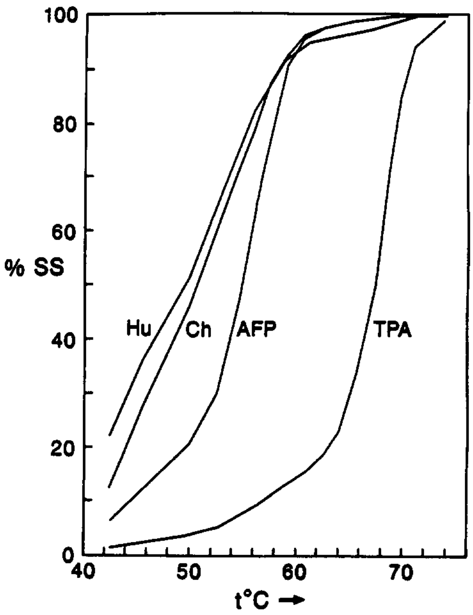


FIG. 4. Melting profiles of a subfamily-specific oligonucleotide. The oligonucleotide (5′ ATCGAGACCATCCCGGCTAAAA 3′) was melted from human (Hu), chimpanzee (Ch), and the TPA [S. J. Friezner Degen, B. Rajput, and E. Reich, *J. Biol. Chem.* **261,** 6972 (1986)] and AFP [P. E. M. Gibbs, R. Zielinski, C. Boyd, and A. Dugaiczyk, *Biochemistry* **26,** 1332 (1987)] *Alu* control DNAs. The underlined bases represent the HS-1 subfamily specific positions. The profile of BLUR 11 was indistinguishable from that of salmon sperm DNA (not shown). Note the high temperature melting component present in the human but not chimpanzee DNA. In this particular set of melting profiles, the filters were charged with 30 μg of each DNA. %SS, Percent single strand. [From A. G. Matera, U. Hellmann, and C. W. Schmid, *Mol. Cell. Biol.* **10,** 5424 (1990).]

TABLE II
EFFECT OF ADDED 3′ AT BASE PAIRS[a]

	Added 3′ AT pairs		
	0	1	2
Measured T_d(°)	40.5	44.5	46.5
Estimated T_d (4 + 2 rule,°)	38	40	42

[a] The oligonucleotide 5′ AGACTCCGTCTCTTTT 3′ is an exact match to the HS subfamily except for the four T residues situated at the 3′ end replacing the A residues normally occupying this position. Measured T_d is the thermal elution temperature of the oligonucleotide (5 × SSPE) from different DNA sequences with exact complements to the first 12 nucleotides and to additional 3′ AT pairs as listed. The 4 + 2 rule [R. B. Wallace, J. Shaffer, R. F. Murphy, J. Bonner, T. Hirose, and K. Itakura, *Nucleic Acids Res.* **6**, 3543 (1979)] predicted values are shown for comparison.

sequence mismatch.[52] Again the exact position and sequence context of a mispair can markedly influence duplex stability, so these generalizations are subject to the peculiarities of any particular oligonucleotide.

In one unfavorable case, we wished to isolate *Alu* members with 3′ ends that terminate in four or more T residues rather than the A-rich region which normally occupies this position. Our strategy was to first determine the dissociation temperatures (T_d) of exact complements having no T, one T, and two T residues, which are summarized in Table II. Interestingly, the thermal stability of these structures increase in about 2° increments for each added T residue, as predicted by the 4 + 2 rule. Using the preliminary calibration shown in Table II and modified hybridization and washing conditions that we use in library screening (see below), we succeeded in isolating *Alu* complements that terminate in four or more 3′ T residues. Based on these experiences, we find it useful to preface library screening with simple filter hybrid melts to define the useful temperature range of the selected oligonucleotide and then to make judicious choices for the library screening conditions as described below.

[52] T. I. Bonner, T. D. Brenner, B. R. Neufeld, and R. J. Britten, *J. Mol. Biol.* **81**, 123 (1973).

