Corruption of genomic databases with anomalous sequence

Edward D. Lamperti, J. Matthew Kittelberger, Temple F. Smith^{1,*} and Lydia Villa-Komaroff*

Department of Neurology, Children's Hospital, Harvard Medical School and ¹Molecular Biology Computer Research Resource (MBCRR), Dana-Farber Cancer Institute, Harvard School of Public Health, Boston, MA 02115, USA

Received March 3, 1992; Revised and Accepted May 8, 1992

ABSTRACT

We describe evidence that DNA sequences from vectors used for cloning and sequencing have been incorporated accidentally into eukaryotic entries in the GenBank database. These incorporations were not restricted to one type of vector or to a single mechanism. Many minor instances may have been the result of simple editing errors, but some entries contained large blocks of vector sequence that had been incorporated by contamination or other accidents during cloning. Some cases involved unusual rearrangements and areas of vector distant from the normal insertion sites. Matches to vector were found in 0.23% of 20,000 sequences analyzed in GenBank Release 63. Although the possibility of anomalous sequence incorporation has been recognized since the inception of GenBank and should be easy to avoid, recent evidence suggests that this problem is increasing more quickly than the database itself. The presence of anomalous sequence may have serious consequences for the interpretation and use of database entries, and will have an impact on issues of database management. The incorporated vector fragments described here may also be useful for a crude estimate of the fidelity of sequence information in the database. In alignments with well-defined ends, the matching sequences showed 96.8% identity to vector; when poorer matches with arbitrary limits were included, the aggregate identity to vector sequence was 94.8%.

INTRODUCTION

In northern blot hybridizations with radiolabeled single-stranded DNA prepared from constructions in the bacteriophage vector M13 (1), we found that probes prepared from the vector alone could hybridize to specific mammalian mRNAs (2, unpublished

data). We sought to identify these messages by searching the rodent section of the GenBank database (Release 63) with the FASTA comparison program (3) for entries with a low level of similarity to the part of the bacteriophage sequence that we had used as a probe. We found instead a group of unrelated rodent entries with extensive matches to M13. We extended our search with M13 to other taxonomic sections of the database. Several entries were M13-related vectors or contained flanking vector that had been identified in source publications. After these matches had been accounted for, there remained a set of nine sequences bearing unusual similarity to M13; all nine of the matches met established criteria for statistical significance (4).

The matching areas in these entries ranged from a 25-base block of identity (DRRSALA, (5)) to a span of greater than 400 bases with 97% identity to M13 (RATTKG6, (6)), a level of similarity normally found only between highly conserved regions of homologous genes in related species. These similarities, however, could not be explained by any evolutionary relationship. Further, for several database entries, including RATTKG6, the matching area was assigned to an intron and showed no similarity to the homologous intron in other species.

The M13-like areas in these entries shared several characteristics that distinguished the matching blocks as a discrete class. All occurred at one end of the reported sequences. All involved the same part of the vector—the border of the M13 polylinker. All were bounded at one end by one of the infrequent-occurring six-base restriction sites of the polylinker. These distinctive features suggested that the matching sequences were not randomly associated with M13 but instead had arisen directly from the vector by a set of events involving the restriction sites at the match boundaries.

This phenomenon was not exclusive to M13. When we extended our original search by using fragments spanning common insertion sites in pUC19 (1), pBR322 (7), lambda gt10 and 11 (8), and lambda EMBL 3/4 (9), we found statistically significant matches (4) to eukaryotic database entries with each

* To whom correspondence should be addressed at: Department of Neurology, Enders 250, Children's Hospital, 300 Longwood Avenue, Boston, MA 02115, USA

⁺ Present address: MBCRR, Molecular Engineering Research Center, Boston University, 36 Cummington St., Boston, MA 02215, USA

vector. The vector-like sequences in these entries shared some or all of the unusual features of our original matches with M13. Several matches did not meet our limit for statistical significance but presented other compelling evidence for accidental vector incorporation. In total, forty-five database entries showed evidence for a significant match to vector sequence (Table 1, refs 10–51).

If the matches in Table 1 represent vector-based artifacts, then the means by which they occurred must have been heterogeneous—all of the matching sequences could not have been generated by a single mechanism. The matches in Table 1 have been categorized by the mechanism which might best explain incorporation in each case. The mechanism in the first category is a simple extension of sequence reading beyond the cloning site into vector. One alignment in this group is depicted in Figure 1. In several cases the matching sequence could be the relic of an editing error in which bases read beyond the insertion point were inadvertently left in the reported sequence. In other cases (including the entry ONGMSPA) a restriction site used for cloning might have been rearranged and thus not recognized as the boundary between vector and insert. The second category consists of matches to vectors that were not used for sequencing but had been present at earlier cloning steps—the matching sequence might have been generated when part of an intermediate vector was unintentionally incorporated with the intended insert into a subclone used for sequencing. The third category involves more extensive subcloning accidents in which multiple vector sequences were incorporated or misrecognized, or in which a flanking fragment of vector had been substituted for an intended insert in a subclone. DRRSALA, depicted in Fig. 2A, bore tandem matches to the polylinkers of the vectors used for its initial cloning and subsequent sequencing (6). TRWKPECOC, in Figure 2B, is one of two trypanosomal kinetoplasts (37) which share regions of extensive sequence identity. One such region was bounded precisely by an EcoRI site. Beyond this site TRWKPECOC bore no significant similarity to its sister kinetoplast but instead matched closely the flank of M13 in an alignment bounded exactly by the M13 EcoRI site; EcoRI was used in the subcloning of this kinetoplast.

The fourth category in Table 1 includes sequences showing evidence of large-scale vector incorporation involving a significant fraction of the total reported sequence, or rearrangement in the early stages of cloning. The matching sequence in entry M24665 (47) was nearly identical to 200 basepairs (bp) of the flank of the insertion point in λ gt10, but in the orientation opposite that expected from a simple error in

Figure 1. Match between HUMVIPMR4 and pBR322. Middle line of sequence is from HUMVIPMR4 (11), bottom line from the complete sequence of the human gene for vasoactive intestinal polypeptide (64). For figures 1–3, sequences and numbering were from GenBank. Alignments were generated by FASTA; multiple alignments were assembled from FASTA output. Vertical lines (|) between bases indicate identity in an aligned area; hyphens (–) mark gaps; (X) marks alignment boundaries defined by FASTA.

recognizing the insertion site. This vector fragment might have been present during the initial cloning and attached to the cDNA before the vector arms were added. The alignments for HUMAMYA1, in Fig. 3A, and the remaining entries in this category deviated in a key feature from other matches in Table 1—they involved areas of vector far from expected cloning sites. Some matches in this category were bounded by restriction sites, but ones not normally used for cloning in the matching vector. Vector DNA may have contaminated reagents used in subcloning these sequences; various fragments of vector were then appended by blunt or compatible ligation to the intended inserts in the subclones.

It is also possible that a new clone might incorporate a contaminating vector fragment with part of a pre-existing insert attached. This is one explanation for the alignment in Fig. 3B. RATADHCY1 is a single chimaeric cDNA containing sequences with greater than 90% identity to parts of the transcripts of two normally separate mitochondrial genes. These regions in RATADHCY1 are joined by a 230bp sequence described as presenting no homology with any known genes (48). This bridging sequence in fact bore 95% identity (in 241bp) with M13. The area of M13 involved was no closer than 40bp to the viral polylinker, and was not bounded by restriction sites commonly used for cloning, suggesting that this construct had been generated by some complicated rearrangement or by ligation of multiple fragments.

The fifth category in Table 1 consists of three sequences that were altered after they had entered the database. We expected that the extent of some of the matching sequences we observed

Figure 2. Matches representing potential inclusion of multiple vector fragments or substitution for intended fragments in subclones. A: DRRSALA / M13L. Bottom sequence line shows an overlapping match to pBluescript SK⁺ (63), an intermediate vector in the cloning of DRRSALA (6). B: TRWKPECOC / M13L rc. Below TRWKPECOC is the sequence of its sister kinetoplast TRWKPBAMC (37).

in Release 63 of GenBank would lead unavoidably to their identification in subsequent releases. Indeed, some entries in Release 67 contained annotation about putative anomalies. In at least three entries, however, vector sequence apparently was

recognized but was then removed without annotation. One of the entries (M25531) was unannotated at the time of our original search; an edited version of the original entry, with the vector sequence excised, was substituted in later editions of GenBank