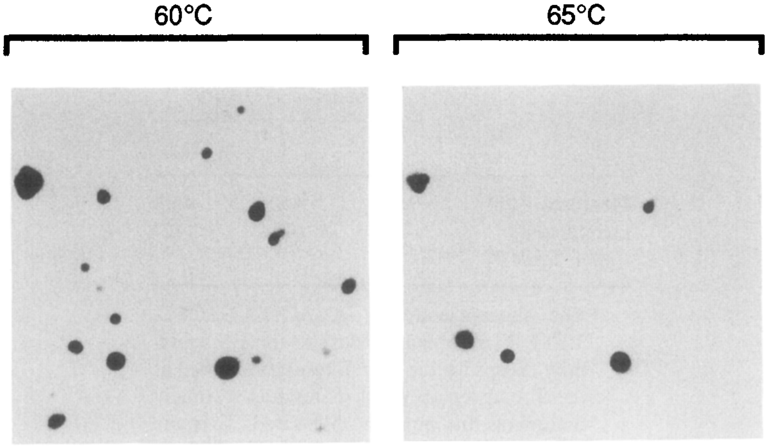


FIG. 5. Effect of varying the stringency of the final washes on the specificity of an oligonucleotide probe. Plaque lifts were performed on a human genomic library. The lifts were hybridized to the human-specific Alu family member oligonucleotide probe (5′ CACCGTTTTAGCCGGGATGG 3′, with the underlined bases representing the HS-1 specific positions) as previously described [M. A. Batzer and P. L. Deininger, *Genomics* **9**, 481 (1991)]. The final washes were then performed using 6× SSC, 0.05% sodium pyrophosphate at 60°, followed by final washes of the same filter at 65°. The autoradiographs were much cleaner after washes at 65° than at 60°, allowing the isolation of exact complements to the oligonucleotide at the higher temperature. Previous studies in our laboratory have shown that the clones which hybridize less intensely at 60° and subsequently disappear result from hybridization of inexact complements to the oligonucleotide.

Alternatively, the exact temperature of stringent washes that facilitate the isolation of perfect complements may be determined using library screening. Using the 4 + 2 rule, the T_d of another oligonucleotide that was used to isolate HS subfamily members (Fig. 5) should be 64°. After hybridization with the human-specific oligonucleotide at 42° overnight to plaque lifts from a human genomic library,[53] a comparison of washes (6× SSC/0.05% sodium pyrophosphate) at two different temperatures was made (Fig. 5). Filters washed at 60° contained both light and dark hybridizing plaques, whereas those washed at 65° contained only dark hybridizing plaques. Experiments in our laboratory have shown that the less intense hybridizations result from imperfect hybrids, whereas clones containing exact complements hybridize more intensely. Using this approach, the ideal temperature of the most stringent wash (65° in this case) to isolate perfect complements for any oligonucleotide probe can be determined.

[53] D. Woods, *Focus* **6**, 1 (1984).

In addition to the washing temperature, the hybridization temperature should also be selected to reduce background.[49] By hybridizing at the highest possible temperature, hybridization to inexact complements is minimized; the subsequent stringent washing then further reduces what is already a diminished background. We typically hybridize at about 5° lower than the elution temperature that was determined in the previously discussed filter hybrid melts (Fig. 4). Whenever possible, positive and negative control lifts of exact and inexact complements should be included in the library screening. After hybridization, usually 4 hr or overnight, we exhaustively wash with several room temperature changes of 5 × SSPE until the wash shows negligible radioactivity compared to the filters, as judged by a hand-held radioactivity monitor. One or more stringent washes are then performed for 5 min with shaking at a temperature just below that of the sharpest rise in the transition for the filter hybrid melting profile of exact sequence complements (e.g., 63° in the example of Fig. 4). Stringent washes are always followed by room temperature washes to dilute any residual radioactivity on the damp filters. Both the positive and negative controls are directly followed by a hand-held monitor during these procedures to ensure the selectivity of the stringent washing. If the background is too high, more stringent washing can be subsequently employed. Theory suggests that multiple stringent washings reduce background more than the signal, although the authentic hybridization signal is also diminished. Again, the internal positive and negative controls provide confidence that the procedures are being appropriately executed. We routinely perform our most stringent final washes at, or even below, the thermal elution temperature of exact sequence complements (e.g., 67° in the example of Fig. 4) The radioautograph for the exact complements following this most stringent wash should be noticeably less intense than that resulting from the previous less stringent washings, and it is hoped that there is no radioautographic exposure resulting for the negative control.