Table 1. Entries in unannotated, vertebrate, and invertebrate sections of Genbank Release 63 containing significant matches to vector, listed by category of potential vector incorporation, as explained in the text. Only the best match for each entry is shown. Vector sequences used in searches: **M13L**—1228 bases, position 5079 (1) through the M13 mp18 polylinker; **M13R**—650 bases, from the M13 polylinker to pos. 6935 (1); **pUC19**—pos. 1 to 1000 (1), containing the pUC19 polylinker; **pBR322**—pos. 3400 to 300 (7); λ gt10 (8,62)—the λ cl 434 gene (286bp), containing the Eco RI site used for cloning; λ gt11—the λ gt11 *lac* gene (210bp); λ EMBL L—the λ *int* gene, pos. 27812 to 28882 (9,62); λ EMBL R—vicinity of gene *N*, pos. 34287 to 35438 in λ . **rc**—reverse complement of vector sequence. Some matches were optimized with vectors related to probes or with sequence beyond probe boundaries. Some entry names were assigned to unannotated sequences after initial searches had been completed. Entries with bracketed names ([]) contained matches below the statistical limit (4) but considered significant because of bounding restriction sites or mismatch to expected homologs. **pBS** is pBluescript phagemid (63). Cited references are drawn from annotations in GenBank. Some references do not include sequence; others do not include areas matching vector. Percentage identity has been recalculated from FASTA output for matches with obvious restriction-site boundaries.

Genbank entry	Matching vector	% match	size (bp)	Function ascribed to matched sequence	ref. #
CATEGORY 1. Sequence reading beyond cloning site					
[HUMVIPMR4]	pBR322	100	17	intron/exon boundary	10
RATUGPSC	pUC19rc	100	26	5' pseudogene flank	11
[HUMPROLA]	pUC12	96.6	29	5' untrans. cDNA	12
ECOEXXBBD	M13L	100	26	part of adjacent gene	13
BSUCOTT	M13 Lrc	97	33	intergenic ?	14
[RATGAMT]	pUC19	85.4	41	5' gene flank	15
M24488	M13 Lrc	100	31	intergenic?	16
M27360	pUC19	100	30	immunoglobulin switch region	17
HUMHPARS2	pBR322	94	34	intergenic?	18
TTVV1TP	M13 L	84.1	63	intergenic?	19
RATCCK3	M13 L	100	34	intron (2nd)	20
M25335	M13 Lrc	100	36	3' gene flank	21
M21052	pUC19	88.2	68	intergenic	22
ONGMSPA	M13 Lrc	89.2	65	3' gene end	23
CHKYES	pUC19	100	46	3' untrans. cDNA	24
DRODSK	M13 L	100	57	3' gene flank	25
HUMIFNB3	pUC19	81.7	131	intergenic	26
HUMPK1	M13 L	100	89	5' pseudogene end	27
M27044	M13 Lrc	100	98	intron (2nd)	28
RATCAT02	pUC19rc	89.4	132	intron (2nd)	29
HUMCOL1A41	M13L	90.3	155	intron (51st)	30
CATEGORY 2. intermediate vector incorporation					
BOVSIG1	λ EMBL Rrc	100	41	5' gene end	31
BOVTG1	pBR322rc	95.2	145	5' gene end	32
PFAHRPC	λ	98	504	overlapping reading frames	33
HUMELA308	λ EMBL R	100	186	3' gene end	34
SHPKERWL	λ	98.9	444	3' gene end	35
CATEGORY 3. multiple vector incorporation or substitution					
[M27011]	pUC/M13	100	34	5' untrans. cDNA	36
DRRSALA	M13/pBS	98.5	65	3' gene end	5
TRWKPECOC	M13 Lrc	80	75	kinetoplast DNA	37
MUSOTC03	pUC9	96	100	intron (2nd)	38
DDID11	pBR322	96	150	3' gene end	39
HAMPRPH29	pUC19	97.6	289	3' gene end	40
RATTKG6	M13 L	97	433	intron (6th)	6
M26306	pUC19	97.5	525	5' gene end	41
M26307	pUC/pBS	86	136	as above	
CATEGORY 4. large-scale incorporation or rearrangement					
MUSNFMG	M13 Rrc	87.9	58	intron (2nd)	42
HUMTCRBSJ	M13 L	96	50	cDNA coding sequence	43
HUMAMYA1	M13 Lrc	77.6	143	5' gene flank	44
HUMTRNSDU	pUC19rc	85.8	169	cDNA spanning chromosomal translocation	45
MUSMB1A	M13 Lrc	98.4	124	cDNA 3' untrans. region	46
M24665	λ gt10rc	98.5	206	5' end of cDNA	47
RATADHCY1	M13 Lrc	94	241	bridge in chimaeric cDNA	48
CATEGORY 5. alterations after entry into database					
M25531	λ	99.4	718	3' gene flank	49
DROENHSPA	pUC19rc	98.8	253	cDNA 5' untranslated region	50
HUMT1418	M13 Lrc	89.7	562	chromosomal translocation	51

under the original accession number. The sequence bearing the entry name DROENHSPA in Release 63 contained an extensive vector match but was not similar to the published sequence cited in its annotation (50); in subsequent database editions the entire original sequence and its accession number could not be found, and the published sequence had been substituted under the name DROENHSPA. Between GenBank Releases 63 and 67, the entry name HUMT1418, together with its annotation, accession number, and its sequence—which included the most extensive match to M13 that we observed—had all vanished.

The presence of bona fide vector sequence would have consequences ranging from negligible to catastrophic for investigations connected with the entries in Table 1. Most matches in the first category were in areas ascribed to introns and gene flanks. In some cases vector sequence would contribute infrequent restriction sites, leading to erroneous expectations for genomic digestions. Vector incorporation in categories 2 and 3 would have affected several alignments in the source articles. Some matching sequence in Table 1 was assigned to splice junctions, chromosomal rearrangements, open reading frames, and to the coding regions of messages.

We expected that vector incorporation was disappearing as computerized sequence compilation grew more sophisticated. The presence in the database of vector fragments arising from editing errors has been widely assumed, and anecdotal reports (52,53) have drawn further attention to the issue. However, when we performed a preliminary search in parts of GenBank Release 67 with the full sequence of M13 (data not shown), in an attempt to find matches to areas of the vector that we had not used in our original search, we identified 50 new additions to the database with significant similarity to our original M13 probe. The addition of all of these sequences would more than double the size of Table 1; by contrast, the sections of the database we used have in aggregate grown by approximately 50% between Releases 63 and 67.

The entries in Table 1 represent only 0.23% of the nearly twenty thousand sequences searched in GenBank Release 63. We did not attempt, however, an updated and exhaustive search for all vector sequences in the current release of the database. We used a limited set of vector sequences—we did not search the database with complete sequences, or with less commonly used or proprietary vectors. In addition, we used a relatively stringent criterion for the significance of a match to these vector fragments, causing us to miss smaller or low-fidelity matches. Our limit for significance will thus produce not an accurate estimate but instead a lower boundary for the frequency of vector incorporation events. The aim of our analysis of the matches in Table 1 was to expose potential mechanisms by which vector had been incorporated; the apparent heterogeneity of those mechanisms indicates that a comprehensive identification of anomalous sequence in the database may be difficult. Our evidence suggests that sequence could be incorporated from vectors not normally used for sequencing and from areas distant from usual cloning sites. Other DNA that is present during cloning manipulations could be incorporated. Bacterial DNA could be introduced in various reagents, including restriction enzymes, or by recombination in a transformed host cell. Matches in the fourth category of Table 1 indicate that rearrangements could accompany incorporation, and that other eukaryotic sequences, including those from existing clones that might contaminate reagents used for cloning, could be integrated. The same heterogeneous mechanisms that have embedded large blocks and scrambled

fragments of recognizable vector sequence in the database could easily have created similar anomalies with other sequences that could not readily be identified. Together with the preliminary evidence that misidentified vector sequence is accumulating in newer releases, this suggests that the frequency of incorporation of anomalous sequence we have observed in an older version of GenBank is a significant underestimate of the extent of the problem in the current database.

Figure 3. Matches arising by possible rearrangements of vector in early cloning stages. **A:** HUMAMYA1 / M13 rc. HUMAMYA1 matches part of HUMAMYAGA, sequenced subsequently from the same human locus (65). The 3' boundary of this alignment is the 5' boundary of the match between HUMAMYA1 and M13. A hexanucleotide at this boundary (GAAATC) is similar but not identical to sequences cleaved by Eco RI under unusual conditions ('star activity') (66). **B:** RATADHCY1 / M13 rc. Matches described to separate mitochondrial genes (48) are shown here to corresponding areas of the rat mitochondrial genome (RATMTRNXX (GenBank Release 68)) (68). The M13 polylinker, not depicted, spans pos. 6230–6290.