Analysis of Orthologous Loci

An alternative to the analysis of random copies of a repetitive DNA family is to study the evolution of a single repetitive DNA family member at a given locus. Such studies have proved very important in eliminating factors such as gene conversion and excision of repetitive sequences as being important considerations in *Alu* evolution.[8,29,30] They are also the most direct measure of the divergence rate seen for repetitive sequence family members. Traditionally these experiments have involved the cloning of a given genetic locus from a number of species and sequence analysis of that region to allow comparison. The PCR approach described below now makes such studies much more rapid and capable of being carried out easily through a wide range of species.

Choice of Primers. The development of the PCR has facilitated the exponential amplification of specific DNA sequences. This technique may be applied to the analysis of orthologous repetitive loci as described below. Initially oligonucleotide primers complementary to unique DNA sequences flanking any repetitive DNA element of interest are chosen manually[11] or with the aid of a computer program such as OLIGO.[54] These primers generally are 25 bases long, contain equal numbers of A, G, C, and T nucleotides, have about the same T_d as calculated by the 4 + 2 rule, and are manually compared to each other to preclude primer–dimer amplification. The primers are then searched against the EMBL/GenBank database (using a program such as QGSEARCH in the PC/GENE suite from Intelligenetics) to determine whether they reside in a previously described region of the genome. For efficient evolutionary PCR, the match of the primers with target DNA at the most 3′ nucleotides is critical for successful amplification.[55] We have previously found that the inclusion of an inosine residue at the 3′ terminal nucleotide mitigates mismatches at the 3′ terminal nucleotide, thereby enhancing the range and reproducibility of evolutionary PCR.[56]

Reaction Conditions and Optimization of Annealing Temperature. Amplification of repetitive loci is typically carried out in a 100-μl reaction consisting of 100 ng of target DNA, 750 ng of each primer, 2.5 units of *Taq* DNA polymerase, a 10× reaction buffer (generally supplied by the manufacturer of the *Taq* polymerase) and 200 μ*M* deoxynucleoside triphosphates (dNTPs). Reactions are carried out for 30 cycles, with each cycle consisting of 1 min at 94° (denaturation), 2 min at an experimentally determined annealing temperature, and 2 min at 72° (extension). One-fifth (20 μl) of the reaction products are subsequently fractionated on a 2% agarose gel containing 0.5 μg/ml ethidium bromide and visualized directly by UV fluorescence. The optimal annealing temperature for any set of primers is determined by amplifying target DNA using different annealing temperatures beginning at 5°–10° below the T_d of either member of the primer pair. The specificity of the reaction increases with increasing temperature, with the reaction products proceeding from a smear of nonspecific amplification products to the amplification of one or a few specific bands.

Amplification of Orthologous Loci. Once the optimal annealing temperature is determined (generally the highest temperature that provides

[54] W. Rychlik and R. E. Rhoads, *Nucleic Acids Res.* **17,** 8543 (1989).

[55] G. Sarker, J. Cassady, C. D. K. Bottema, and S. S. Sommer, *Anal. Biochem.* **186,** 64 (1990).

[56] M. A. Batzer, J. E. Carlton, and P. L. Deininger, *Nucleic Acids Res.* **19,** 5081 (1991).