Our observations have some bearing on more general issues of the management and fidelity of database information. The entries in the fifth category of Table 1 raise questions about the treatment of recognized errors. The changes in these entries violated precepts of the use of annotation numbers, which should be irreversibly linked to specific sequences. Edited sequences should receive new numbers, and the changes should be noted, to leave an 'audit trail' which can be used to trace a final edited entry back to its initial submission. It should not be acceptable to erase a sequence from the database without leaving a record of its existence, just as an erroneous publication cannot be erased without trace from the literature. In both cases, an immutable record is required to ensure that analyses based on existing sequences or other data can be either reproduced or reinterpreted subsequently.

There is an obvious advantage in recognizing and preventing anomalous sequence from entering the database. Programs now in use (54) can identify vector-based artifacts in new sequences as they are compiled and used for searches. The identification of vector fragments in new sequence would be made easier by comparison to a comprehensive database of vector sequences; the current GenBank version is outdated. Such a tool would be useful to those generating new sequences, but it is necessary that all new sequence be screened in a single uniform process by the managers of the database. The elimination of anomalously incorporated eukaryotic sequences would require more stringent measures. Almost all such artifacts would be exposed by generating overlapping sequence from both strands of a clone and by comparing sequence-based restriction maps with those generated by digestion of clones and of genomic DNA in Southern blots. Our evidence suggests that these necessary corroborative steps may frequently be ignored and that anomalous sequence continues to pass into the database. If GenBank continues to be subjected to centralized editing and evaluation (55), some of these sequences can at least be recognized and annotated. However, the continuing accumulation of anomalies, even if recognized, may ultimately compromise some uses of the database. It may be easy to find vector-like sequences, but for many of the entries in Table 1 it is not possible—in the absence of direct experimental evidence—to demonstrate beyond a reasonable doubt that the matching sequences originated from vectors. The nature of an incorporation artifact may be obvious when pointed out during the compilation of a sequence but impossible to prove long after the data was generated and the investigators had moved to other interests. Without such proof, a suspect sequence could not be decisively rejected in favor of another version. This would complicate the goal of establishing a concordance sequence in which errors are eventually eliminated. The presence of anomalous eukaryotic sequences, some of which might be indistinguishable from bona fide polymorphisms or transpositions, would make a concordance more difficult to achieve. This problem would be further aggravated by possible changes in the management of GenBank to a primarily archival accumulation of sequence information, with minimal evaluation, from the published literature (56). The assumption that the review process for journal publications is sufficiently rigorous to ensure sequence accuracy is unwarranted—most of the entries in Table 1 either came from published sequence or were based on publications showing part of the sequence.

If we treat incorporated vector fragments not as anomalies but instead as a set of test sequences embedded in the database, then the matches in Table 1 can be used for a crude measure of

sequence fidelity. Such an estimate carries significant qualifications: the matches in Table 1 are clearly a skewed sample of the database, vector fragments may have been overlooked during sequence compilation specifically because their borders had been rearranged, and the source of incorporated vector may not have been the same as that of our probe sequences. Nevertheless, for the alignments with well-defined 5' and 3' boundaries, the aggregate match to vector showed 96.8% identity (5532 of 5712bp). The best among these (M25531) had only two mismatches with the λ genome in a span of 718 basepairs. Table 1 includes, however, a discrete set of lower-fidelity matches whose identity to vector could not be quantitated exactly because the alignments did not have a well-defined boundary at one or both ends. Several of these sequences showed a disturbing pattern with an almost perfect match near one end but increasing numbers of mismatches with increasing distance from that end, extending to areas with no recognizable similarity either to vector or to expected homologous sequence from other sources. If these matches—with boundaries set by the comparison program—were included, the aggregate identity to vector sequence would be 94.8% (6519 of 6871bp).

The utility of database information for some purposes might not be seriously diminished if its overall accuracy fell in the range we have observed for incorporated vector. Searches for nucleotide similarity are tolerant of gaps and mismatches, and similarities among translation products can be identified even in the midst of frameshift errors (57). Conversely, small mistakes around splice junctions have compromised the identification of consensus sequences (58). The discussion of this issue has focused on dispersed and largely predictable errors involving individual bases (59). Our findings suggest that more extensive errors have arisen from large-scale, heterogeneous, and unpredictable incorporation of anomalous sequence. Such larger-scale anomalies will exacerbate problems in recognizing low-level similarities and consensus sequences, but they may also introduce misleading new similarities. A misplaced, unrecognized gene fragment in a database entry might cause it to match a sequence used to search the database, leading to erroneous expectations for the function or evolutionary relationships of the probe sequence. We identified the matches we found as potential artifacts because they were so numerous and similar; a single rearranged and accidentally incorporated eukaryotic sequence will not be so easily recognized in a database match. The combination of single-base errors and larger-scale anomalies may have their greatest impact on investigations involving polymerase chain reaction (PCR)-mediated amplification of gene fragments. PCR amplification may fail with oligonucleotides that are built to match erroneous sequence, and successful products may be considered artifacts if their sizes or restriction sites do not match expectations based on sequence with errors or anomalous inclusions. Our evidence supports the assumption that such problems will be less likely for amplification targets that are in the coding regions of well-studied genes, much more likely for an untranslated region used to identify a unique transcript or an intergenic region targeted as a likely site of polymorphisms. These problems are likely to increase dramatically with the large-scale employment of PCR in the proposed sequence tagged sites (STS) approach to the dissemination of human genomic mapping information (60, 61). Most of these problems could be avoided if the managers of the database applied a rigorous and uniform screening procedure, using analytical tools already well established, to all of the sequence entering the database and removed the accumulated

artifacts in a way that preserves the database as an immutable record. Simultaneously, efforts should be increased to develop a concordance of sequence information to make the database a more reliable resource.

ACKNOWLEDGEMENTS

This work is dedicated to the memory of Salvador Luria. FASTA and GenBank were provided by the Molecular Biology Computer Research Resource (MBCRR) at the Dana Farber Cancer Institute, Boston. We thank Regena Bradeen for clerical help, Paul Newmann for discussion and suggestions during this study, and Nina Irwin, Walter Gilbert, Pat Hogan, and Kathy Buckley for reviewing earlier versions of our manuscript. This work was supported by grants from the Council for Tobacco Research and NIH (NS27832, Mental Retardation Center Grant HD 18655). A more extensive version of Table 1 is available on request.