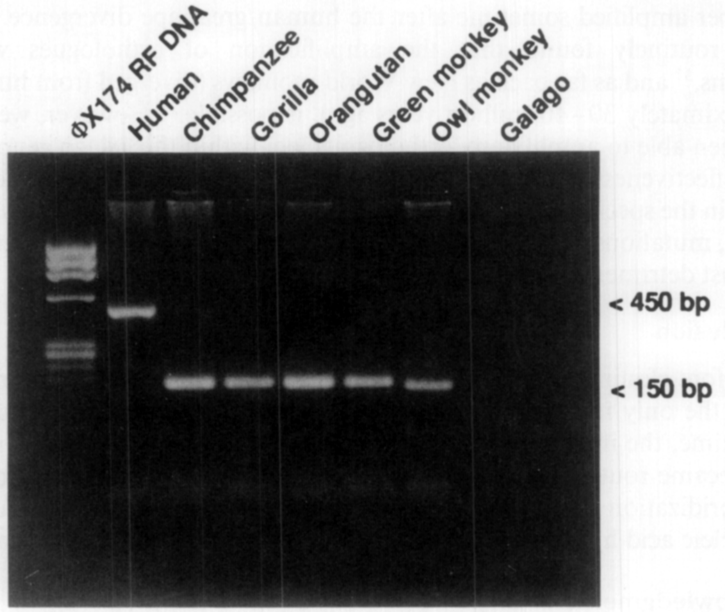


FIG. 6. PCR analysis of an individual *Alu* family member at orthologous loci within primate genomes. PCR amplification was carried out with unique primers (schematically demonstrated in Fig. 2, primers 3 and 4) flanking the HS C4N4 *Alu* family member [M. A. Batzer and P. L. Deininger, *Genomics* **9,** 481 (1991)]. Products resulting from the amplification of *Alu* subfamily member HS C4N4 were run on an agarose gel containing ethidium bromide and visualized by UV fluorescence. A 450-bp band is present if the *Alu* family member is located within the chromosome, whereas heterozygotes produce both bands; genomes that do not contain an *Alu* family member insertion produce only the 150-bp band. The marker was φX174 RF DNA digested with *Hae*III. The analysis shows that this *Alu* family member is located only within the human genome and is absent from the genomes of chimpanzee, gorilla, orangutan, green monkey, and owl monkey. No fragment was amplified from the galago genome, indicating that the galago was too divergent for the effective use of primers predicted from a gene located within the human genome.

sufficient specific product), orthologous loci can generally be amplified. The amplification of an *Alu* HS subfamily member (HS C4N4) locus is shown in Fig. 6. We can see that a 450-base pair (bp) fragment (indicating that a 300-bp *Alu* family member inserted between the two oligonucleotide primers) is present only in human DNA. Amplification of other ape (chimpanzee, gorilla, and orangutan), Old World monkey (green monkey), and New World monkey (owl monkey) DNAs resulted in the amplification of only a 150-bp fragment (no *Alu* family member present). The amplification of DNA from the prosimian galago resulted in no specific amplification products at this temperature. These data suggest that this *Alu* family member amplified sometime after the human/great ape divergence.[20] We have routinely found that the amplification of orthologues within humans,[57] and as far back as New World monkeys (diverged from humans approximately 30-40 million years ago), is possible.[21] However, we have not been able to amplify any orthologous loci within the galago genome.[21] The effectiveness of this procedure will be dependent on the molecular clock in the species of interest, as well as the random location of mutations. Again, mutations occurring at, or near, the 3′ end of one of the primers will be most detrimental to amplification.

Conclusion

Before cloning and routine sequence analysis, DNA renaturation provided the only method for comparing nucleic acid sequences. Naturally, for a time, the importance of this approach decreased as the newer methods became routine. However, the availability of oligonucleotides for use as hybridization probes and PCR primers has reinvigorated the usefulness of nucleic acid hybridization as a complement to DNA sequence analysis.

Acknowledgments

We thank Drs. Morris Goodman and Jerry Slightom for providing orangutan DNA, Ms. Utha Hellmann-Blumberg for the use of her unpublished data, and the department of photographic services at the Louisiana State University Medical Center for assistance. This research was supported by U.S. Public Health Service Grants GM-21346 (C.W.S.) and RO1 HG 00340 (P.L.D.) as well as a grant from the Cancer Crusaders (M.A.B.). Work by M.A.B. was also performed by Lawrence Livermore National Laboratory under the auspices of the U.S. Department of Energy Contract No. W-7405-ENG-48.

[57] M. A. Batzer, V. A. Gudi, J. C. Mena, D. W. Foltz, R. J. Herrera, and P. L. Deininger, *Nucleic Acids Res* **19**, 3619 (1991).

[17] DNA-DNA Hybridization of Single-Copy DNA Sequences

By MARK S. SPRINGER and ROY J. BRITTEN

Introduction

Single-copy DNA hybridization techniques have seen widespread application to problems in systematics. Most notably, Sibley and Ahlquist[1] have produced a phylogeny for many of the birds of the world. Other taxa

[1] C. G. Sibley and J. E. Ahlquist, "Phylogeny and Classification of Birds." Yale Univ. Press, New Haven, Connecticut, 1990.