REFERENCES

- Yanisch-Perron, C., Vieira, J., & Messing, J. (1985) *Gene* **33**, 103–119.
- Lamperti, E.D., Rosen, K.M., & Villa-Komaroff, L. (1991), *Mol. Brain Res.* **9**, 217–231.
- Pearson, W.R. & Lipman, D.J. (1988), *Proc. Natl. Acad. Sci. (USA)* **85**, 2444–2448.
- Estimates of maximum match length expected at random ($L_{\max}$) between probes and the database came from modifications (Smith, T., Waterman, M.S., & Burks, C. (1985) *Nucleic Acids Res.* **13**, 645–656) of formulations of Erdos, P. & Renyi, A. (1970) *J. Analyse Math.* **22**, 103–111. Similar estimates are obtained from the equations of Karlin, S. & Ghandour, G. (1985) *Proc. Natl. Acad. Sci. (USA)* **82**, 5800–5804, who suggest that a match be considered statistically significant if its length exceeds $L_{\max}$ by at least two standard deviations. For the matches in Table 1, we set an arbitrary significance limit of 25 contiguous matching nucleotides, corresponding to a FASTA score of 100. This limit exceeds $L_{\max}$ by more than two standard deviations for all probes used. Equivalent significance limits, with larger $L_{\max}$, were used for alignments with mismatches.
- Reuter, D., Schuh, R., & Jackle, H. (1989) *Proc. Natl. Acad. Sci. (USA)* **86**, 5483–5486.
- Anderson, K.P., Croyle, M.L., & J.B. Lingrel, J.B. (1989) *Gene* **81**, 119–128.
- Sutcliffe, J. (1979), *Cold Spring Harbor Symp. Quant. Biol.* **43**, 77–90.
- Huynh, T.V., Young, R.A., & Davis, R.W. (1985) in *DNA Cloning: A Practical Approach* ed. Glover, D.M. (IRL Press, Oxford) **1**, p. 49–78.
- Frischauf, A.M., Lehrach, H., Poustka, A., & Murray, N. (1983) *J. Mol. Biol.* **170**, 827–842.
- Bodner, M., Fridkin, M., & Gozes, I. (1985) *Proc. Natl. Acad. Sci. (USA)*, **82**, 3548–3551.
- Saba, A., Busch, H. & Reddy, R. (1985) *Biochem. Biophys. Res. Comm.* **130**, 828–834.
- Joseph, L.J., Chang, L.C., Stamenkovich, D., & Sukhatme, V.P. (1988) *J. Clin. Invest.* **81**, 1621–1629.
- Eich-Helmerich, K., & Braun, V. (1989) *J. Bacteriol.* **171**, 5117–5126.
- Aronson, A.I., Song, H.-Y., & Bourne, N. (1989) *Mol. Microbiol.* **3**, 437–444.
- Ogawa, H. & Fujioka, M. (1988) *Nucleic Acids Res.* **16**, 8715–8716.
- Ming Li, J., Russell, C.S., & Cosloy, S.D. (1989) *Gene* **75**, 177–184.
- Stockinger, H., Schmidtke, J., Bostock, C., & Epplen, J.T. (1986) *Hum. Genet.* **73**, 104–109.
- Maeda, N. (1985) *J. Biol. Chem.* **260**, 6698–6709.
- Newmann, H., Zillig, W., Schwass, V., & Eckerskorn, C. (1989) *Mol. Gen. Genet.* **217**, 105–110.
- Deschenes, R.J., Haun, R.S., Funckes, C.L., & Dixon, J.E. (1985) *J. Biol. Chem.* **260**, 1280–1286.
- Tepler, I., Shimizu, A., & Leder, P. (1989) *J. Biol. Chem.* **264**, 5912–5915.
- Son, H.J., Cook, G.A., Hall, T. & Donelson, J.E. (1989) *Mol. Biochem. Parasitol.* **33**, 59–66.
- Scott, A.L., Dinman, J., Sussman, D.J., Yenbutr, P., & Ward, S. (1989) *Mol. Biochem. Parasitol.* **36**, 119–126.
- Sudol, M., Kieswetter, C., Zhao, Y.-H., Dorai, T., Wang, L.-H., & Hanafusa, H. (1988) *Nucleic Acids Res.* **16**, 9876.
- Nichols, R., Schnewly, S.A., Dixon, J.E. (1988) *J. Biol. Chem.* **263**, 12167–12170.
- May, L.T., Landsberger, F.R., Inouye, M., & Sehgal, P.B. (1985) *Proc. Natl. Acad. Sci. (USA)* **82**, 4090–4094.
- Van den Ouweland, A.M.W., Van Duijnhoven, H.L.P., Deichmann, K.A., Van Groningen, J.J.M., de Leij, L., & Van de Ven, W.J.M. (1989) *Nucleic Acids Res.* **17**, 3829–3843.
- Gharib, S.D., Roy, A., Wierman, M.E., & Chin, W.W. (1989) *DNA* **8**, 339–349.
- Nakashima, H., Yamamoto, M., Goto, K., Osumi, T., Hashimoto, T., & Endo, H. (1989) *Gene* **79**, 279–288.
- Soininen, R., Huotari, M., Ganguly, A., Prockop, D.J., & Tryggvason, K. (1989) *J. Biol. Chem.* **264**, 13565–13571.
- Creighton, T.E., & Charles, I.G. (1987) *J. Mol. Biol.* **194**, 11–22.
- de Martynoff, D., Pohl, V., Mercken, L., van Ommen, G.-J., & Vassart, G. (1987) *Eur. J. Biochem.* **164**, 591–599.
- Lenstra, R., d'Auriol, L., Andrieu, B., Le Bras, J., & Galibert, F. (1987) *Biochem. Biophys. Res. Comm.* **146**, 368–377.
- Tani, T., Ohsumi, J., Mita, K., & Takiguchi, Y. (1988) *J. Biol. Chem.* **263**, 1231–1239.
- Wilson, B.W., Edwards, K.J., Sleight, M.J., Byrne, C.R., & Ward, K.A. (1988) *Gene* **73**, 21–31.
- Rentier-Delrue, F., Swennen, D., Prunet, P., Lion, M., & Martial, J.A. (1989) *DNA* **8**, 261–270.
- Ponzi, M., Birago, C., & Battaglia, P.A. (1984) *Mol. Biochem. Parasitol.* **13**, 111–119.
- Scherer, S.E., Veres, G., & Caskey, C.T. (1988) *Nucleic Acids Res.* **16**, 1593–1601.
- Barklis, E., Pontius, B., & Lodish, H.F. (1985) *Mol. Cell. Biol.* **5**, 1473–1479.
- Ann, D.K., Gadbois, D., & Carlson, D.M. (1987) *J. Biol. Chem.* **262**, 3958–3963.
- Xu Y., Pitcovski, J., Peterson, L., Auffray, C., Bourlet, Y., Gerndt, B.M., Nordskog, A.W., Lamont, S.J., & Warner, C.M. (1989) *J. Immunol.* **142**, 2122–2132.
- Levy, E., Liem, R.K.H., D'Eustachio, P., & Cowan, N.J. (1987) *Eur. J. Biochem.* **166**, 71–77.
- Freimark, B., Pickering, L., Concannon, P., & Fox, R. (1989) *Nucleic Acids Res.* **17**, 455.
- Handy, D.E., Larsen, S.H., Karn, R.C., & Hodes, M.E. (1987) *Mol. Biol. Med.* **4**, 145–155.
- Begley, C.G., Aplan, P.D., Davey, M.P., Nakahara, K., Tchorz, K., Kurtzberg, J., Hershfield, M.S., Haynes, B.F., Cohen, D.I., Waldmann, T.A., & Kirsch, I.R. (1989) *Proc. Natl. Acad. Sci. (USA)* **86**, 2031–2035.
- Sakaguchi, N., Kashiwamura, S., Kimoto, M., Thalmann, P., & Melchers, F. (1988) *EMBO J.* **7**, 3457–3464.
- Vanderslice, P., Craik, C.S., Nadel, J.A., & Caughey, G.H. (1989) *Biochem.* **28**, 4148–4155.
- Corral, M., Baffet, G., & Defer, N. (1988) *Nucleic Acids Res.* **16**, 10935.
- Greslin, A.F., Prescott, D.M., Oka, Y., Loukin, S.H. & Chappell, J.C. (1989) *Proc. Natl. Acad. Sci. (USA)*, **86**, 6264–6268.
- Hartley, D.A., Preiss, A., & Artavanis-Tsakonas, S. (1988) *Cell* **55**, 785–795.
- Cited in GenBank Release 63 as Ganguly, S. & Chaganti, R.S.T. (1988) *Nucleic Acids Res.*, in press.
- Hodgson, C.P. (1990) *Biotechniques* **9**, 54–55.
- Lopez, R., Kristensen, T., & Prydz, H. (1992) *Nature* **355**, 211.
- One example is Auto-search developed by Kathleen Klose of the MBCRR, Boston, MA, incorporating features of FASTA (ref. 4) the comparison program BLAST (Altschul, S.F., Gish, W., Miller, W., Myers, E.W., & Lipman, D.J. (1990) *J. Mol. Biol.* **215**, 403–410).
- Cinkosky, M.J., Fickett, J.W., Gilna, P., & Burks, C. (1991) *Science* **252**, 1273–1277.
- Fact sheets, April May 1991, National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD.
- States, D.J. & Botstein, D. (1991) *Proc. Natl. Acad. Sci. (USA)* **88**, 5518–5522.
- Brunak, S., Engelbrecht, J., & Knudsen, S. (1990) *Nature* **343**, 123.
- Roberts, L. (1991) *Science* **252**, 1255–1256.
- Olson, M., Hood, L., Cantor, C., & Botstein, D. (1989) *Science*, **245**, 1434–1435.
- Roberts, L. (1989) *Science*, **245**, 1438–1440.

62. Sequences, maps, numbering came from: Irwin, N. (1989) *Molecular Cloning* eds. Sambrook, J., Fritsch, E.F. & Maniatis, T. (Cold Spring Harbor Laboratory Press, New York) ed. 2, ch. 2; the database entry LAMCG notation, other entries as specified, in GenBank Release 63; from *Lambda II* (1983) ed. Hendrix, R.W., Roberts, J.W., Stahl, F.W., & Weisberg, R.A. (Cold Spring Harbor Laboratory Press, New York), app. II, III.
63. Short, J.M., Fernandez, J.M., Sorge, J.A. & Huse, W.D. (1983) *Nucleic Acids Res.* **16**, 7583–7600.
64. Yamagami, T., Ohsawa, K., Nishizawa, M., Inoue, C., Gotoh, E., Yanaihara, N., Yamamoto, H., and Okamoto, H. (1988) *Ann N.Y. Acad. Sci.* **527**, 87–102.
65. Emi, M., Horii, A., Tomita, N., Nishide, T., Ogawa, M., Mori, T., and Matsubara, K. (1988) *Gene* **62**, 229–235.
66. Gardner, R.C., Howarth, A.J., Messing, J. & Shepherd, R.J. (1982) *DNA* **1**, 109–115.
67. Gadaleta, G. et al. (1989) *J. Mol. Evol.* **28**, 497–516.